Modèles statistiques précoces et robustes pour l'estimation de la concentration d'agents biologiques dans un système de surveillance en continu dans l'environnement

Abou Keita

Laboratoire Informatique,
du Traitement de l'Information et des Systèmes

Normandie Université
INSA de Rouen

T H È S E
en vue de l'obtention du titre de
Docteur en Sciences de l'INSA de Rouen

Présentée et soutenue par
Abou KEITA

****

Modèles statistiques précoces et robustes pour l'estimation de la concentration d'agents biologiques dans un système de surveillance en continu dans l'environnement

****

Thèse dirigée par Stéphane CANU et Romain HÉRAULT

préparée au LITIS de l'INSA de Rouen, Projet GENEASE

15/12/2014

Jury :

Rapporteurs :	NUZILLARD Danielle	-	CReSTIC, Université de Reims Champagne Ardenne
	BEAUSEROY Pierre	-	LM2S, UTT Troyes
Directeur :	CANU Stéphane	-	LITIS, INSA de Rouen
Encadrant :	HÉRAULT Romain	-	LITIS, INSA de Rouen
Examinateurs :	PORTIER Bruno	-	LMI, INSA Rouen
	BRUN Luc	-	GREYC, ENSICAEN Caen

« Cela semble toujours impossible, jusqu'à ce qu'on le fasse »

Nelson Mandela

À ma femme,
À ma famille et
À mes ami(e)s.

Remerciements

Tout d'abord, je souhaite exprimer toute ma gratitude à mon directeur de thèse Monsieur CANU Stéphane et à mon encadrant Monsieur HÉRAULT Romain. Ils ont dirigé et encadré l'ensemble de mes travaux de recherches. Je salue la pertinence de leurs suggestions et remarques très constructives, leurs précieux conseils et expertises dans de nombreux domaines techniques et scientifiques : informatique, statistiques, optimisation, traitement d'image, etc... Leur disponibilité et leur étroite collaboration durant la réalisation de cette thèse m'ont permis d'approfondir mes connaissances dans leurs domaines d'expertises.

Je souhaite également remercier :

— les membres du jury pour leur disponibilité et l'intérêt qu'ils portent à mes travaux de recherches ;

— les membres du Laboratoire Informatique du Traitement de l'Information et des Systèmes (LITIS) pour leur appui.

Je donne une mention spéciale à chacune des personnes qui ont œuvré de leur meilleure façon pour me fournir du support et de l'assistance dans l'aboutissement de mes études et tout le long de mon séjour en France pour concrétiser mon projet d'études supérieures. Sans aucun doute, je n'oublierai pas les meilleurs moments que nous avons partagé dans un cadre soit amical, soit universitaire, soit professionnel. Je tiens à dire une fois de plus, merci à Yannick TCHAPTCHET, Babacar NDIAYE, Olivier OKITANGOMO SHONDA, Adja Maimouna SOUMARE, Coumba BA, Cheikh Mouhamed DIOP, Boubacar TRAORE, Mr et Mme KA, Leaticia ADER YA, Cynthia GIRANEZA, Jean Claude TWAGIRAMUNGU, Alioune GUEYE, Birane GAYE, Mr et Mme BAL, Alioune NDIAYE, Moussa DIOP et Maniang GUISSÈ.

Je remercie tout particulièrment ma famille, principalement à mes adorables parents sans qui cette expérience n'aurait jamais été possible. Merci à mes frères et sœurs pour leur réconfort et leur encouragement.

Je garde le meilleur pour la fin en remerciant du fond du cœur ma femme, Ndeye Marianne KEITA, pour sa présence, son appui moral, sa patience et ses corrections apportées à ce manuscrit. Merci mon astre !

Abou KEITA

Table des matières

Table des figures

Liste des tableaux

Symboles et notations

Notations basiques

Notation	Signification
c	Numéro de cycle
j	Indice liée au numéro de cycle
m , p	Pente et ordonnée à l'origine de la droite de régression
$\mathbb{R}$	Ensemble des nombres réels
y	Variable de sortie
$x \in \mathbb{R}^n$	Vecteur de variable d'entrée
$\mathbf{x} \in \mathbb{R}^n$	Observation de x
X	Variable aléatoire
ε	Bruit

Notation pour la probabilité et les statistiques

Notation	Signification
$\mathcal{H}_0$	Hypothèse nulle
$\mathcal{H}_1$	Hypothèse alternative
α	Probabilité de rejeter à tord $\mathcal{H}_0$
θ	Paramètre caractérisant un changement sur les données
$p_\theta(.)$	Densité de probabilité de paramètre θ
$L(X;\theta)$	Fonction de Vraisemblance
$L_{\mathcal{H}}(X;\theta)$	Fonction de Vraisemblance sous l'hypothèse $\mathcal{H}$
$s(X)$	Log rapport de vraisemblance

Notation pour la fluorescence

Notation	Signification
PCR	*Plymerase Chain Reaction*
qPCR	PCR quantitative
F_c	Fonction représentant la fluorescence
O_c	Fonction représentant la fluorescence par notre méthode
Q_c	Quantité molécule d'ADN au cycle c
Q_0	Quantité de molécule d'ADN initiale
E	Efficacité d'amplification
$c_{1/2}$	Cycle repère par la méthode sigmoïde
C_{y_0}	Cycle repère par la méthode de Richards
SCM	*Sigmoïd Curve fitting Method*

Notation pour la détection de changement

Notation	**Signification**
CUSUM	CUmulative SUMme (somme cumulée)
c_{max}	Cycle maximum
t_{reg}	Cycle limite où le pré traitement est effectué
θ_0	Paramètre avant changement
θ_1	Paramètre après changement
c_r	Cycle repère
c_a	Cycle ou instant d'alarme
δ	Retard à la détection
$\Phi_{c_r}(j)$	Indicateur du CUSUM (log rapport de vraisemblance des observations de c_r à j)
M_j	Valeur maximale de $\Phi_{c_r}(j)$
g_j	Statistique de test ou fonction de décision
d	Règle de décision
h	Seuil de la règle du CUSUM
$\mu_{0,1}$	Espérance avant et après changement
$\sigma^2_{0,1}$	Variance avant et après changement

Autres

Notation	**Signification**
IQR	Écart interquartile
Q_1	1^{er} quartile
Q_3	3^e quartile
err	Erreur d'estimation de la fluorescence
$\bar{e}$	Erreur d'estimation moyenne de la fluorescence
R^2	Cœfficient de détermination
I_{ref}	Image fixe ou de référence
I_{reca}	Image en mouvement ou à recaler
ξ	Résidu moyen de la log régression
e_k	Erreur d'estimation du k^{eme} observation
MSE_{loocv}	Erreur de la cross validation moyenne
e	Résidu
β	Paramètre de la fonction sigmoïde

Introduction

CETTE THÈSE s'inscrit dans le cadre du projet ANR[4] *Génétic EquipemeNt for biothrEat enviroNmental Analysis and SurveillancE* (GENEASE) en partenariat avec Bertin Technologies, CEA[5] et le LHVP[6].

Contexte et positionnement de la problématique[7]

L'émergence de la menace biologique est devenue une réalité, avec notamment les attaques à l'Anthrax en 2001, qui ont conduit en Octobre 2001 à la mise en place du plan Biotox complétant le dispositif ORSEC (Organisation de la Réponse de Sécurité Civile) de gestion des risques et de la crise face à une menace biologique.

Dans ce cadre, la détection la plus précoce d'une agression et sa caractérisation de manière fiable permet une réponse optimale en terme d'organisation de secours et de dispositions à adopter afin d'en limiter les conséquences. Au moment de l'élaboration du projet, il n'existait pas un tel système de surveillance et d'analyse biologique de l'environnement.

Ainsi, le projet GENEASE porte sur l'étude d'un système de surveillance et d'analyse biologique en continu de l'environnement, compact et sensible. Cette analyse se fait par une méthode de biologie moléculaire par la détection et l'identification de plusieurs agents biologiques simultanément. Le dispositif visé doit intégrer l'ensemble des fonctions de la collecte d'échantillons à la prise de décision. Ce projet s'inscrit essentiellement dans l'axe « gestion de crise » puisqu'il porte sur l'étude d'un équipement mobile et portable de détection et d'identification biologique.

De plus, cet équipement participe à la caractérisation de la typologie (gravité) des victimes d'un attentat (conventionnel ou non conventionnel), et l'identification du pouvoir contaminant des substances impliquées. Ceci est réalisé afin de dimensionner des contremesures médicales et techniques en fonction des critères de gravité et du nombre de victimes concernées.

4. Agence Nationale de la Recherche
5. Commisariat à l'Énergie Atomique
6. Laboratoire d'Hygiène de la Ville de Paris
7. Tiré du document de soumission du projet ANR, GENEASE

Présentation de la problématique et objectif

Le laboratoire LETI[8] du CEA développe des solutions technologiques miniaturisées qui sont basées sur la micro-fluidique digitale par électromouillage. Cette technologie permet de réaliser facilement les étapes fluidiques indispensables à la réalisation de protocoles complexes. Parmi ces protocoles, nous avons le déplacement de gouttes (fragments d'ADN[9]), la formation de gouttes d'échantillons ou de réactifs de volume contrôlé et le mélange rapide de deux gouttes uniquement par contrôle électrique des électrodes.

Un signal de fluorescence issu de l'immuno-analyse est obtenu via une caméra qui filme la puce « Smart Drop » : laboratoire sur puce. Ce signal est fonction de la concentration de l'espèce à qualifier dans la goutte analysée et correspondant à la quantification par fluorescence d'une réaction d'amplification par PCR (*Polymerase Chain Reaction*). Ce dernier est une technique d'amplification d'un fragment d'ADN par une réaction enzymatique (obtention suffisante d'ADN pour pouvoir effectuer des analyses classiques).

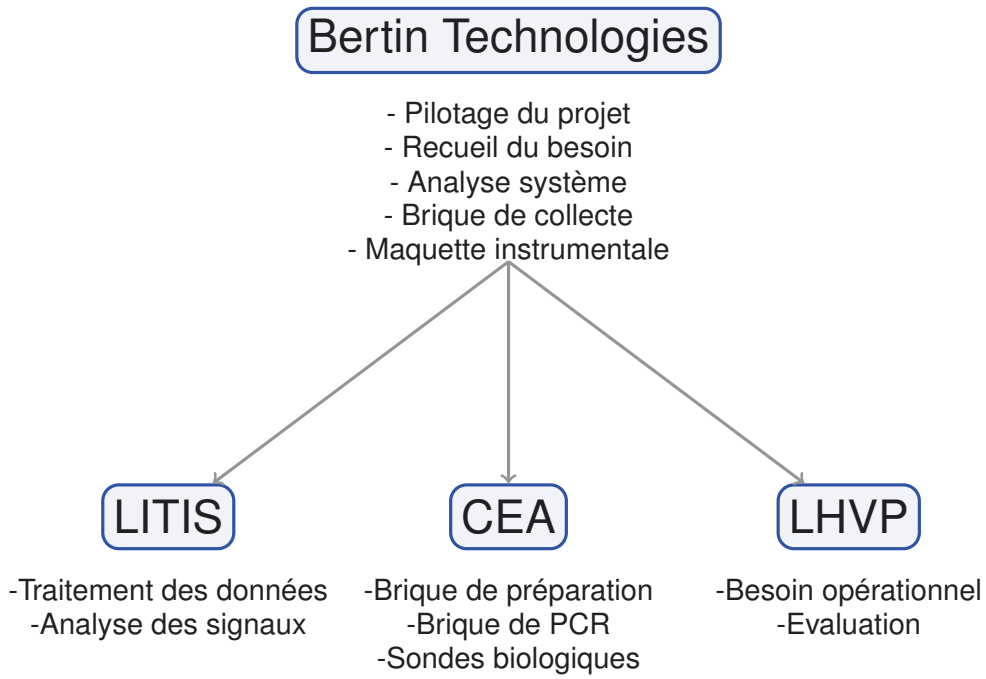


FIGURE 1 – Schéma institutionnel.

8. Laboratoire au sein du CEA qui concentre son activité sur les micro et nano technologies et leurs applications aux systèmes et composants de communication sans fil. Leur activité s'étale aussi dans à la biologie, à la santé, à l'imagerie, et aux Micro-Nano Systèmes (MNS).

9. Acide DésoxyriboNucléique

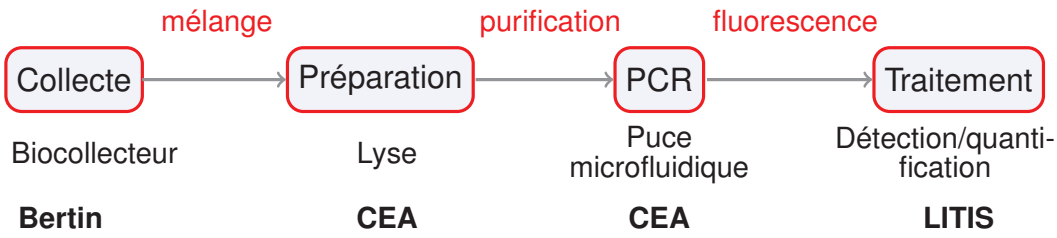


FIGURE 2 – Schéma fonctionnel.

Bertin Technologies a développé la brique de collecte associée à un système de transfert de fluide. La collecte de l'échantillon est réalisée par un biocollecteur miniaturisé qui aspire et concentre les particules contenues dans l'air dans un échantillon liquide. Le liquide est régulièrement injecté via une pompe périlstatique reliée à un réservoir de manière à avoir un volume final d'échantillon toujours égal à $1,5\ mL$. Un système de transfert de l'échantillon permet ensuite via un piston d'aspirer le collectât en même temps que la solution de billes magnétiques afin d'effectuer le mélange. Une fois l'ensemble du collectât aspiré, le mélange est envoyé vers la brique de préparation de l'échantillon.

Le CEA Leti a développé une maquette permettant de traiter un échantillon de $1\ mL$ afin d'en purifier l'ADN puis de la détecter par PCR quantitative. Un protocole basé sur l'utilisation de billes magnétiques avec des volumes adaptés à la maquette a ainsi été développé. Le système complet a été validé sur deux espèces modèles : les bactéries Escherichia coli et Serratia marcescens. La préparation a été effectuée avec une lyse chimique.

La brique d'analyse PCR, développée par le CEA, est effectuée avec une solution technologique miniaturisée basée sur la micofluidique digitale par électromouillage. Cette technologie permet de réaliser facilement les étapes fluidiques comme le déplacement, la formation et le mélange rapide de gouttes par contrôle électriques d'électrodes. Une fois la PCR réalisée sur cette puce, un signal de fluorescence est obtenu.

Notre tâche (LITIS) au sein du projet GENEASE est de mettre en place un système qui détecte (ou non) la présence de l'espèce puis d'estimer sa concentration. Cette estimation est faite à partir des données de fluorescence via un indicateur que nous appelons cycle repère (ou instant de rupture). Il s'agit donc d'intégrer dans la même approche la discrimination pour détecter et la régression pour qualifier.

En outre, le but est de construire un modèle statistique qui, à partir de la seule observation de la fluorescence d'une goutte, calcule une estimation du cycle repère et en déduire une estimation de la concentration de l'espèce à qualifier tout en minimisant le retard à la détection. Le travail constitue à poser un modèle statistique

pour déterminer en temps continu à partir de quel instant on observe ce cycle repère sur la fluorescence.

Contributions

Pour mettre en place une règle de décision sophistiquée pour la détection ou non d'une espèce biologique recherchée, cette étude est scindée en 4 chapitres : de la définition de quelques notions biologiques, en passant par une présentation des méthodes de l'état de l'art et en évoquant les problématiques jusqu'à la présentation de la méthode utilisée et des résultats obtenus.

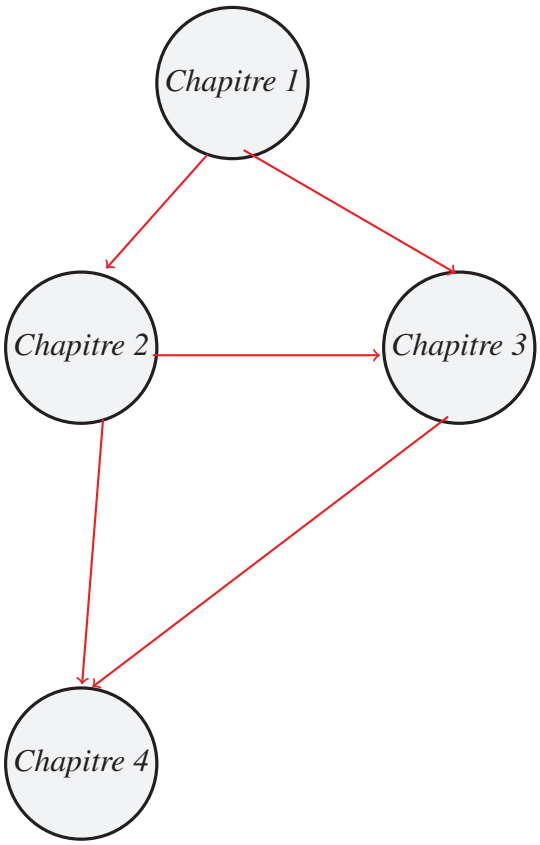


FIGURE 3 – Plan du thèse.

Chapitre 1 **:** Dans ce chapitre, nous présentons la réaction en chaine par polymérase (PCR) et la fluorescence issue d'une PCR sur un appareil standard. Nous évoquons aussi comment un signal de fluorescence peut être obtenu sur une puce microfluidique et comment les gouttes sur la puce peuvent être guidées par recalage d'image.

Chapitre 2 **:** Ce chapitre présente différentes techniques utilisées dans l'état de l'art pour déterminer un cycle repère de la fluorescence, étape nécessaire à la quantification.

Chapitre 3 **:** Ce chapitre étudie un pré-traitement pour sélectionner et supprimer des erreurs de mesures sur la fluorescence en analysant les résidus de la régression L_1 et L_2 par la méthode de la boite à moustache. En plus, nous allons proposer un modèle analytique représentatif d'une fluorescence en tenant compte de la décroissance que l'on peut observer sur la fluorescence au début de la quantification de la PCR.

Chapitre 4 **:** Ce chapitre étudie la méthode statistique de somme cumulée, le CUSUM, pour construire une règle de décision sophistiquée pour la détection (ou non) de la présence d'une espèce biologique recherchée tout en minimisant le retard à la détection.

CHAPITRE 1

Présentation du signal de fluorescence issu d'une réaction en chaine par polymérase (*Polymerase Chain Reaction*) et techniques de détection du cycle repère

Sommaire

DURANT CETTE THÈSE, nous avons travaillé sur des signaux de fluorescence issus d'une réaction en chaine par polymerase. Selon le comportement de cette fluorescence, nous pouvons par exemple déterminer la quantité initiale de molécule d'ADN d'une espèce biologique recherchée. Seule une séquence particulière de l'ADN est ciblée par le fluorophore émettant la lumière. Ainsi nous pouvons mesurer spécifiquement la quantité d'ADN (donc la concentration) d'un virus, d'une bactérie, d'un anticorps, etc., même si différents ADN se trouvent dans l'échantillon initial.

1.1 Principe de la réaction en chaine par polymérase

Au début des années 80, le diagnostic des maladies infectieuses et génétiques étaient basées sur des sondes. Ces techniques nécessitent une culture biologique pour augmenter le nombre de cellules ou d'organismes à déceler, mais cette culture n'est pas toujours facile ou réussie [Schochetman et al., 1988]. Le processus chimique de réaction en chaine par polymerase (PCR) a permis de ne plus dépendre du processus biologique de culture. La PCR consiste en l'amplification in vitro de l'ADN ou de l'ARN [1] [Erlich, 1989]. C'est un outil universel dans le domaine de la biologie pour la détection des acides nucléiques. Avant de détailler le principe de la PCR, nous allons d'abord définir quelques termes biologiques utilisés.

1.1.1 Définitions de quelques notions biologiques

L'unité fondamentale de tout organisme vivant est la cellule. Pour les eucaryotes (ce qui nous concernent), la cellule comprend essentiellement deux parties : le cytoplasme et le noyau. Ce dernier contient les chromosomes où se trouve une molécule d'Acide DesoxyriboNucléique (ADN). Dans cette dernière est contenue l'information génétique qui constitue le génotype d'un organisme. La molécule d'ADN est composée de deux brins face à face formant une double hélice. Chaque hélice est caractérisée par la succession de nucléotides ou bases azotées : Adénine (A), Thymine (T), Guanine (G), Cytosine (C) (fig. 1.1). Entre chaque hélice, on observe une complémentarité des bases azotées : T associée à A et inversement ; G associée à C et inversement.

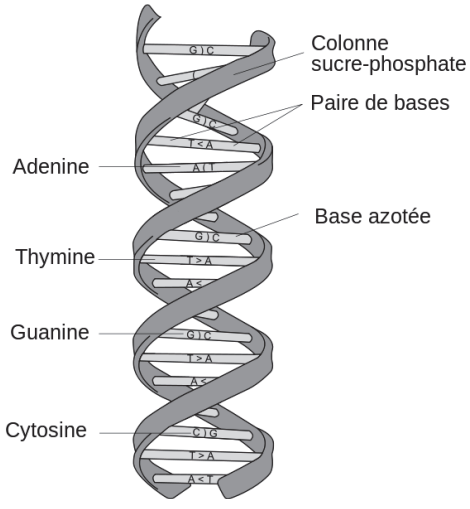


FIGURE 1.1 – Fragment d'ADN[2]

1. Acide RiboNucléique

1.1.1.1 Expression génétique

L'expression génétique est un processus biochimique qui permet de traduire un gène en protéines. Ceci se fait en deux phases (fig. 1.2) :

— La transcription : de l'ADN vers l'Acide RiboNucléique (ARN). Un seul brin (ou hélice) d'ADN est utilisé pour la transcription. La molécule d'ARN est construite par complémentarité au brin d'ADN.

— La traduction : de l'ARN vers une protéine.

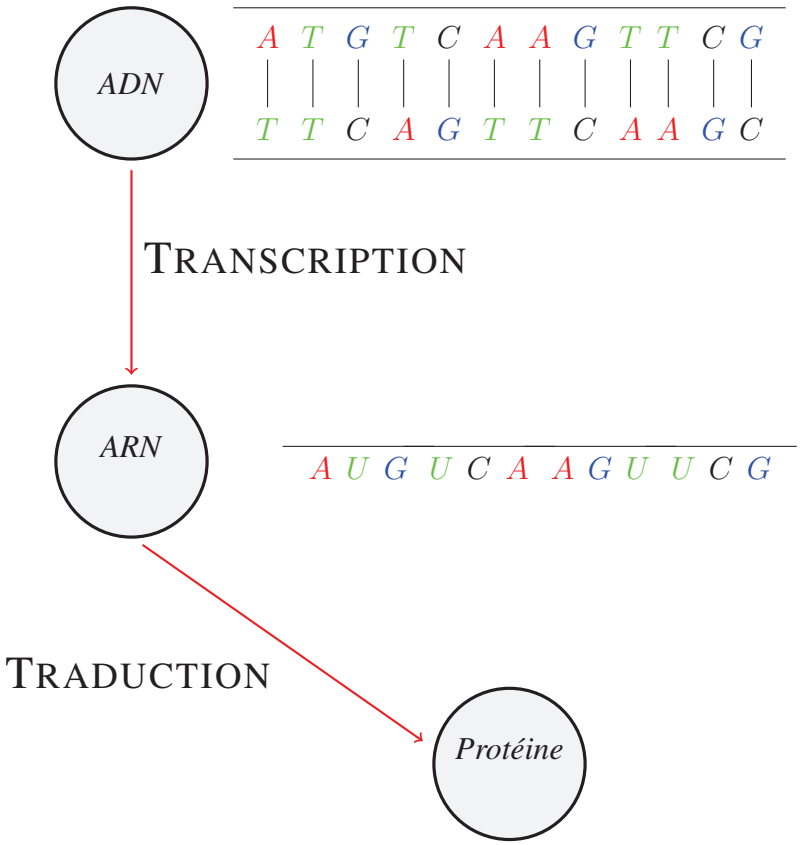


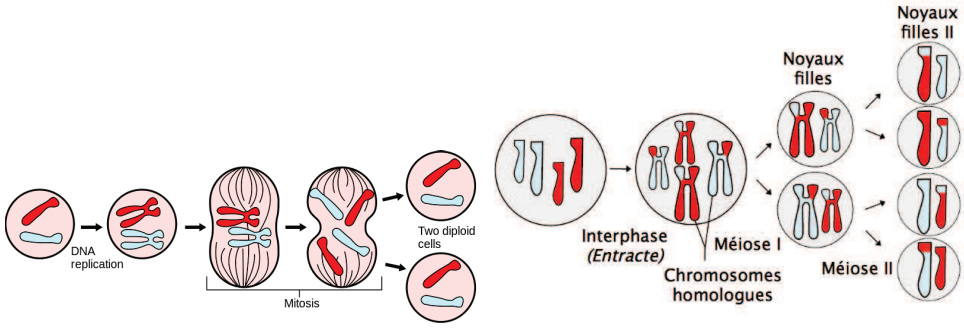
FIGURE 1.2 – Illustration de l'expression génétique.

L'ARN est aussi composée d'une succession de nucléotides : Uracile (U), Adénine (A), Guanine (G) et Cytosine (C). On note que la thymine (T) du brin de l'ADN devient l'uracile (U) en ARN.

2. DNA structure and bases PLsvg. Utilisable sous les conditions GNU Free Documentation License.

1.1.1.2 Reproduction cellulaire

La reproduction ou division cellulaire est le mode de multiplication de toute cellule. Elle comporte deux catégories : la mitose (fig. 1.3(a)) et la méiose (fig. 1.3(b)). La mitose assure la naissance de cellules identiques à la cellule mère et la méiose aboutit à la production de cellules sexuelles. Ainsi durant une reproduction cellulaire, l'information génétique contenue dans l'ADN est recopiée.



(a) Étape du processus mitose (b) Étape du processus méiose

FIGURE 1.3 – Reproduction cellulaire [3]

Dans ce qui suit, nous allons détailler le concept de la PCR. Reproduisant les phénomènes à l'œuvre lors de la division cellulaire, l'ADN est multiplié ou répliqué grâce à l'ADN polymérase qui est un ADN assurant la recopie de la même information génétique de la molécule d'ADN matrice à une nouvelle molécule avec quelques erreurs de recopie.

1.1.2 Cadre géneral de la PCR

La réaction en chaine par polymérase appelée *Polymerase Chain Reaction* (PCR) en anglais, est un concept technologique de multiplication en chaîne découvert par Kary Mullis en 1985 [Mullis and Faloona, 1987]. Une PCR permet d'obtenir, à partir d'un échantillon d'ADN peu abondant, d'importantes quantités d'ADN. En quelques heures, le nombre de molécules d'ADN obtenues peut atteindre des millions de copies suffisant pour une analyse ultérieure [Muyzer et al., 1993].

La PCR consiste en une répétition de réactions chimiques en cycle qui aboutissent à chaque cycle au doublement de la quantité d'ADN grâce aux polymérases. Ces dernières sont des enzymes qui permettent la synthèse du nouveau brin d'ADN.

3. National Institutes of Health, mitose, meiose, Wikipédia. Licence : domaine public.

Avant la réalisation du premier cycle de la PCR, l'ADN est placé à température ambiante et il est sous sa forme en double hélice. Cet état est appelé *conditions natives* (étape 0 sur la fig. 1.4). Il s'en suit l'étape intitulée *Dénaturation initiale* qui ne sera réalisée qu'une seule fois avant le tout premier cycle (étape $1'$ sur la fig. 1.4). Elle consiste à chauffer à 95 ^{o}C pendant 10 à 15 min environ jusqu'à déshybrider (séparer) les ADN double brin, de casser les structures secondaires, d'homogénéiser le milieu par agitation thermique, d'activer les polymérases et de dénaturer d'autres enzymes qui pourraient être présents dans la solution.

Ensuite comme nous le montre la figure (1.4), nous répètons les 3 opérations suivantes pour chaque cycle d'environ 1 min :

1. *Dénaturation* : Cette étape permet de déshybrider les ADN, d'enlever les polymérases restantes et d'homogénéiser le milieu réactionnel avec une température maintenue à 95 ^{o}C.
2. *Hybridation* : On fait varier la température entre 40 et 65 ^{o}C pour permettre aux amorces de s'hybrider (se lier) aux brins d'ADN matrices.
3. *Élongation* : Cette étape permet aux polymérases de synthétiser le brin complémentaire de leur ADN matrice à une température qui remonte légèrement jusqu'à 72 ^{o}C.

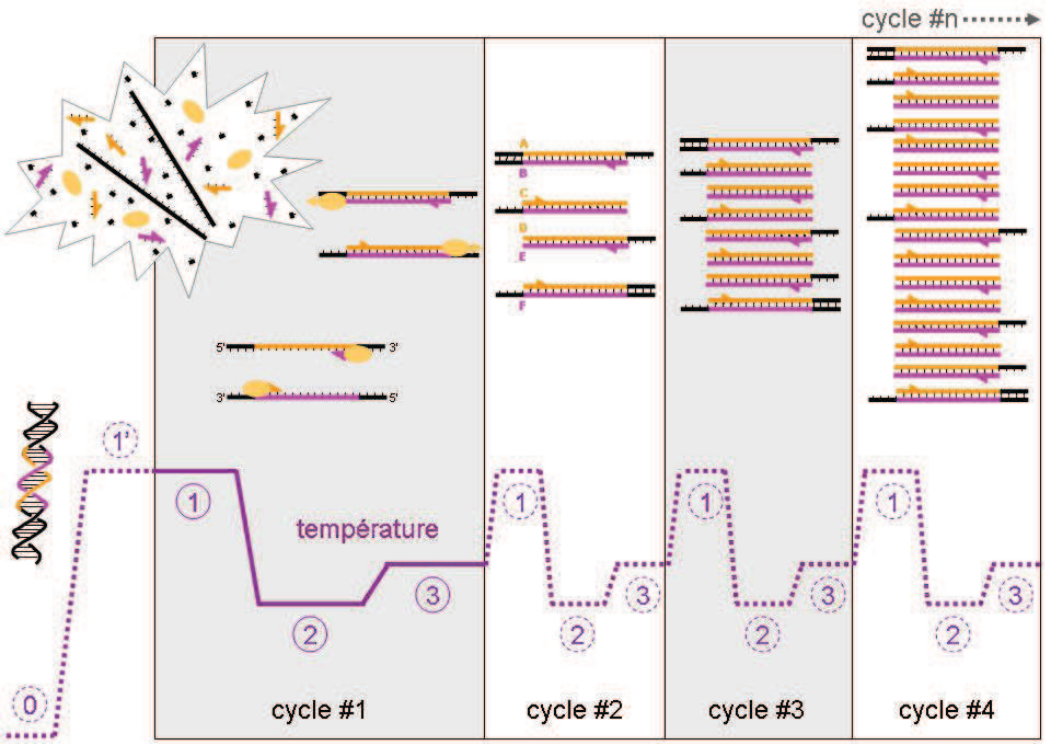


FIGURE 1.4 – Etapes d'un cycle d'une PCR et enchainement des cycles [4].

1.1.3 Techniques associées

Il existe plusieurs techniques associées à la PCR. Parmi celles-ci nous pouvons citer :

a) *Reverse Transcriptase* PCR (RT-PCR) : Il s'agit d'une PCR "classique" réalisée sur un ADN complémentaire (ADNc), qui est une copie d'un ARN, obtenue par une transcription inverse [5]. Elle a été mise en place pour utiliser les ARN comme matrice d'amplification de la PCR [Liu and Saint, 2002]. Elle peut être utilisée pour la construction de plusieurs ADNc et pour la construction de sondes d'ADN.

b) PCR multiplex : C'est un processus qui, avec l'utilisation de trois amorces (courtes séquences d'ARN ou d'ADN) au moins par réaction de PCR, permet d'amplifier plus d'un ADN à la fois [Markoulatos et al., 2002]. Comme exemple d'utilisation, la PCR-multiplexe a été utilisée pour l'analyse des microsatellites (séquence d'ADN formée par une répétition continue de motifs composés de 2 à 10 nucléotides) et les polymorphismes d'un seul nucléotide (variation d'une seule paire de bases du génome entre les individus d'une même espèce) [Hayden et al., 2008].

c) PCR en point final : C'est la première technique inventée pour mesurer la quantité d'ADN. La mesure se fait à la fin de tous les cycles d'où le nom de point final. Elle est nommée ainsi que depuis l'apparition de la PCR quantitative (cf point suivant). Néanmoins, il n'y a pas de différence du point de vue biologique entre ces deux concepts.

d) PCR quantitative (en temps réel) : Cette technologie est basée sur la détection et la quantification d'un agent biologique pendant la réaction de la PCR à partir d'un traceur fluorescent (voir sous partie suivante) [Ramakers et al., 2003].

Pour notre travail, nous nous plaçons dans le cadre de la PCR quantitative que nous allons détailler dans la sous partie qui suit.

1.1.4 PCR quantitative (qPCR)

C'est une technique destinée à mesurer la quantité d'ADN initialement présente dans un échantillon. À chaque cycle d'amplification, la quantité d'ADN totale (amplicon) est mesurée grâce à un marqueur fluorescent. Un suivi en temps réel de la qPCR à chaque cycle est réalisé et le signal d'émission de la fluorescence obtenu est représenté par une courbe (fig. 1.5) [Livak and Schmittgen, 2001]. Ce signal permet d'obtenir une quantification absolue ou relative de l'ADN cible.

4. Ygonaar, PCR, Wikipédia. Licence d'utilisation : GFDL + CC-BY-SA
5. C'est la synthèse d'un brin d'ADN à partir d'une matrice ARN.

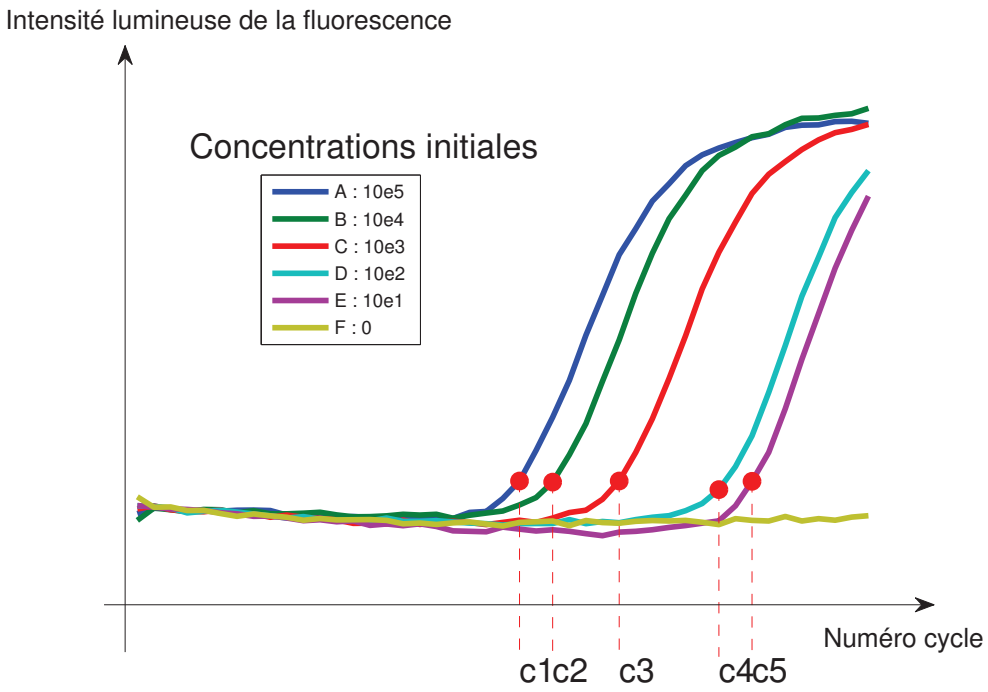


FIGURE 1.5 – 6 échantillons de fluorescences disponible au LHVP de concentration initiale de différente formant d'une même gamme.

La quantité d'ADN initiale influe sur la position de la montée de la fluorescence (cycle repère), les $c_r^i, i = 1, \ldots, 5$, représentés sur la figure (1.5). Plus il y a de quantité d'ADN initiale, plus la montée apparait plutôt. Il n'y a pas de montée s'il n'y a pas d'espèces biologiques recherchées, comme nous pouvons l'observer sur l'échantillon F où la quantité initiale est égale à 0 (fig. 1.5). Nous allons utiliser cette propriété de la montée afin de déterminer la quantité initiale d'ADN.

Matériels de PCR en temps réel : Le thermocycleur est un appareil automatisant la réaction en chaine par polymérase. Il est couplé à un système de détection de fluorescence appelé fluoromètre. Parmi les matériels, nous pouvons citer le Light Cycler^{TM} qui est commercialisé par Roche Diagnostics. Il y a aussi les ABI- Prism^{TM} 7000 qui sont fabriqués par Applied Biosystems et le RotorGene Q commercialisé par Qiagen.

Dans la section suivante, nous allons présenter la fluorescence issue de la quantification par PCR, ainsi que différentes méthodes pour déterminer un cycle repère de la fluorescence.

1.2 Fluorescence

1.2.1 Présentation d'un signal de fluorescence issu d'une qPCR

Une fluorescence est un signal obtenu après un suivi en temps réel d'une qPCR (Réaction en Chaine par Polymérase quantitative). Les données de la fluorescence sont collectées à chaque cycle de la PCR et représentent la quantité de produits amplifiés au total [Ginzinger, 2002].

Si un signal de fluorescence possède une montée, on annote la montée par un cycle repère ou cycle caractéristique. Ce cycle repère permet d'estimer la quantité initiale Q_0 de la molécule d'ADN amplifié.

Modèle théorique : Théoriquement, la quantité de molécules produites par PCR double à chaque cycle d'amplification suivant le modèle ci-dessous [Tse and Capeau, 2003] :

$$Q_c = Q_0\, 2^c \tag{1.1}$$

où Q_0 est le nombre de molécules présentes à l'origine, c est le nombre de cycles d'amplification et Q_c est le nombre de produits PCR présents au cycle c.

Observations : Le modèle théorique (1.1) représente mal la réalité expérimentale. Le suivi en temps réel d'une réaction quantitative de la PCR observée, nous donne une courbe schématisée sur la figure (1.6).

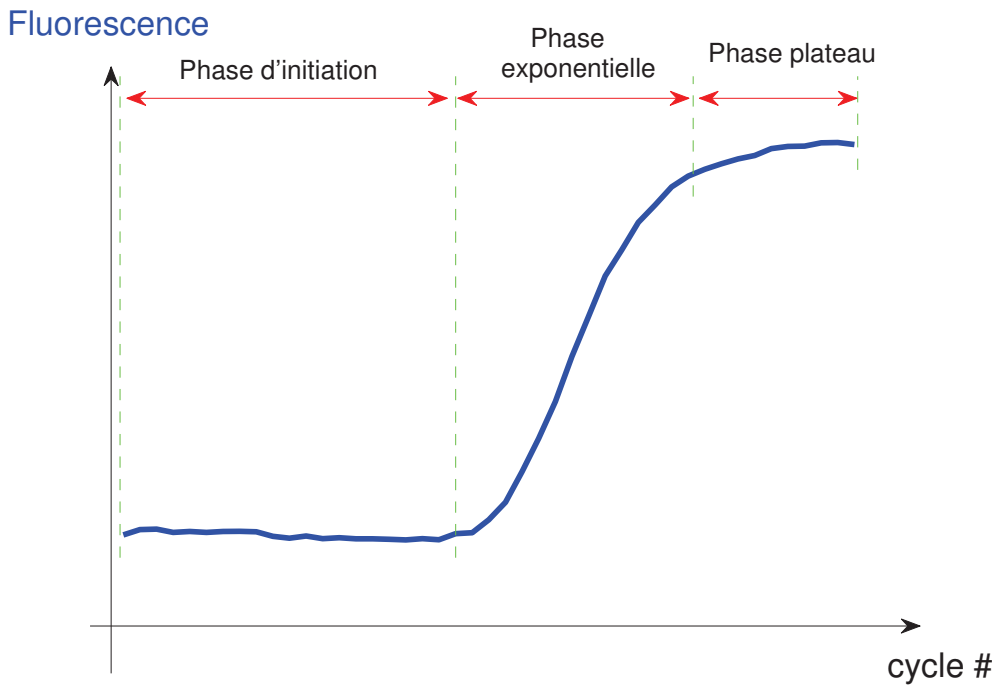


FIGURE 1.6 – Allure d'une fluorescence issu d'une réaction de qPCR.

La fluorescence obtenue, est composée de trois phases : phase initiale (ou ligne de base représentant l'intensité du bruit de fond de fluorescence), phase exponentielle et phase plateau. Au début de la réaction et durant toute la phase d'initiation, les réactifs ont une concentration trop faible. Puis durant la phase exponentielle, l'amplification est réalisée de façon constante avec un doublement du nombre de produits PCR à chaque cycle. Suit enfin la phase plateau au cours des derniers cycles d'amplification où le taux d'amplification décroît à près de zéro générant très peu d'amplicons. Cette décroissance est due à l'épuisement des différents réactifs de la PCR et l'inactivation thermique partielle de l'ADN polymérase. On l'appelle aussi l'effet plateau ou saturation.

Dans le cadre de cette thèse, nous avons aussi travaillé sur des fluorescences issues de la puce microfluidique du CEA-LETI. La brique d'analyse PCR, développée par le CEA, est composée d'une instrumentation mais aussi d'un consom-

mable très spécifique permettant de réaliser les réactions à des échelles de l'orde de nL (gouttes de $64\ nL$). La méthode utilisée pour déplacer des gouttes, repose sur le principe de l'électromouillage sur diélectrique, appelé EWOD (*Electrowetting On Dielectric*). Les forces utilisées sont des forces électrostatiques. La goutte se déplace sur un réseau d'électrodes, dont elle est isolée par une couche diélectrique et une couche hydrophobe. Lorsque l'électrode à proximité de la goutte est activée, la goutte est attirée de façon électrostatique sur la surface de cette électrode. Une PCR est alors réalisée sur cette puce grâce à la microfluidique qui est la science de la manipulation des fluides à l'échelle micrométrique.

1.2.2 Fluorescences issues de la puce microfluidique

À chaque cycle de réaction, une photo est prise sur la puce microfluidique. Nous n'obtenons pas alors directement la fluorescence. Pour ce faire, nous allons utilisé le principe de recalage d'image pour trouver la position des gouttes sur la puce (voir ci dessous).

Nous disposons d'une série d'images provenant de la puce miniaturisée dont le schéma de surface est représentée sur la figure 1.7.

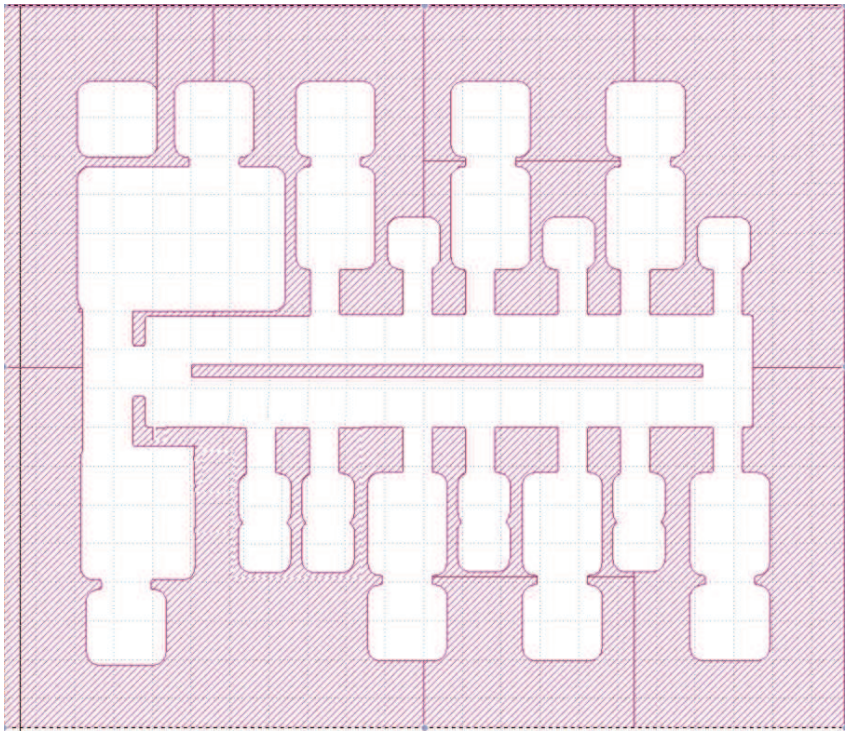


FIGURE 1.7 – Représentation de la surface d'une puce microfluidique du CEA-LETI.

Sur cette puce sont représentés différents compartiments (puits) dont 6 en allant de droite à gauche sur sa partie supérieure et 7 sur sa partie inférieure. Ces 13 compartiments sont séparés par un bus (ligne horizontale au milieu de la puce) qui est un passage pour les gouttes.

Un ensemble d'électrodes est répartie sur la surface de la puce. Les gouttes sont transportées jusqu'aux puits par guidage électrostatique où sera réalisée la réaction par électro-mouillage. La Réaction en Chaine par Polymérase (PCR) est réalisée à l'intérieur de ces puits où sont placés des réactifs séchés. En effet, la fluorescence est obtenue en calculant la moyenne sur une zone d'intérêt de la goutte à chaque cycle de réaction.

Nous ne disposons pas directement du signal de fluorescence mais d'une série d'images de la plaque où est effectuée la PCR (par exemple fig. 1.8).

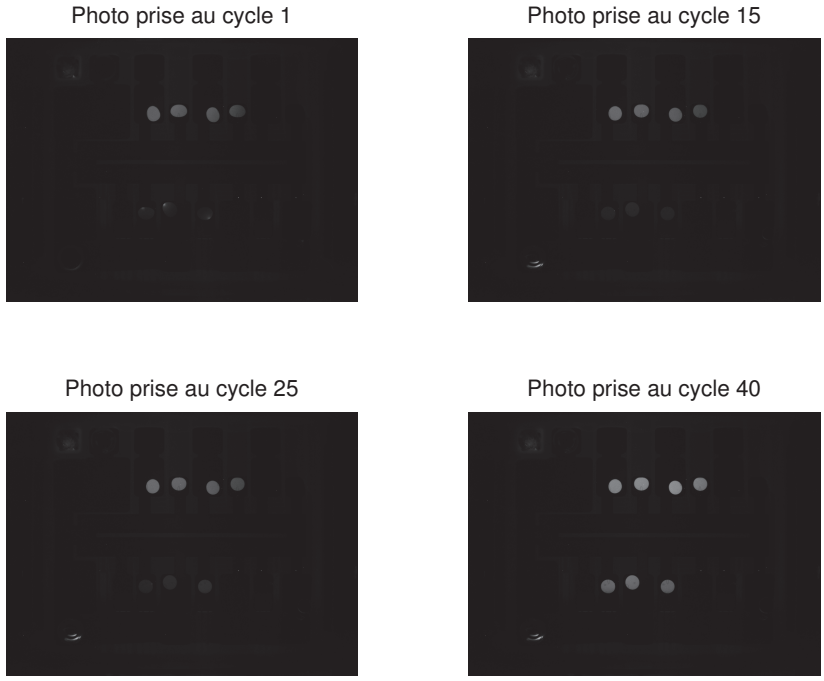


FIGURE 1.8 – Exemple d'une série d'images expérimentales de la puce prise au cycle 1, 15, 25 et 40

Ici les gouttes ont déjà été guidées jusqu'aux puits contenant les réactifs séchés. Un intervalle d'un cycle sépare les images I_n et I_{n+1} ($n \geq 1$). Sur ces images

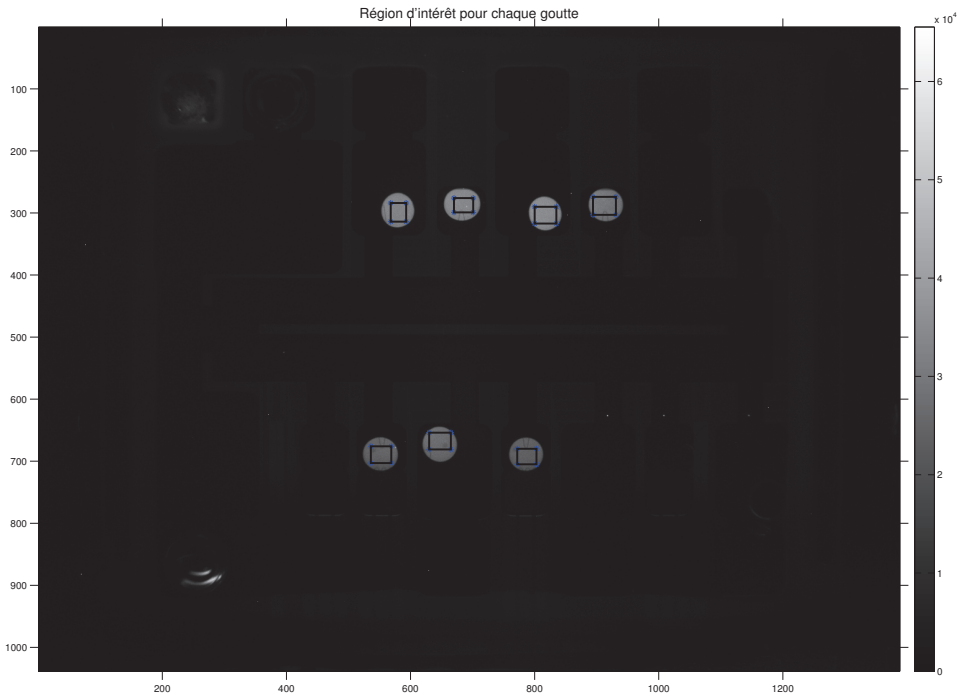


FIGURE 1.9 – Exemple de ROI pour chaque goutte d'une photo prise sur la puce lors d'une expérimentation.

(fig. 1.8), nous observons que l'intensité lumineuse des gouttes augmente au fur du temps. S'il y a la présence de l'espèce recherchée, la luminosité de la goutte marque une montée sur la fluorescence obtenue. Cet instant de montée caractérise la quantité initiale de l'information génétique recherchée.

NB : Le cycle maximum est fixé souvent à 40 cycle. En effet, la qPCR en elle même ne contient que 40 cycles, les cycles supplémentaires servent de contrôle de qualité et appartiennent à un autre test qui est réalisé après les 40 cycles d'amplification. Puis, il y a une prise de vue qui est réalisée à la fin de chaque cycle.

Le signal de fluorescence est obtenu via la puce da la façon suivant :

— fixation d'une *Region Of Interest* (ROI) sur la goutte (fig. 1.9),

— puis, on calcule la moyenne de la fluorescence sur la ROI.

Cependant, il nous semble intéressant d'avoir accès à plus d'information que seulement cette moyenne. Ainsi, d'autres indicateurs statistiques sur la luminosité seraient utiles comme la médiane ou l'écart-type ou les quartiles pour calculer l'instant de rupture. La figure (1.10(a)) représente les indicateurs énumérées précédemment. Notons que l'axe des ordonnés représente la moyenne, la médiane, etc., de

l'intensité lumineuse des pixels de la zone cible à chaque cycle. La médiane est moins sensible au bruit et aux erreurs de mesure (problème de chauffe ou mauvais positionnement de la ROI) que la moyenne.

Nous connaissons la position des électrodes sur la puce grâce à son plan. Les gouttes se déplaçant d'électrodes en électrodes, si on est capable d'avoir la superposition du plan de la puce avec les photos prises lors de la PCR alors nous pouvons déduire la position des gouttes. Pour ce faire, on va utiliser le principe de recalage pour aligner les photos avec le plan de la puce.

Une fois le recalage réalisé, nous ne devons pas explorer l'ensemble d'une image pour la détection des gouttes mais uniquement là où se trouvent les électrodes. On déduit alors la ROI pour chaque goutte et un ou plusieurs indicateurs statistiques sont calculés (par exemple l'écart type, la médiane, les quartiles, etc.). Afin de pouvoir traiter tous ces indicateurs, nous devons savoir où se trouve les gouttes sur notre dispositif de la PCR (puce). Pour cela nous allons appliquer un recalage d'images.

1.2.3 Principe du recalage d'image

1.2.3.1 Approche

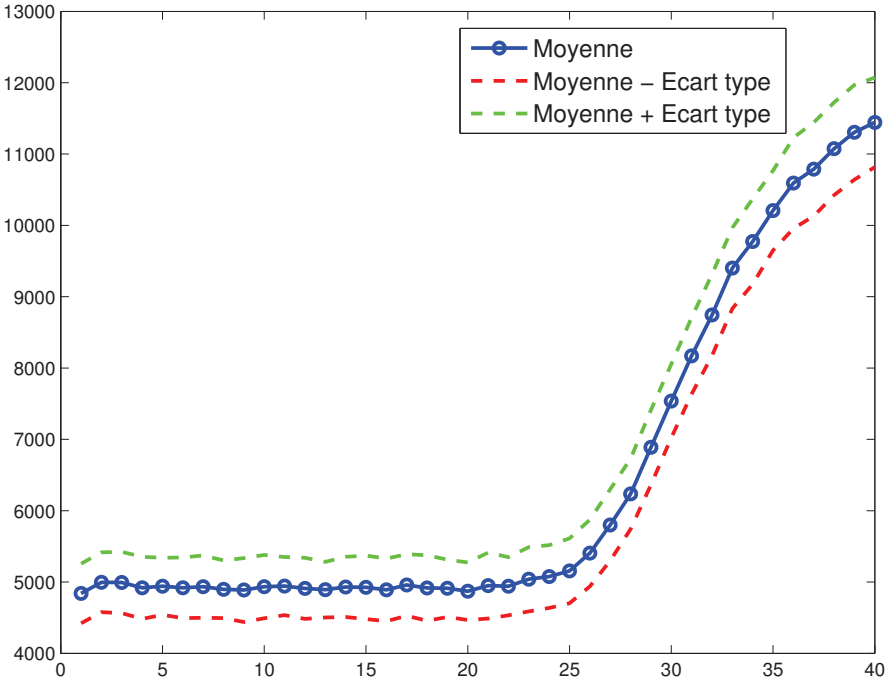
Le recalage d'images consiste à la mise en correspondance de deux différentes images [Reddy and Chatterji, 1996; Wolberg and Zokai, 2000; Sarvaiya et al., 2009]. En disposant de deux ou plusieurs images issues d'une même source mais pouvant être de modalités différentes, le recalage consiste à faire une transformation permettant de passer d'une image source ou flottante à une autre appelée image cible ou référence. Dès lors, un certain nombre de questions se posent : quelles informations utiliser pour guider le recalage ? Comment définir la ressemblance entre les images ? Comment déformer une image ? Comment trouver la meilleure déformation ? Ces questions mènent aux quatre critères caractérisant une méthode de recalage [Brown, 1992; Barillot et al., 1994] :

— les attributs : ce sont les caractéristiques, extraites des images, qui permettent de guider le recalage. On distingue les attributs extrinsèques (par exemple des marqueurs externes fixés à la main) et les attributs intrinsèques (information issue de l'image, comme par exemple les niveaux de gris ou des primitives géométriques extraites) ;

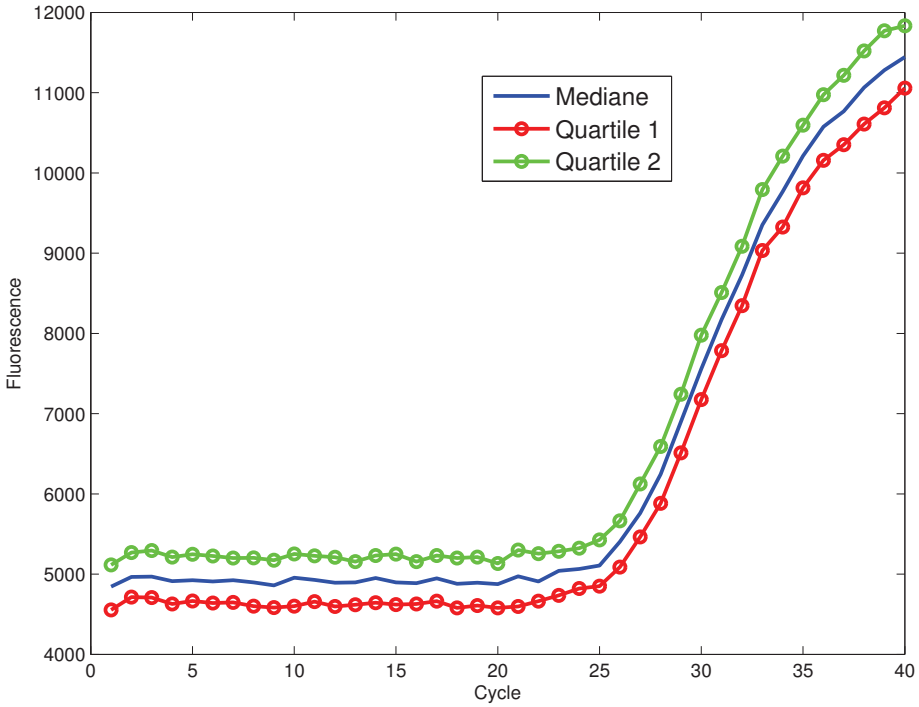
— le critère de similarité : il définit une certaine distance entre les attributs des images afin de quantifier la notion de ressemblance ;

— le modèle de déformation : il conditionne la manière dont l'image est modifiée.

— la stratégie d'optimisation : c'est la méthode qui permet de déterminer la



(a) Moyenne + ou − écart type des zones d'intérêts d'une goutte à chaque cycle.



(b) Médaine et quartiles des zones d'intérêts d'une goutte à chaque cycle.

FIGURE 1.10 – Fluoescences obtenues en calculant la moyenne, la médiane, l'écart type et les quartiles des zones d'intérêts d'une même goutte de la puce (le même résultat est obtenu pour les autres gouttes).

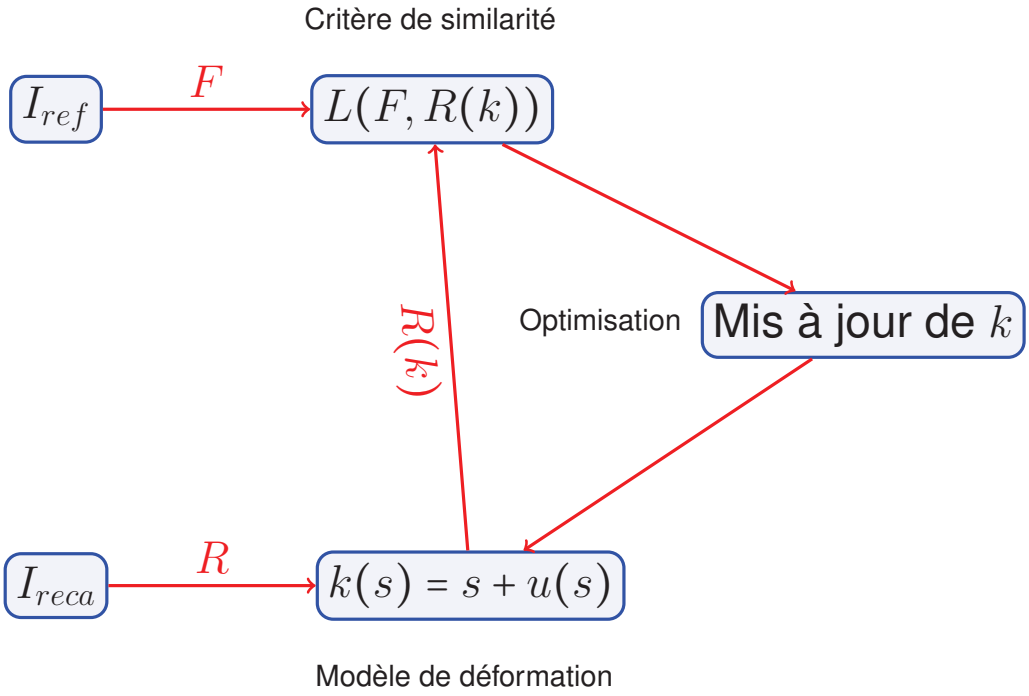


FIGURE 1.11 – Principe général du recalage de l'image à recaler I_{reca} sur l'image référence I_{ref}.

meilleure transformation au sens d'un certain critère de similarité dans l'espace de recherche défini par le modèle de déformation.

Notre objectif n'est pas d'étudier dans les détails le recalage d'images mais de faire une présentation générale et présenter quelques techniques utilisées sur les images dont nous disposons.

1.2.3.2 Techniques générales

Les composantes du système de recalage et leurs interconnexions sont présentées sur la figure (1.11). Les données d'entrée de base pour la procédure de recalage sont deux d'images : une qui définit l'image fixe ou de référence (I_{ref}) et l'autre l'image en mouvement ou à recaler (I_{reca}). Le recalage d'images est traité comme un problème d'optimisation avec le but de trouver la représentation spatiale qui aligne l'image à recaler sur l'image fixe.

Sur la figure (1.11), les images F et R sont construites par extraction des attributs des images I_{ref} et I_{reca}. L'image R est déformée grâce à une transformation k. Un critère de similarité L permet de quantifier la ressemblance entre l'image de référence et l'image déformée. La phase d'optimisation permet de trouver la transformation optimale qui minimise L.

En considérant le recalage d'une image $I_{reca} : \Omega_r \longmapsto \mathbb{R}$ sur une image fixe $I_{ref} : \Omega_f \longmapsto \mathbb{R}$. Les supports Ω_r et Ω_f sont des sous-ensembles de $\mathbb{R}^d$ ($d = 2$ dans le cas d'images $2D$, 3 dans le cas d'images $3D$, etc.). s désigne un point du domaine Ω_f. Le problème de mise en correspondance des deux images consiste à l'estimation d'une transformation $k : \Omega_f \longmapsto \Omega_r$ qui à chaque point de s de l'image fixe associe les coordonnées $k(s) = s + u(s)$ dans l'image à recaler ; u représentant le champ de déformation.

Une étape primordiale est l'extraction d'informations pertinentes (attributs) des images brutes (I_{ref} et I_{reca}) permettant de guider le recalage. F et R sont donc construites respectivement à partir de I_{ref} et I_{reca}. Une fois les informations extraites, il s'agit de définir une fonction L permettant d'associer à F et à $R(k)$ une valeur permettant de quantifier leur proximité ou bien leur ressemblance. Cette fonction L, appelée aussi critère de similarité, devrait théoriquement être minimale (ou maximale selon le critère) lorsque l'image de référence et l'image à recaler sont en parfaite correspondance.

La phase d'optimisation consiste enfin de trouver la transformation optimale $\hat{k}$ qui minimise (ou maximale) la fonction L sur un espace d'ensemble de transformations H. Elle est formulée de la manière suivante :

$$\hat{k} = \arg\min_{k \in H} \; L(F, R(k)). \tag{1.2}$$

Un état de l'art et une étude approfondie en matière de recalage d'images sont présentés dans la thèse de Vincent NOBLET [Noblet, 2006].

1.2.3.3 Méthodes utilisées

Nous nous sommes limités ici aux transformations rigides et nous nous contentons qu'aux déformations suivantes :

— de position c'est-à-dire qu'une des images est décalée par translation affine par rapport à l'autre

— angulaire qui équivaut à un effet de rotation de l'une des images par rapport à l'autre

— d'échelle c'est-à-dire qu'une des images a un effet d'agrandissement (zoom) par rapport à l'autre

L'image peut avoir subi une ou toutes combinaisons de ces déformations.

Pour effectuer le recalage, nous allons utiliser la fonction *imregister* dite *Intensity based image registration*[6] se trouvant dans *Image Processig Toolbox* de

6. Fonction de recalage d'images : transformé une image 2D ou 3D, I, de sorte qu'elle soit recalée sur une image J.

matlab [7].

Elle nécessite une mesure ou métrique, un optimiseur et un type de transformation (fig. 1.12) :

1. La métrique définit la mesure de similarité d'image pour évaluer la précision du recalage. Cette mesure de similarité d'image prend deux images et retourne une valeur scalaire qui décrit la ressemblance de ces deux images.
2. L'optimiseur définit la méthodologie pour minimiser ou maximiser la métrique de similarité.
3. Le type de transformation définit la représentation géométrique (translation, affine, rigide et similarité) que porte le défaut d'alignement de l'image à recaler avec l'image de référence.

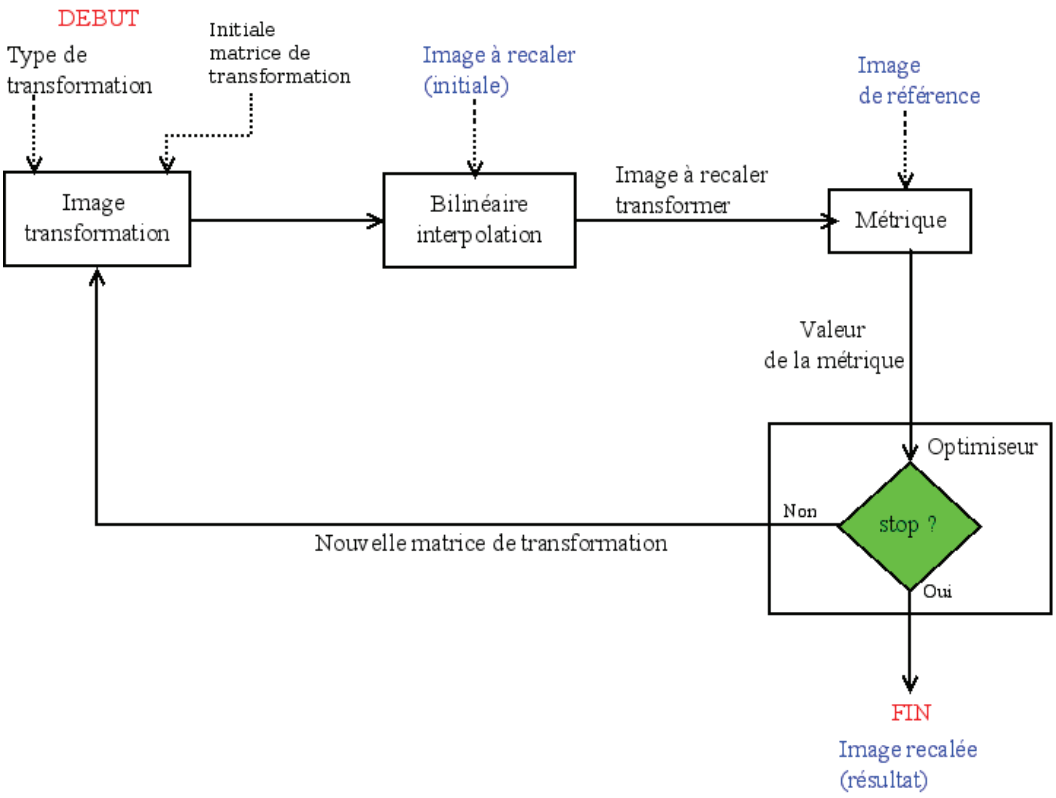


FIGURE 1.12 – Principe du recalage d'image par imregister.

À noter que cette méthode considére toute l'image pour faire le recalage. Comme nous pouvons le voir sur la figure (1.12), le processus commence en définissant un type de transformation et une matrice de transformation déterminée en interne. Ensemble, ils déterminent la transformation d'image spécifique qui est appliquée

7. http ://www.mathworks.fr/fr/help/images/ref/imregister.html

à l'image à recaler avec une interpolation bilinéaire. Ensuite, la métrique compare l'image à recaler transformée de l'image de référence et une valeur de métrique est calculée. Enfin, l'optimiseur vérifie une condition d'arrêt (processus a atteint un nombre d'itérations maximum atteint).

1.2.3.4 Traitement des images

Si le dispositif permettant de prendre les photos lors de la PCR est stable, c'est-à-dire que la caméra permettant de prendre les photos ne bouge pas, alors le recalage se fera uniquement sur la première photo. Dans le cas contraire, par exemple vibration sur la caméra ou déplacement du dispositif en cours d'expérimentation, il faudra faire le recalage à chaque nouvelle image.

En annexe A, nous avons fait un recalage d'images sur un exemple d'image avec : une déformation angulaire, un effet d'agrandissement et un effet de rotation et d'agrandissement.

Applications sur les images de la puce : Voici ci-dessous le plan initial de la puce et une photo prise sur cette dernière lors d'une expérience (fig. 1.13(a) et fig. 1.13(b)). En superposant les deux images (fig. 1.13(c)), nous voyons bien que les gouttes de la photo ne sont pas sur les puits du plan. Un recalage d'images entre les deux est alors nécessaire (fig. 1.14).

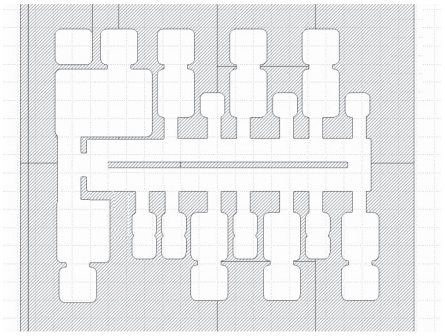
Le plan initial de la puce est une image de 519×694 pixels et les photos de la puce de 530×700 pixels. La photo (image source) a un effet d'agrandissement par rapport au plan de la puce (image cible).

La figure (1.14) montre les résultats du recalage par la méthode *imregister*. Nous constatons que le recalage de la photo sur le plan de la puce est bien réalisé. Le même résultat est obtenu sur les autres photos du même expérience.

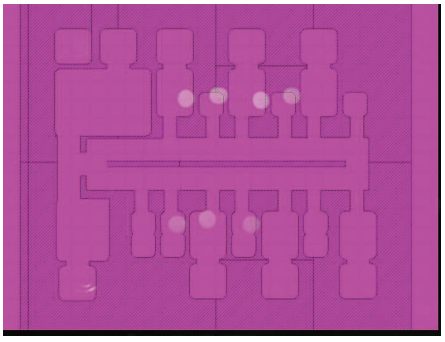
Nous représentons dans ce qui suit d'autres résultats du recalage d'images. Nous disposons de 9 dossiers (données du CEA) contenant chacune des photos prises sur la puce à chaque cycle de la réaction de la PCR. En faisant le recalage d'images entre une photo et le plan de la puce, nous obtenons les résultats suivants sur 3 dossiers (fig. 1.15, fig. 1.16, fig. 1.17).

Comme précédemment, le recalage d'images des photos sur l'image de la puce est bien réalisé. Néanmoins, certaines gouttes peuvent ne pas être dans leurs puits (fig. 1.16). Ainsi, elles ne se placent pas sur la position des électrodes correspondantes. La construction de leurs zones d'intérêt se fera alors à la main.

(a) Photo



(b) Plan



(c) Superposition des deux images avant recalage

FIGURE 1.13 – Présentation des images avant le recalage.

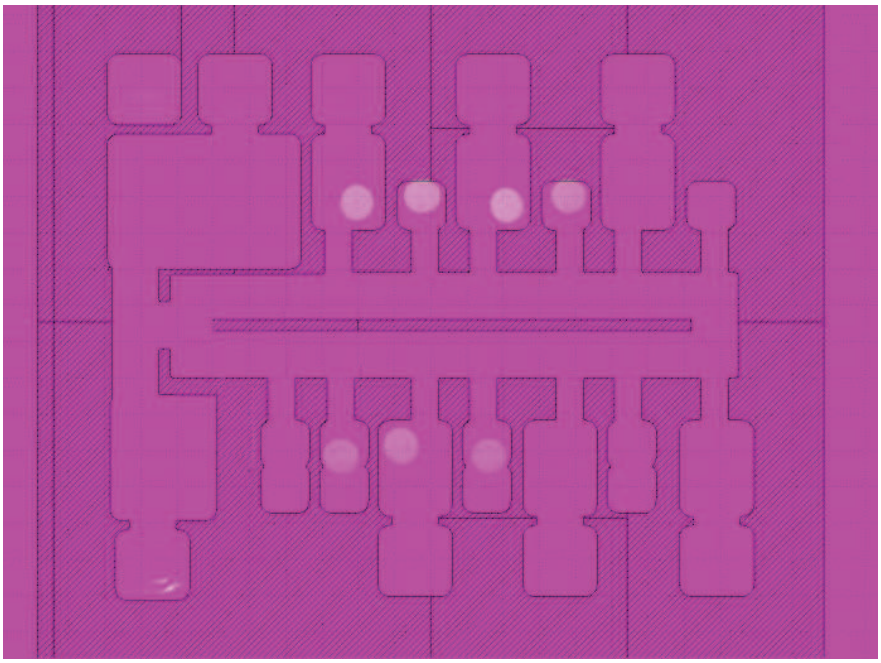


FIGURE 1.14 – Résultat du recalage de l'image source (photo) par rapport à l'image cible (plan puce)

1.3 Conclusion

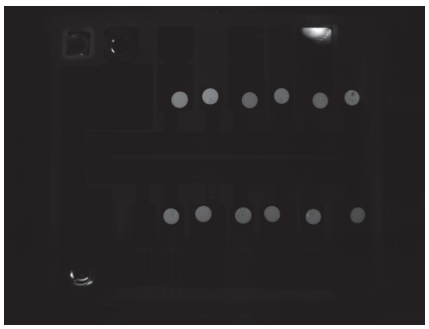
Nous avons présenté la réaction en chaine par polymérase (PCR) qui permet d'obtenir plusieurs molécules d'ADN à partir d'une seule séquence particulière d'ADN. Différents types de PCR existent notamment la PCR en temps réel ou quantitative (qPCR) qui est utilisée pour mesurer la quantité initiale Q_0 de l'ADN.

Grâce à un marqueur fluorescent, émettant de la lumière, qui réagit à l'ADN à chaque cycle d'amplification, un signal de fluorescence est obtenu après un suivi d'une PCR en temps réel. La fluorescence est composée de 3 phases (initiations, exponentielles et plateau) et caractérisée par un cycle repère c_r qui est l'instant du début de la phase exponentielle. À partir de ce cycle repère, nous pouvons estimer la quantité initiale d'ADN grâce à la relation log-linéaire entre Q_0 et c_r.

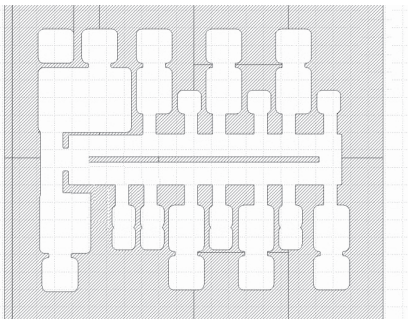
Nous avons présenté le signal de fluorescence issu d'une puce microfluidique du CEA-LETI. Puis, nous avons utilisé la technique de recalage d'image basée sur l'intensité des pixels pour obtenir les fluorescences en utilisant d'autres indicateurs que la moyenne de la goutte (médiane, écart type et les quartiles). Les images à recaler sont des photos prises sur la puce à chaque cycle et l'image de référence est

le plan de la puce. De bons résultats sur le recalage sont obtenus en l'appliquant sur les images.

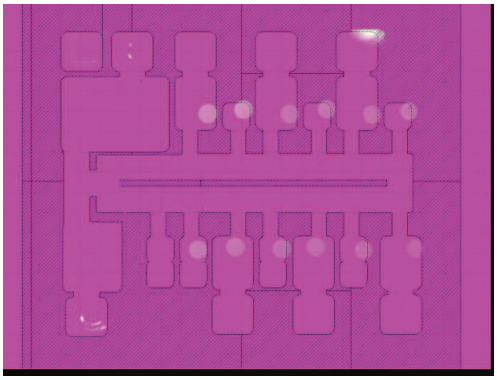
Dans le chapitre suivant, nous allons présenter différentes types de méthodes pour calculer un cycle repère de la fluorescence.



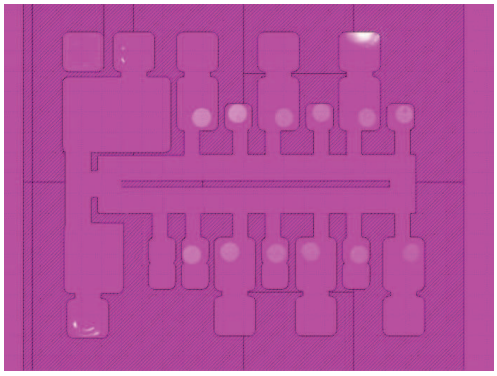
(a) Photo



(b) Plan

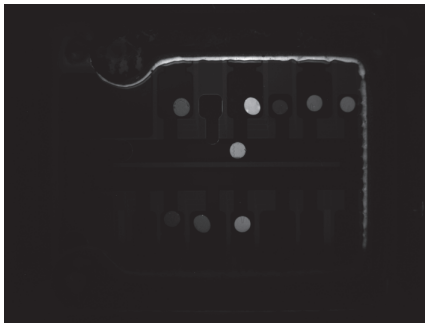


(c) Superposition des deux images avant recalage

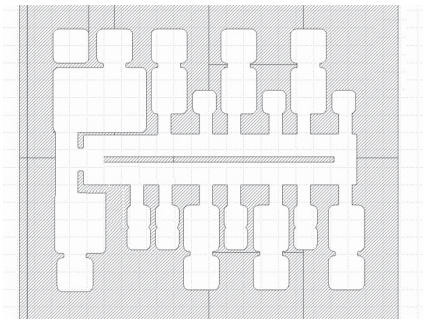


(d) Superposition des deux images après recalage

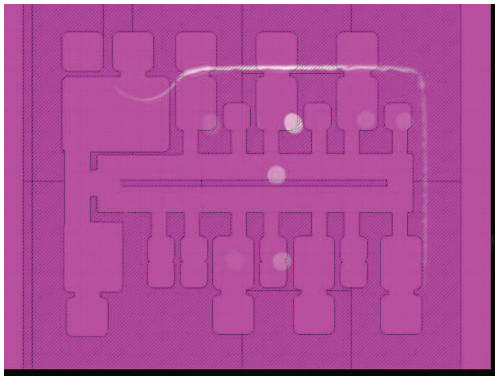
FIGURE 1.15 – Résultat du recalage d'images sur un exemple de photo.



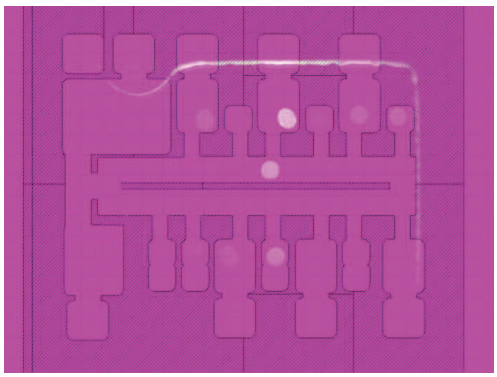
(a) Photo



(b) Plan

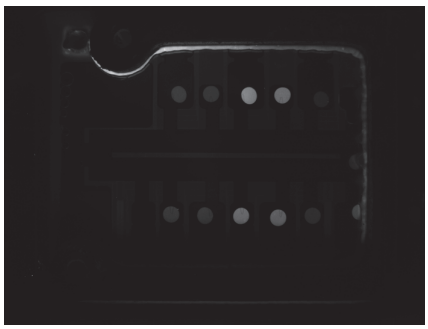


(c) Superposition des deux images avant recalage

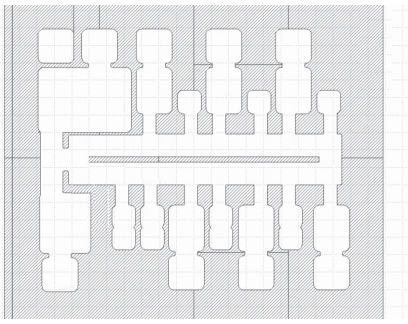


(d) Superposition des deux images après recalage

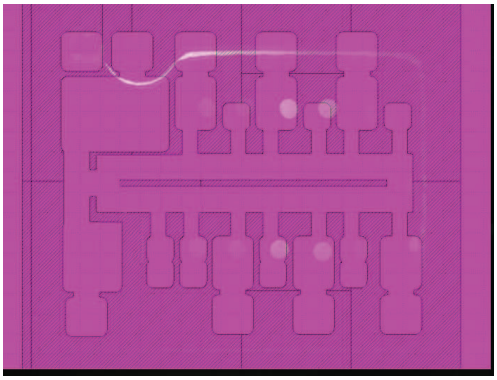
FIGURE 1.16 – Résultat du recalage d'images un exemple de photo.



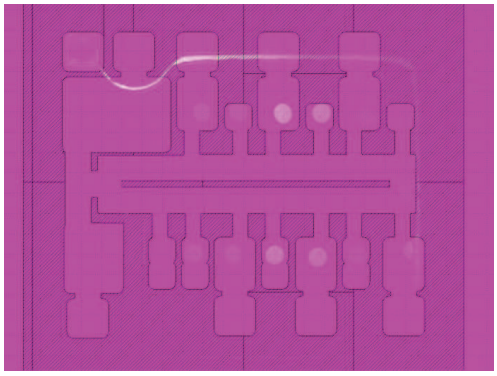
(a) Photo



(b) Plan



(c) Superposition des deux images avant recalage



(d) Superposition des deux images après recalage

FIGURE 1.17 – Résultat du recalage d'images un exemple de photo.

CHAPITRE 2

Modèles et techniques de détermination d'un cycle repère

Sommaire

Dans le chapitre précédent, nous avons présenté la fluorescence issue d'une réaction en chaine par polymérase. Une fluorescence est caractérisée par un cycle repère qui nous indique la présence d'un agent biologique recherché. À l'inverse, s'il n'y a pas de cycle de repère, alors il n'y a pas de présence de l'espèce recherché. Dans ce chapitre, nous allons alors étudier différentes méthodes de l'état de l'art qui calculent un cycle repère de la fluorescence. Nous présenterons leurs avantages et leurs inconvéinents.

2.1 Modèle théorique et estimation de la quantité de molécules cibles

La quantité de produit formé dépend d'un facteur d'efficacité d'amplification (E). Ce facteur est défini comme étant la proportion moyenne des molécules d'ADN cible se dupliquant à chaque cycle d'amplification. En réalité, l'effet plateau reflète une baisse de l'efficacité E d'amplification. Elle résulte en partie de l'inactivation thermique de l'ADN polymérase au cours des derniers cycles du fait que les produits nécessaires à l'amplification sont limités. L'efficacité d'amplification est comprise entre 0 (aucune amplification ne s'est produite) et 1 (après chaque cycle PCR, chaque molécule d'ADN cible a généré deux amplicons, voir fig. 1.4).

2.1.1 Modèle théorique

L'introduction de ce facteur permet de déduire un modèle représentant au mieux la phase exponentielle par l'équation suivante [Peirson et al., 2003] :

$$Q_c = Q_0(1 + E_c)^c \tag{2.1}$$

où Q_0 représente la quantités de molécules initiale, E_c l'efficacité d'amplification u cycle c et Q_c la quantité de molécules amplifiée au cycle c.

Par convention E_c est inférieur à 1 et varie entre 0.78 et 0.97 selon le gène amplifié.

Si on considère que E_c est constant quelque soit le cycle c alors on pose $E := E_c$. En notant c_r un cycle repère, l'équation (2.1) devient,

$$Q_{c_r} = Q_0(1 + E)^{c_r} \tag{2.2}$$

En appliquant le $\log$ à gauche et à droite, nous obtenons

$$\log Q_{c_r} = \log(Q_0(1 + E)^{c_r}) \tag{2.3}$$
$$\log Q_{c_r} = \log Q_0 + c_r \log(1 + E) \tag{2.4}$$
$$\log Q_0 = -\log(1 + E).c_r + \log Q_{c_r} \tag{2.5}$$

où Q_{c_r} est le nombre de molécules amplifiées au cycle repère c_r.

Nous retrouvons une relation $\log$-linéaire entre la quantité de molécules cibles Q_0 et le cycle repère c_r [Rutledge, 2004; Pfaffl, 2001]. De ce fait si on connait au moins 2 cycles repères de 2 courbes avec leurs quantités initiales, on peut retrouver le Q_0 correspondant au c_r d'une nouvelle courbe réalisée dans les mêmes conditions. La détermination de la quantité de molécules cibles Q_0 sera alors faite à l'aide d'une gamme de fluorescence (voir ci-dessous).

2.1.2 Estimation de la quantité de molécules initiales

Pour estimer la quantité cible d'une molécule d'ADN, on estime les paramètres inconnus E et Q_{c_r} du modèle (2.5) par régression linéaire. En effet, nous disposons d'une gamme de fluorescences brutes pour lesquelles on connait leurs quantités de molécules cibles Q_0 et leurs cycles repères associés. Puis, on déduit pour une nouvelle expérience et sous les mêmes conditions de la quantification par PCR, la quantité de molécules initiale. Ceci est illustré par l'équation (2.6) et la figure (2.1).

$$\log Q_0 = m\ c_r + p \tag{2.6}$$

$$Q_0 = exp(m\ c_r + p) \tag{2.7}$$

où m et p sont respectivement les estimations de $-\log(E+1)$ et $\log(Q_{c_r})$.

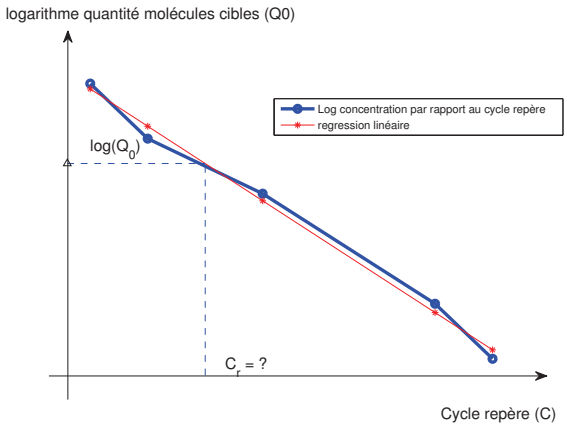


FIGURE 2.1 – Illustration du log linéarité de la quantité d'ADN initiale par rapport aux cycles repères sur la gamme d'étalonnage de la figure (1.5).

Il n'existe pas de définitions uniques du c_r et ce dernier n'est pas forcément un entier. Les méthodes de détermination de ce dernier ne font pas toutes références au même point sur la courbe. Ce qui est important, est qu'une fois une méthode choisie, on reste constant (cohérent) de la création de la gamme à la mesure. Dans ce qui suit, nous présentons différents types de méthodes pour la détermination d'un cycle repère avec leurs avantages et leurs inconvénients.

2.2 Comment déterminer un cycle repère ?

L'étape primordiale lors de la quantification d'une PCR est de déterminer un cycle repère défini de façon identique pour chaque courbe traitée afin d'estimer la

concentration ou la quantité cible de la molécule d'ADN, cf équation (2.6). Nous présenterons deux familles de méthodes déterminant le cycle repère. Une première famille basée sur la géométrie de la courbe comprenant la *threshold method*, le *Crossing point* et la *fit point method*. Une deuxième famille, appelée modèles globaux, basée sur la modélisation de la fluorescence : *sigmoïde curve method* et C_{y_0} *method*. Ces différentes méthodes seront présentées avec leurs avantages, leurs inconvénients et leurs comportements sur diverses fluorescences. À noter que le cycle c_r peut être dans un espace continu même si la fluorescence est discrète (intensité d'amplification de la fluorescence mesurée une fois par cycle).

Comme illustration, ces différents types de méthodes seront appliqués sur une même fluorescence, représentée sur la figure (2.2).

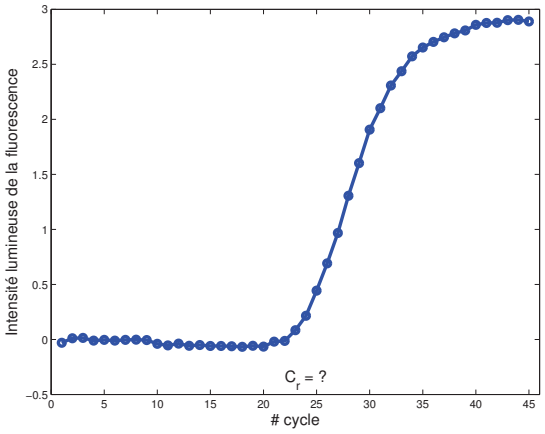
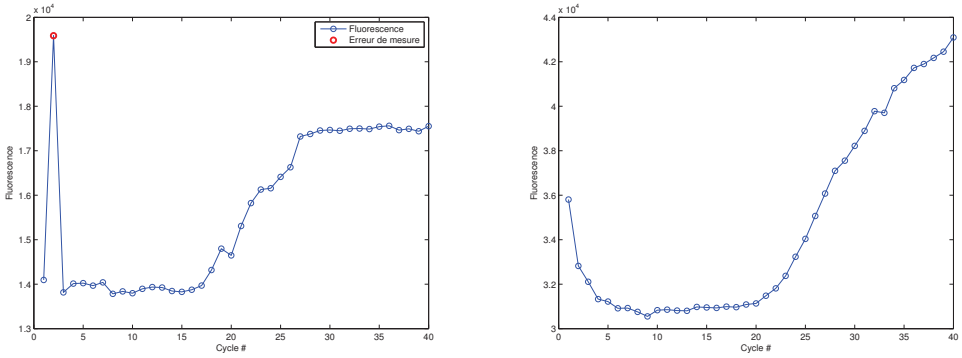


FIGURE 2.2 – Signal de fluorescence sur lequel sont appliquées les différentes techniques énumérées ci-dessus pour déterminer c_r.

Définitions : Avant de détailler les méthodes ci-dessus, nous allons définir quelques termes.

1. On appelle *détection en ligne* d'un cycle repère, la détermination en temps réel de ce cycle repère sans la connaissance du comportement complet de la fluorescence.
2. Inversement, on appelle *détection hors ligne*, la détermination d'un cycle repère avec la connaissance du comportement complet de la fluorescence.
3. On dit qu'il y a une *fausse détection*, lorsqu'un cycle repère est déterminé sachant qu'il n'y en a pas.
4. On dit qu'il y a *non détection*, lorsqu'un cycle repère n'est pas déterminé sachant qu'il y en a.



(a) Avec une erreur de mesure (b) Avec une décroissance au début

FIGURE 2.3 – Fluorescences non standards observées.

5. Le *retard à la détection* δ d'un cycle repère c_r est le décalage de cycle entre c_r et le nombres de cycles supplémentaires qu'il faut pour le détecter.

Sur les fluorescences expérimentales, nous pouvons observer deux problématiques : une présence d'erreurs de mesures (fig. 2.3(a)) et/ou une décroissance au début de la fluorescence (fig. 2.3(b)). Les erreurs de mesures sont des vraies mesures et peuvent être causées par un problème de chauffe (température) lors de la PCR. Ainsi, pour chaque méthode, nous allons illustrer son comportement sur ces deux problématiques (fig. 2.3).

2.2.1 Modèles géométriques

Les méthodes du modèle géométrique déterminent le point repère en étudiant les propriétés géométriques de la courbe de fluorescence.

2.2.1.1 Méthode par seuillage

La méthode par seuillage (*threshold method* en anglais) est une technique définissant le cycle repère comme étant le nombre de cycles correspondant à l'intersection entre la courbe de la fluorescence et une ligne de seuil établie *à priori* (fig. 2.4) [Kubista et al., 2006]. La ligne de seuil est réglable et placée au dessus de l'intensité de fluorescence des premières cycles de la qPCR. Elle correspond au seuil de détection optique au delà duquel les observations sont dans la phase exponentielle. Pour la figure (2.4), nous obtenons un cycle repère c_r égale à $24,3$ et un retard à la détection δ nul (un cycle repère est déterminé dès que la ligne coupe la fluorescence).

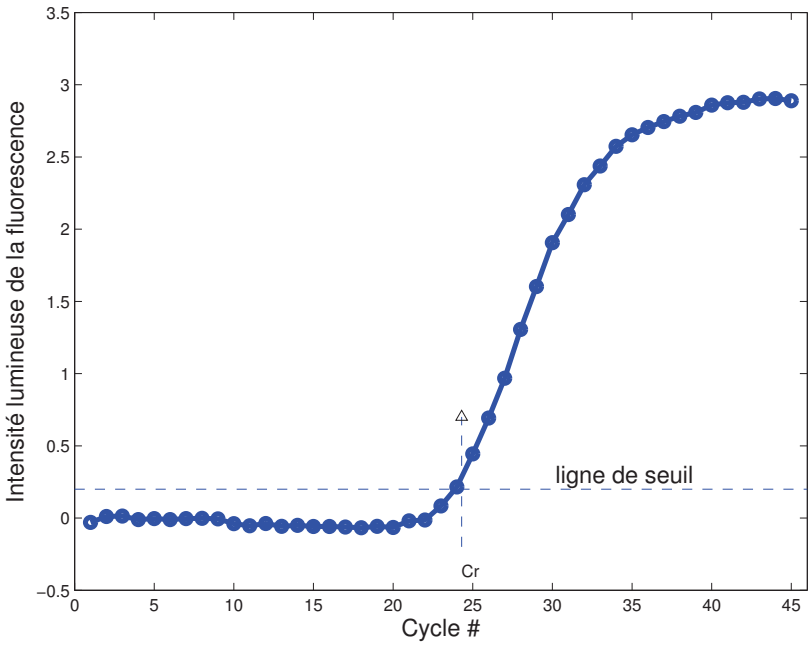
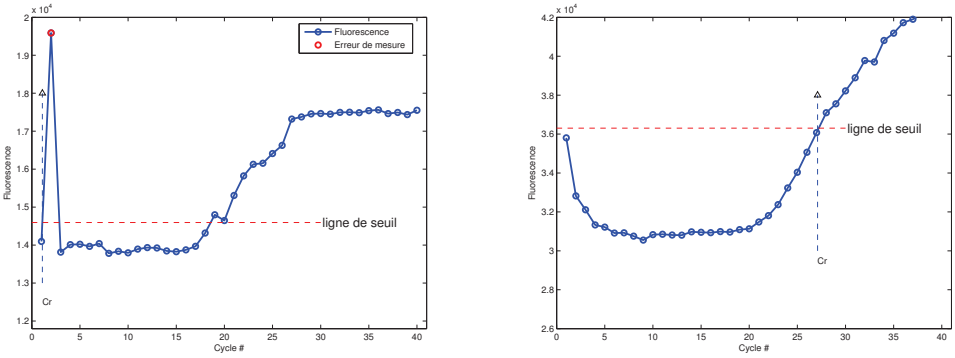


FIGURE 2.4 – Threshold method, $c_r = 24,3$.

Nous notons que pour appliquer cette méthode nécessite l'hypothèse selon laquelle l'efficacité E doit être identique dans tous les échantillons d'une même expérience pour pouvoir fixer la ligne de seuil. En effet, plusieurs fluorescences sont obtenues sur une même gamme d'expérience et alors une seule ligne de seuil est fixée pour déterminer les cycles repères correspondants [Durtschi et al., 2007]. La ligne de seuil est fixée par l'utilisateur et elle varie d'une gamme d'expérience à une autre.

Avantages : La détermination d'un cycle repère par la méthode du seuil peut se faire en temps réel. Aussi, étant donné que la ligne de seuil est placée au dessus des premières observations, cette méthode détermine un c_r pour des fluorescences avec une décroissance au début de la PCR (fig. 2.5(b)).

Inconvénients : Nous pouvons avoir des fausses alarmes sur des fluorescences où on observe des montées avant le cycle repère ou avec des erreurs de mesures comme nous pouvons l'observer sur la figure (2.5(a)). Dans ce cas la fixation d'une ligne de seuil n'est pas appropriée. Il sera alors nécessaire de faire un pré traitement pour supprimer les erreurs de mesures qui apparaissent sur la fluorescence (voir chapitre 2).



(a) Méthode du seuil sur une fluorescence avec présence d'erreur de mesure.

(b) Méthode du seuil sur une fluorescence avec décroissance au début.

FIGURE 2.5 – Comportement de la méthode du seuil sur une fluorescence avec une erreur de mesure et une avec une décroissance au début.

2.2.1.2 Maximum de la dérivée seconde

La méthode du maximum de la dérivée seconde (*crossing point*) est définie comme étant le cycle correspondant au maximum de la dérivée seconde de la fluorescence [Durtschi et al., 2007]. Nous ne connaissons pas *à priori* la dérivée seconde et nous avons une courbe de fluorescence discrète. Nous allons utiliser les techniques de différence finie basées sur les formules de Taylor pour estimer la dérivée seconde de la fluorescence [Euvrard, 1994].

En appliquant cette méthode sur la fluorescence (2.2), nous obtenons le résultat schématisé sur la figure (2.6). Nous avons un cycle repère $c_r = 22$ et un retard à la détection $\delta = 45 - c_r = 23$. C'est une méthode très utilisée dans le cas hors ligne.

Avantages : Elle a pour avantage de ne pas présupposer l'hypothèse d'égalité de E.

Inconvénients : La méthode du maximum de la dérivée seconde ne s'applique pas en temps réel car le maximum n'est certain que lorsque tout le signal est connu. En plus, la méthode montre ces limites pour des fluorescences avec la présence d'erreurs de mesures ou avec une décroissance au début de la quantification (voir fig. 2.7).

2.2.1.3 Modélisation des observations par la méthode des points ajustés

La méthode des points ajustés (*fit point method*) consiste à construire une régression linéaire au $\log$ de la phase exponentielle de la fluorescence. Le cycle ca-

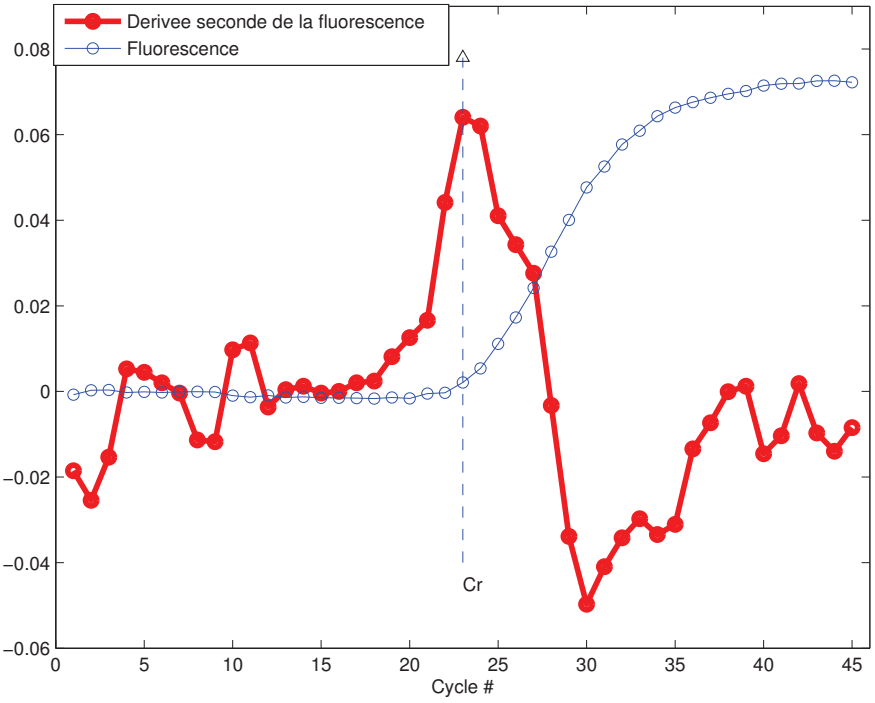
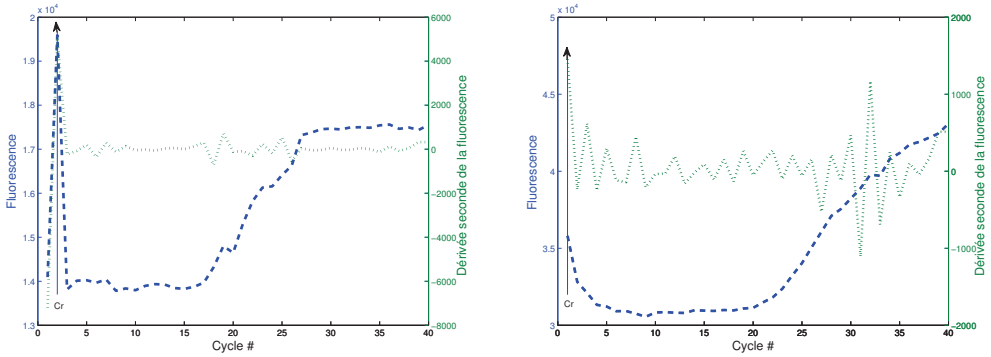


FIGURE 2.6 – Méthode du maximum de la dérivée seconde, c_r = 22.



(a) Maximum de la dérivée seconde sur une fluorescence avec présence d'erreur de mesure.

(b) Maximum de la dérivée seconde sur une fluorescence avec décroissance au début.

FIGURE 2.7 – Comportement de la méthode du *crossing point* sur une fluorescence avec une erreur de mesure et une avec une décroissance au début.

ractéristique c_r, par la *fit point method*, correspond à l'intersection entre la droite de régression et une ligne de seuil établie par l'utilisateur comme indiqué sur la figure (2.8).

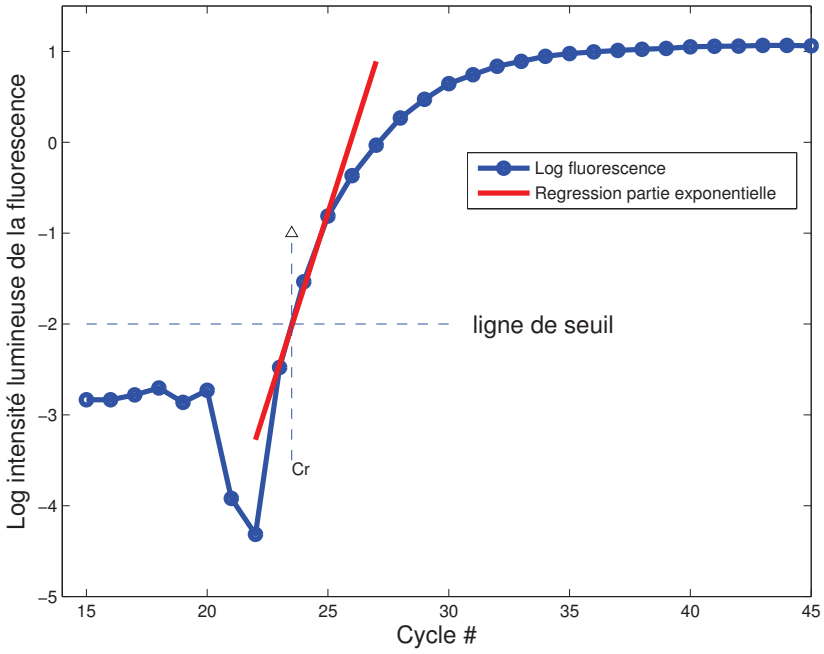


FIGURE 2.8 – Exemple illustratif du Fit point method, nous trouvons un cycle repère $c_r = 23,5$.

Sur cette dernière, en appliquant la méthode des points ajustés, nous trouvons un cycle repère $c_r = 23,5$ et un retard à la détection $\delta = 21,5$ cycles.

Comme pour la méthode du seuil, la fixation d'une ligne de seuil nécessite l'hypothèse selon laquelle l'efficacité d'amplification E est la même dans tous les échantillons d'une même d'expérience [Durtschi et al., 2007].

Nous allons donner le résultat sur la régression linéaire par la méthode des moindres carrés et le détails du calcul sont fournis en annexe B.

— **Régression linéaire par la méthode des moindres carrés :** Soient $(z_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, n$; où n est la taille des données. On cherche une fonction f telle que $y = f(z)$. Dans ce cas, le modèle linéaire s'écrit :

$$y_i = a_0 + a_1 z_i + \varepsilon_i \quad \forall\ i = 1, \ldots, n \tag{2.8}$$

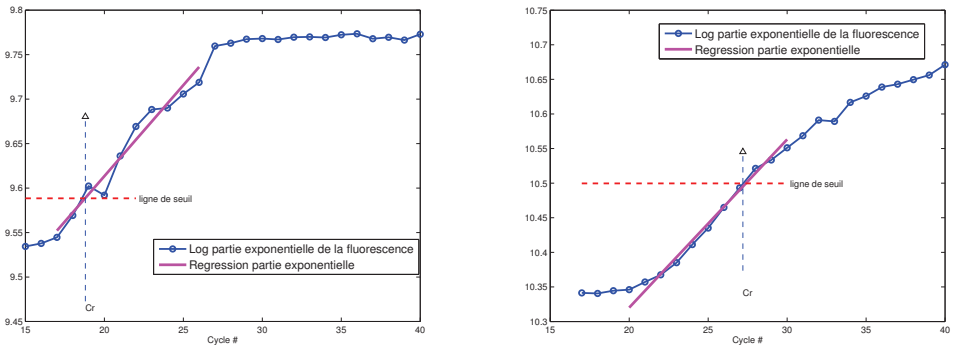
où $\varepsilon = (\varepsilon_i)_{i=1,\ldots,n}$ est un bruit gaussien centré ; a_0 et a_1 sont des paramètres réels inconnus.

En posant $a^t = (a_0, a_1)$, l'estimateur des moindres $\hat{a}_{MC}$ de a donne le résultat suivant,

$$\hat{a}_{MC} = (Z^t Z)^{-1} Z^t y \tag{2.9}$$

où $Z = \begin{pmatrix} 1 & z_1 \\ \vdots & \vdots \\ 1 & z_n \end{pmatrix}$ et $y^t = (y_1, \ldots, y_n)$.

Avantages : Avec comme objectif la détection hors ligne du cycle repère, la méthode des points ajustés peut être utilisée. L'avantage de cette méthode est l'utilisation de l'ensemble du segment exponentiel pour déterminer le cycle caractéristique. À noter aussi, que cette méthode marche sur des fluorescences avec une décroissance au début ou avec une erreur de mesure (fig. 2.9) car elle n'utilise que la partie exponentielle.



(a) *Fit point method* sur une fluorescence avec présence d'erreur de mesure.

(b) *Fit point method* sur une fluorescence avec décroissance au début.

FIGURE 2.9 – Comportement de la méthode du *Fit point method* sur une fluorescence avec une erreur de mesure et une avec une décroissance au début.

Inconvénients : Si on veut déterminer le cycle repère en temps réel tout en minimisant le retard à la détection, cette méthode ne sera pas adéquate. En effet, il faudra plusieurs points d'amplification de la fluorescence pour identifier sa partie exponentielle, construire sa courbe logarithmique et y faire une régression.

2.2.2 Modèles globaux

La détermination du cycle repère c_r par des modèles globaux suit une démarche identique :

1. Nous formulons d'abord un modèle théorique du signal de fluorescence $f(c, \beta)$ fonction du numéro de cycle c et du paramètre β.
2. À partir des observations disponibles, nous estimons les paramètres β du modèle.
3. Enfin, des paramètre β nous en déduisons le cycle repère c_r.

De part la forme en S de la fluorescence, nous déduisons qu'elle appartient à la famille des fonctions sigmoïdes. La fonction sigmoïde (dite aussi courbe en S) est définie par :

$$f : c \longmapsto f(c, \beta) = \frac{1}{1 + e^{-\beta c}}$$

avec $\beta > 0$ représentant la pente de la partie exponentielle de f.

On remarque que quand c tends vers $+\infty$, cette fonction f prend une valeur maximale égale à 1 et lorsque c tends vers $-\infty$, f est nulle.

Parmi ces modèles, nous avons les modèles

— Sigmoide (SCM),

— Richards (C_{y_0} method),

— Gompertz,

— Hill et

— Chapman.

Nous allons présenter le modèle Sigmoide et le modèle de Richards parce qu'ils donnent une meilleure approximation de la fluorescence que les autres modèles en se basant sur les travaux de Guescini et al. [2008].

2.2.2.1 Méthode d'ajustement de la courbe sigmoïde

La méthode d'ajustement de la courbe sigmoïde ou *Sigmoïd curve fitting method* (SCM) décrite par Rutledge [2004] utilise le modèle (2.10) suivant :

$$F_c = F_b + \frac{F_{\max}}{1 + e^{-\frac{c - c_{1/2}}{k}}} + \varepsilon_c, \forall\ c \tag{2.10}$$

où

— F_c est la fluorescence issue d'une qPCR.

— $F_{\max}$ est la valeur maximale de la fluorescence.

— F_b est le fond de la réaction de fluorescence (*background fluorescence*).

— c est le numéro de cycle.

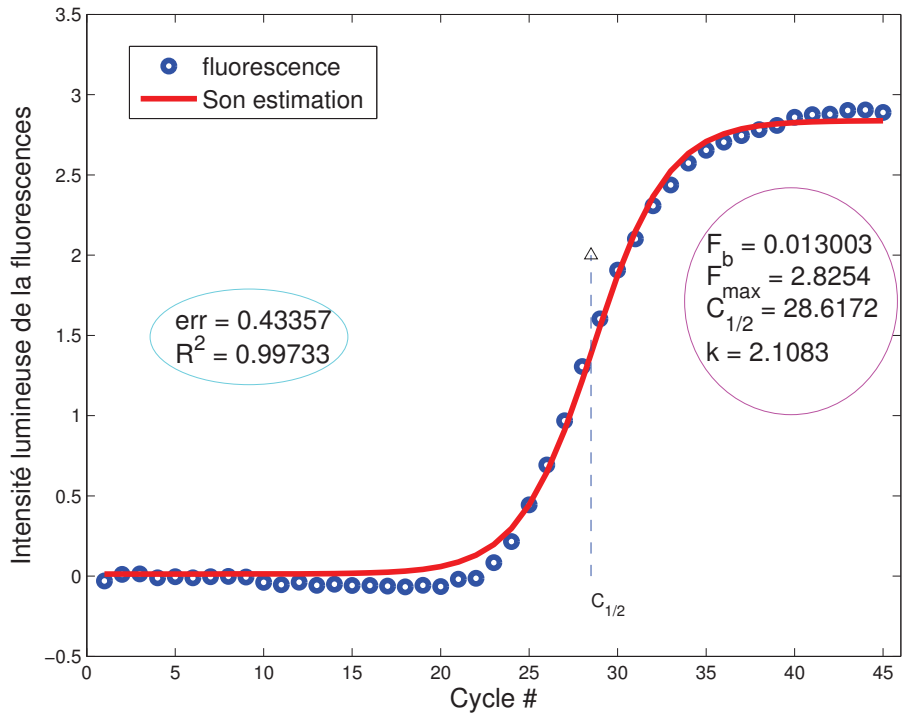


FIGURE 2.10 – Illustration de la méthode *sigmoïdal curve fit method* sur un exemple de fluorescence.

— $c_{1/2}$ est le numéro de cycle du point d'inflexion de la courbe sigmoïde. Le point d'inflexion est un point où on observe un changement de concavité d'une courbe. En mathématiques, c'est le point qui annule la dérivée seconde, si elle existe, de la courbe.

— k est la pente de la partie exponentielle de la courbe.

— $\varepsilon = \{\varepsilon_c\}_{c=1,2,...}$ est un bruit.

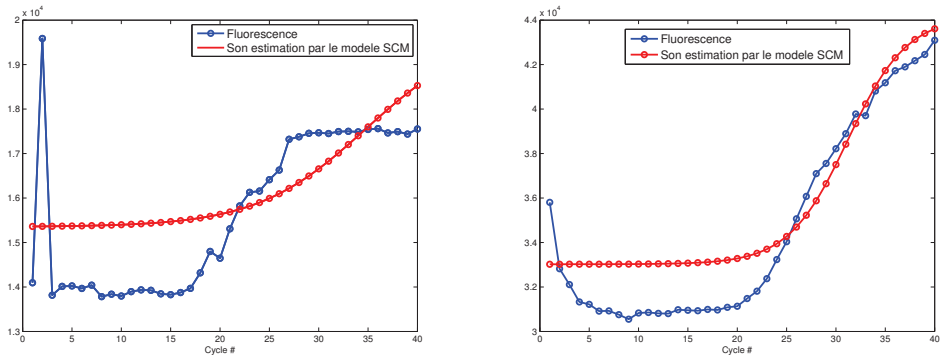
La *sigmoïd curve fitting method* définit le cycle repère comme étant égal au paramètre $c_{1/2}$ [Durtschi et al., 2007].

Les paramètres inconnus du modèle (2.10) seront estimés en minimisant l'erreur quadratique de ε par la méthode de descente du gradient conjugué. La figure (2.10) illustre l'estimation des paramètres du modèle SCM sur l'exemple de la fluorescence (2.2). Nous obtenons, sur cet exemple, un coefficient de corrélation R^2 proche de 1 et une faible erreur d'estimation de la fluorescence. Le modèle (2.10) représente alors bien ce signal de fluorescence. Nous obtenons un cycle repère égale à $28,62$ cycles et un retard à la détection égal à 16.38 cycles

(= 40 – 28, 62).

Avantages : L'avantage de cette méthode est l'utilisation complète de la fluorescence donc l'identification d'un cycle repère est précis. On note aussi que cette méthode ne nécessite pas l'hypothèse d'égalité de E.

Inconvénients : La méthode SCM n'est pas applicable en ligne car l'estimation des paramètres du modèle nécessite une connaissance complète de la fluorescence. Comme nous pouvons le voir sur la figure (2.11), cette méthode représente mal la fluorescence qui a une erreur de mesure ou une forme de décroissance au début de la quantification.



(a) *Sigmoïde curve fitting method* sur une fluorescence avec présence d'erreur de mesure.

(b) *Sigmoïde cure fitting method* sur une fluorescence avec décroissance au début.

FIGURE 2.11 – Comportement de la méthode du *Fit point method* sur une fluorescence avec une erreur de mesure et avec une décroissance au début.

2.2.2.2 Modèle de Richards (C_{y_0} method)

La méthode C_{y_0} est une généralisation de la méthode sigmoïde, elle considère la fonction de Richards [Guescini et al., 2008; Rutledge, 2004] qui modélise la courbe de fluorescence par,

$$F_c = F_b + \frac{F\text{max}}{\left(1 + e^{-\frac{c - c_{1/2}}{k}}\right)^d} + \varepsilon_c, \forall\ c \tag{2.11}$$

où d est le coefficient de Richards ($d > 0$). Pour $d = 1$, on obtient le modèle Sigmoide précédent.

Comme précédemment, on estime les paramètres $(F_b, F\text{max}, c_{1/2}, k, d)$ par la méthode du gradient conjugué. Le cycle repère $c_r = C_{y_0}$ est l'intersection entre

l'axe des abscisses et la tangente du modèle au point d'inflexion $c_{1/2}$ de la courbe de Richard. Le cycle repère C_{y_0}, est égal à (voir annexe pour le détails des calculs) :

$$C_{y_0} = \frac{-F_b - F\max\left(\frac{d}{d+1}\right)^d \left(1 - \frac{1}{k}\left(\frac{d}{d+1}\right)(k\log d + c_{1/2})\right)}{\frac{F\max}{k}\left(\frac{d}{d+1}\right)^{d+1}} \tag{2.12}$$

En appliquant cette méthode sur la fluorescence (2.2), nous obtenons les résultats schématisés sur la figure (2.12).

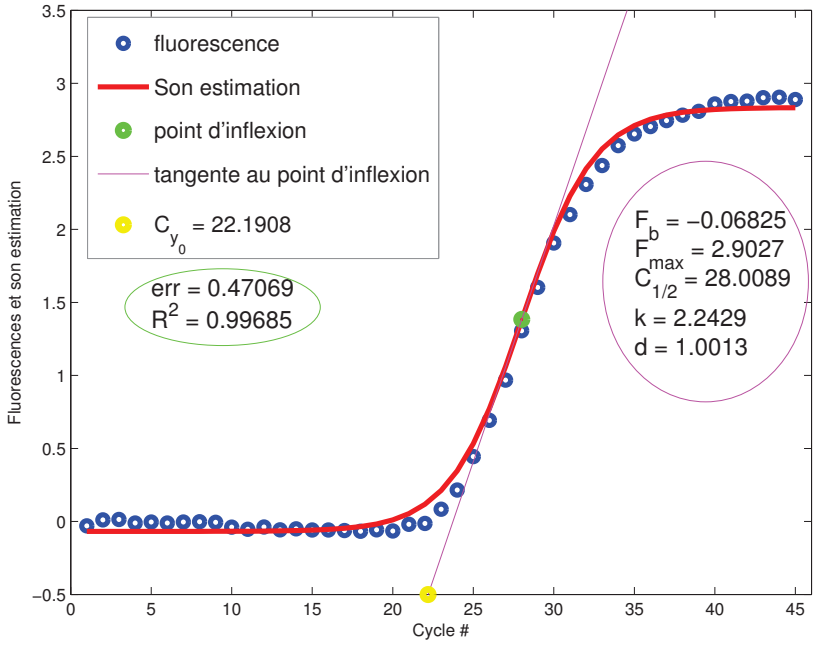
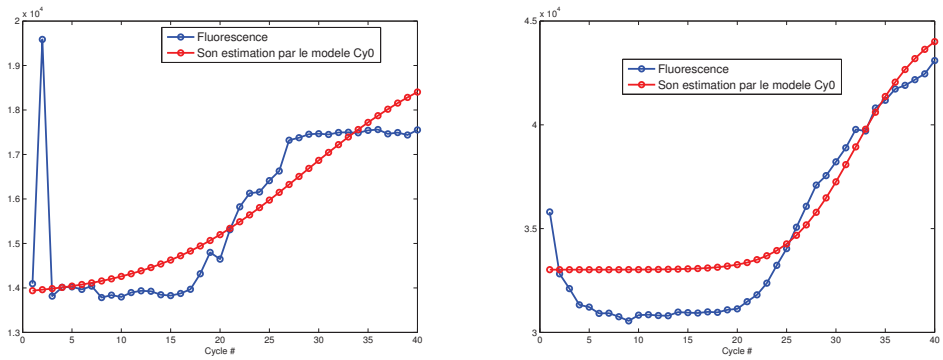


FIGURE 2.12 – Estimation de la fluorescence par le modèle de Richards, $c_r = C_{y_0} = 22.19$.

Comme pour la méthode SCM, la fonction de Richards représente bien ce signal de fluorescence issu de la qPCR. En effet, nous obtenons un coefficient de corrélation proche de 1 et une faible erreur d'estimation de la fluorescence. La méthode C_{y_0} détermine un cycle caractéristique c_r égale à 22.19 et un retard à la détection de 17.81 cycle.

Avantages : Comme pour la méthode SCM, la méthode C_{y_0} ne nécessite pas l'hypothèse d'égalité de l'efficacité d'amplification E. Aussi, l'estimation d'un cycle repère est faite en considérant toute la fluorescence, d'où la précision du cycle repère.

Inconvénients : L'application de la méthode C_{y_0} nécessite la connaissance complète de la fluorescence pour bien estimer les paramètres de leurs modèles. Ainsi elle n'est pas efficace pour l'appliquer en ligne. On observe aussi que cette méthode représente mal la fluorescence avec une erreur de mesure (fig. 2.13(a)) ou avec une décroissance au début de la quantification (fig. 2.13(b)).



(a) C_{y_0} *method* sur une fluorescence avec présence d'erreur de mesure.

(b) C_{y_0} *method* sur une fluorescence avec décroissance au début.

FIGURE 2.13 – Comportement de la méthode C_{y_0} sur une fluorescence avec une erreur de mesure et une avec une décroissance au début.

2.3 Conclusion

Plusieurs méthodes existent pour déterminer ce cycle repère avec leurs avantages et leurs inconvénients (tab. 2.1).

Nous pouvons observer deux difficultés sur les données réelles de fluorescences, à savoir la présence d'erreurs de mesures et la décroissance au début de la courbe. Excepté la *fit point method*, les autres méthodes de l'état de l'art étudiées ne sont pas robustes sur des fluorescences avec ces deux difficultés. De ce fait, dans le chapitre 3, nous analyserons les résidus de la régression L_1 et L_2 par la méthode de la boite à moustaches pour sélectionner et supprimer les erreurs de mesures que l'on peut apercevoir sur des fluorescences. Puis, nous allons présenter un modèle analytique qui estime la fluorescence en tenant compte de la décroissance au début de la courbe.

Ces différentes méthodes de l'état de l'art étudiées, ne donnent pas toutes références le même cycle repère. Le plus important est que les cycles repères pour chaque méthode soient cohérents sur une même gamme de fluorescence. Chacune de ces méthodes a des avantages et des inconvénients. Néanmoins, aucune ne permet de déterminer le cycle repère en ligne tout en minimisant le retard à la détec-

	Détection c_r en ligne	**Hypothèse d'égalité de E**	**Minimise retard à la détection**	**Robustesse** Fluorescence avec erreur de mesure	Fluorescence avec décroissance début
Méthode du seuil [Kubista et al., 2006]	oui	oui	oui	☹	☹
Maximum dérivée seconde [Durtschi et al., 2007]	non	non	non	☹	☹
Fit point method [Durtschi et al., 2007]	non	oui	non	☺	☺
SCM [Rutledge, 2004]	non	non	non	☹	☹
C_{y_0} method [Guescini et al., 2008]	non	non	non	☹	☹
CUSUM [cf chapitre 3]	oui	non	oui	☺	☺

TABLE 2.1 – Tableau récapitulatif des avantages, des inconvénients et la robustesse des différentes méthodes de l'état de l'art sur la détermination du cycle repère.

tion. On étudiera dans le chapitre 4 une méthode statistique robuste, le CUSUM, qui tient compte de ces contraintes.

CHAPITRE 3

Filtrage des erreurs de mesures et modèle analytique des courbes fluorescences

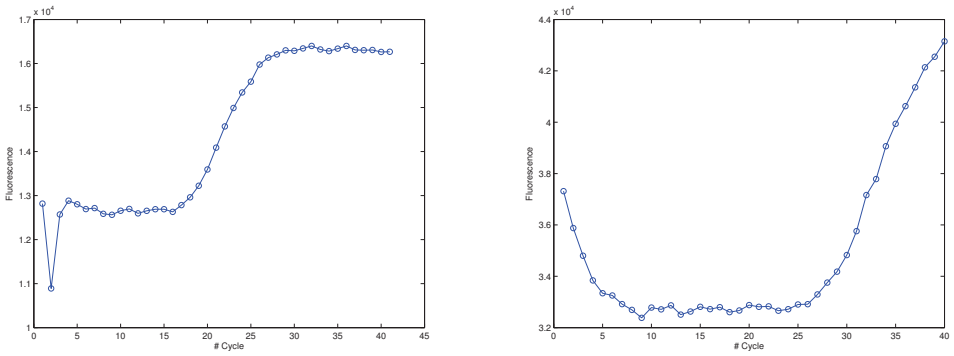
Sommaire

À la fin du chapitre 1, nous avons constaté deux difficultés sur une fluorescence issue d'une qPCR (erreurs de mesures et décroissance). Ainsi, dans ce chapitre, nous allons proposer une méthode robuste pour supprimer les erreurs de mesure que l'on peut rencontrer sur la fluorescence. Ensuite, nous allons proposer un modèle analytique pour représenter la fluorescence en tenant compte de la décroissance au début de la fluorescence.

3.1 Modèle représentatif de la fluorescence issue de la qPCR

3.1.1 Problématique

Nous avons vu au chapitre précédemment, qu'en appliquant une PCR associée à un marqueur fluorescent sur un échantillon à analyser, nous observons une émission lumineuse appelée fluorescence qui est fonction de la quantité de molécules initiales de l'agent biologique recherché. Les biologistes affirment que les fluorescences peuvent présenter des erreurs de mesures ou observations aberrantes et une décroissance au début de la quantification lors de la phase d'initiation, due à une dissociation du double brins d'ADN. De plus, la présence d'erreurs de mesure peut entrainer une mauvaise estimation du cycle repère. Ainsi, deux problématiques peuvent être observées sur les fluorescences, comme nous pouvons le voir sur la figure (3.1) : la présence de points aberrants et la décroissance au début des fluorescences.



(a) Fluorescence avec erreur de mesure (b) Fluorescence avec décroissance au début

FIGURE 3.1 – Illustration des deux problématiques liées au pré-traitement des données : présence erreurs de mesures (a), décroissance au début de la fluorescence (b).

Nous allons d'abord proposer une méthode robuste pour élimer les erreurs de mesures (fig. 3.1(a)) et proposer ensuite un modèle représentant la fluorescence en tenant compte de la décroissance initiale (fig. 3.1(b)).

La présence d'erreurs de mesure ou points aberrants avant le cycle repère peut entrainer une mauvaise estimation de ce dernier, donc une estimation erronée de la quantité initiale de l'agent biologique recherché. Nous proposons de faire un prétraitement sur les $t_{reg} = 15$ premiers cycles pour supprimer les erreurs de mesure qui peuvent apparaître sur la fluorescence. Les erreurs de mesures sont souvent observées à la phase d'initiation de la fluorescence lors de la PCR quantitative. Cette

partie de la fluorescence a une tendance linéaire, nous allons alors faire une régression linéaire et analyser ses résidus par la méthode de la boîte à moustaches pour supprimer les points hors épure (valeurs à l'extérieur des moustaches), considérés comme des erreurs de mesure.

3.1.2 Filtrage des observations aberrantes

Nous proposons de faire une régression linéaire sur les 15 premières observations de la fluorescence avant l'instant de rupture. Nous allons étudier deux types de régression : la régression des moindres carrées ou la régression L_2 (*least squares regression*) et la régression L_1 (*least absolute deviation*). Nous avons aussi testé le *lasso* (*Last Absolut Shrinkage and Selection Operator*, détaillé dans l'annexe C) mais sans obtenir de meilleur résultat.

3.1.2.1 Régression L_2 et L_1

Soit $(c_i, f_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, n$, où n est la taille des données ($n = 15$ dans notre cas). c_i représente le cycle i et f_i l'intensité de fluorescence amplifiée au cycle c_i. On cherche une fonction g telle que $f = g(c)$. Dans ce cas, le modèle linéaire s'écrit :

$$f_i = a_0 + a_1 c_i + \varepsilon_i \quad \forall\, i = 1, \ldots, n \tag{3.1}$$

où $\varepsilon = (\varepsilon_i)_{i=1,\ldots,n}$ est un bruit ; a_0 et a_1 sont des paramètres réels inconnus.

a) La méthode des moindres carrés (régression L_2) est la régression la plus populaire et elle remonte à Gauss et Legendre (voir [Stigler, 1981], pour une discussion historique). Comme nous l'avons vu dans le chapitre 2, en posant $a^t = (a_0, a_1)$, l'estimateur des moindres carrés des paramètres est égal à :

$$\hat{a}_{MC} = (C^t C)^{-1} C^t f \tag{3.2}$$

b) La régression L_1 détaillée dans [Rousseeuw and Leroy, 2005] consiste à minimiser la norme 1 du bruit, $||\varepsilon||_1 = \sum_{i=1}^{n} |\varepsilon_i|$. C'est à dire,

$$\min_{a_0,a_1} \sum_{i=1}^{n} |\varepsilon_i| = \min_{a_0,a_1} \sum_{i=1}^{n} |f_i - (a_0 + a_1 c_i)|\,, \tag{3.3}$$

En simplifiant l'équation 3.3, nous obtenons l'équation suivante :

$$\min_{a} ||Ca - f||_1 = \min_{a} \sum_{i=1}^{n} |\tilde{c}_i^t a - f_i|. \tag{3.4}$$

où $\tilde{c}_i^t = (1, c_i)$

La fonction $J(a) := \sum_{i=1}^{n} |c_i^t a - f_i|$ n'est pas différentiable en a, donc on ne peut pas utiliser la même technique que la méthode des moindres carrés pour estimer le paramètre a. Pour ce faire, nous allons poser :
$\beta_i := c_i^t a - f_i$ avec $\beta_i = \beta_i^+ - \beta_i^-$ où $\beta_i^+ \geq 0$ et $\beta_i^- \geq 0 \quad \forall\ i$.
Avec ces notations, le problème revient à :

$$(P1) : \begin{cases} \min\limits_{a,\beta} & \|\beta\|_1 \\ \\ \text{sc :} & \beta = Ca - f \end{cases} \tag{3.5}$$

$(P1)$ est équivalent à $(P2)$,

$$(P2) : \begin{cases} \min\limits_{a,\beta^+,\beta^-} & e^t(\beta^+ + \beta^-) \\ \\ \text{sc :} & \beta^+ - \beta^- = Ca - f \\ & \beta^+ \geq \mathbf{0}\ ;\ \beta^- \geq \mathbf{0} \end{cases} \tag{3.6}$$

où $e^t = (1, \ldots, 1)$.
En posant $a = a^+ - a^-$, avec $a_j^+ \geq 0$ et $a_j^- \geq 0$ (pour $j = 0, 1$), $(P2)$ est un programme linéaire qui peut s'écrire sous la forme suivante :

$$(P3) : \begin{cases} \min\limits_{a^+,a^-,\beta^+,\beta^-} & H^t\gamma \\ \\ \text{sc :} & A\gamma = b \\ & \gamma \geq \mathbf{0} \end{cases} \tag{3.7}$$

avec

$$H^t = \boxed{1, \ldots, 1} \boxed{1, \ldots, 1} \boxed{0,0} \boxed{0,0}$$

un vecteur de taille $(1, 2n + 4)$;

$$A = \boxed{I_n} \boxed{-I_n} \boxed{-C} \boxed{C}$$

une matrice de taille $(n, 2n + 4)$;

$$b = \boxed{-f}$$

un vecteur de $\mathbb{R}^n$ et

$$\gamma^t = \boxed{\beta^+} \boxed{\beta^-} \boxed{a^+} \boxed{a^-}$$

un vecteur $(1, 2n + 4)$.
Pour résoudre un tel problème, nous avons utilisé la fonction *linprog* de matlab qui résout un problème linéaire.

Notons $e_i,\ i = 1, \ldots, n$, les résidus. Ils sont estimés par :

$$\hat{e}_i = y_i - \hat{y}_i \quad \forall\ i \tag{3.8}$$

où $\hat{y}_i$ est la valeur estimée ou ajustée de y_i, $\forall\ i$.

Comme illustration de ces deux méthodes de régression, nous allons présenter les résultats de la régression que nous obtenons sur une même fluorescence avec erreur de mesure (fig. 3.2(a)). Sur cette dernière, nous constatons que la régression par la méthode des moindres carrés est moins robuste que la régression L_1. En effet la régression des moindres carrés estime les observations de telle sorte que le résidu moyen soit proche de zéro [Rousseeuw and Leroy, 2005]. Ainsi les erreurs de mesure influencent fortement la régression de moindres carrés tandis qu'elles agissent moins sur la régression L_1. La méthode des moindres carrés est alors très sensible aux erreurs de mesure.

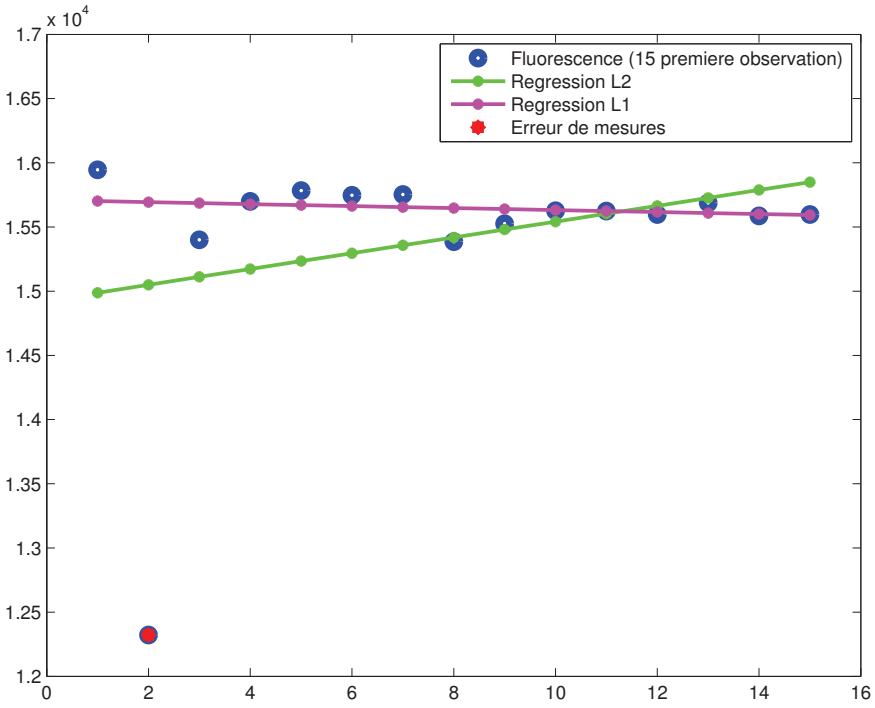
La figure (3.2(b)) montre le diagramme des résidus obtenus par la méthodes des moindres carrés et la régression L_1.

Nous remarquons que la deuxième valeur des résidus est une valeur largement inférieure aux autres. Cette valeur correspond à l'observation aberrante de la fluorescence (fig. 3.2(a)). Nous voulons mettre en place une règle de décision pour déterminer les résidus atypiques. Ainsi, dans la suite, nous allons déterminer un intervalle à partir duquel nous considérons les observations des résidus n'appartenant pas à cet intervalle comme des erreurs de mesure. Pour cela, nous allons utiliser la méthode de la boite à moustaches que nous détaillerons ci-dessous. Une étude plus détaillée, de l'analyse des résidus de la régression L_1 et L_2 par la méthode de la boîte à moustaches, sera présente à la fin de ce chapitre.

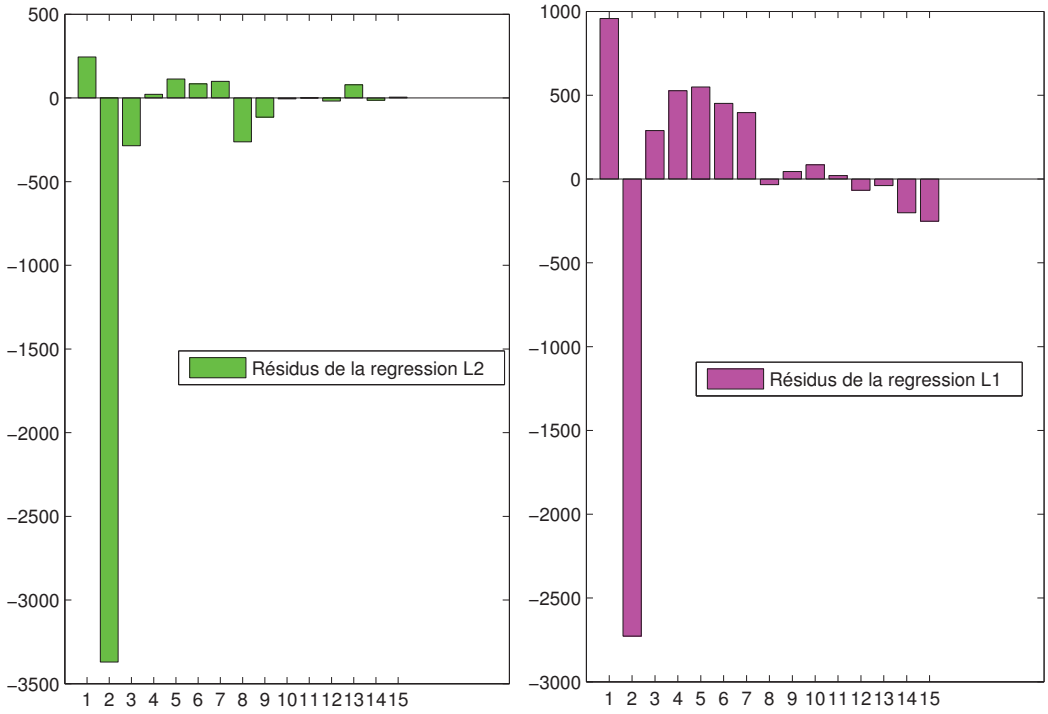
3.1.2.2 Boite à moustaches

La boîte à moustaches est une méthode basique pour l'interprétation des données. La boîte à moustaches (*Box & Whiskers Plot*), est une invention de [Tukey, 1977] permettant de représenter graphiquement un ensemble de données. Cette méthode graphique peut être utilisée pour l'identification des mesures atypiques ou extérieures sur les données.

Elle est constituée principalement des 3 quartiles (quartile 1, quartile 2 ou médiane et quartile 3) et les extrémités qui délimitent la taille des moustaches. La figure (3.3) représente un exemple de boite à moustache. Les signes + représentent les valeurs extérieures de la boite à moustaches que nous considérons comme des erreurs de mesure dont leur identification est donnée ci-dessous. Avant de mettre en place la règle de décision pour déterminer les valeurs externes, nous allons définir la notion d'écart interquartile.



(a) Illustration sur un exemple de fluorescence avec une erreur de mesure.



(b) Diagramme en bar des résidus estimés par la méthode des moindres carrés (gauche) et par la régression L_1 (droite).

FIGURE 3.2 – Illustration sur un exemple de fluorescence avec une erreur de mesure et le graphe des histogrammes des résidus estimés par la méthode des moindres carrés et par la régression L_1.

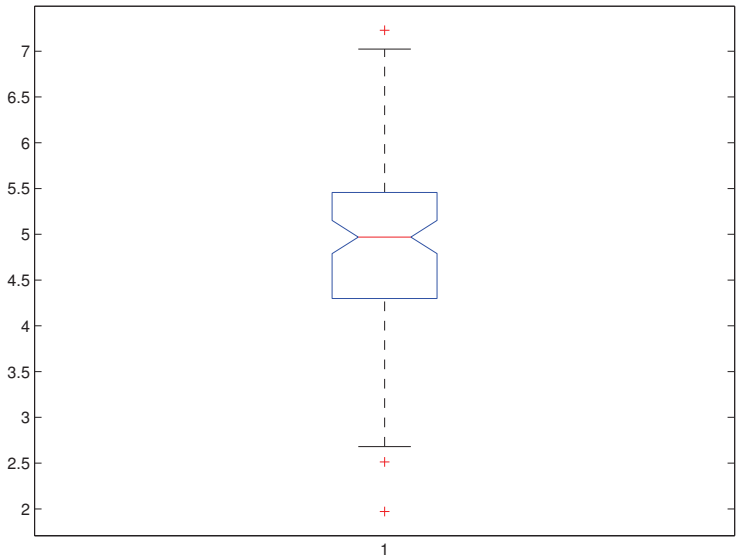


FIGURE 3.3 – Boîte à moustaches d'un exemple de données.

Définition : L'**écart interquartile** IQR est la longueur de l'intervalle interquartile $[Q_1, Q_3]$ (l'intervalle dont les bornes sont le 1^{er} et le 3^e quartiles) :

$$IQR = Q_3 - Q_1$$

En autre, l'écart interquartile est une mesure de la dispersion des 50% d'observations centrales.

Définissons les deux valeurs pivots suivantes :

$$p_g = Q_1 - \lambda\, IQR$$

$$p_d = Q_3 + \lambda\, IQR$$

où $\lambda \in \mathbb{R}^+$.

p_g et p_d sont situées à une distance de λ fois l'écart interquartile de part et d'autre de la boîte. Nous considérons qu'un résidu en dehors de l'intervalle $[p_g, p_d]$ comme valeur extérieure de la boîte à moustaches. Ainsi, l'observation associée à ce résidu est une erreur de mesure. Une fois une observation aberrante détectée, elle est remplacée par une nouvelle valeur égale à la moyenne des valeurs de ces deux plus proches observations.

Usuellement, on prend λ égale à 1.5 et on considère les valeurs en dehors de $[Q_1 - 1.5\ IQR,\ Q_3 + 1.5\ IQR]$ comme des valeurs extérieures de la boite à mous-

taches. Il n'existe pas une valeur optimal de λ et les résultats en analysant les résidus des régression L_1 et L_2 par la méthode de la boite à moustaches sont sensibles à λ (tab. 3.2). Dans la partie 3.2.2, nous présenterons les résultats obtenus avec deux valeurs de λ (1.5 et 2).

Nous présenterons dans la section résultats et expériences, l'analyse des résidus des régressions (L_2 et L_1) couplée avec la méthode de la boite à moustaches sur des données de fluorescences avec ou sans erreurs de mesure. Dans ce qui suit, nous allons présenter un modèle représentatif des fluorescences en tenant compte de la décroissance au début.

3.1.3 Modèle représentatif de la fluorescence

La modélisation des signaux de fluorescence par un modèle mathématique est faite pour s'affranchir du calcul intermédiaire du cycle repère et estimer directement la concentration de l'espèce biologique recherchée. De ce fait, nous allons proposer un modèle modifiant la fonction sigmoïde C_{y_0} prenant en compte la décroissance du début de la fluorescence. Nous avons aussi étudié un modèle auto-régressif en considérant que la phase initiale de la fluorescence comme un modèle auto-régressif (Annexe D) mais nous n'avons pas obtenu de bons résultats.

3.1.3.1 Approche

Nous avons vu que lors de la quantification par PCR, une décroissance peut être remarquée sur certains signaux de fluorescences (fig. 3.1(b)). En effet la fluorescence est due à la présence d'ADN double brin. De ce fait, la décroissance de la fluorescence au début peut être causée par une dissociation de ce double brin d'ADN lors de la quantification.

Comme nous l'avions vu sur le chapitre 1 (fig. 2.11(b) et fig. 2.13(b)), les méthodes sigmoïdes (SCM) et C_{y_0} estiment mal la décroissance de la fluorescence au début de la quantification. Ainsi, nous proposons un modèle tenant compte cette problématique.

3.1.3.2 Modèle

En tenant compte de la décroissance au début de la quantification, la phase initiale n'est pas constamment linéaire. Nous pouvons la scinder en deux phases : une partie décroissante et une partie à tendance constante (fig. 3.4). Ceci nous conduit à définir un nouveau modèle représentatif des signaux de fluorescence en considérant ces deux phases. Le comportement de cette phase initiale nous conduit à faire l'hypothèse qu'il existe un cycle τ tel qu'avant ce dernier on observe une certaine

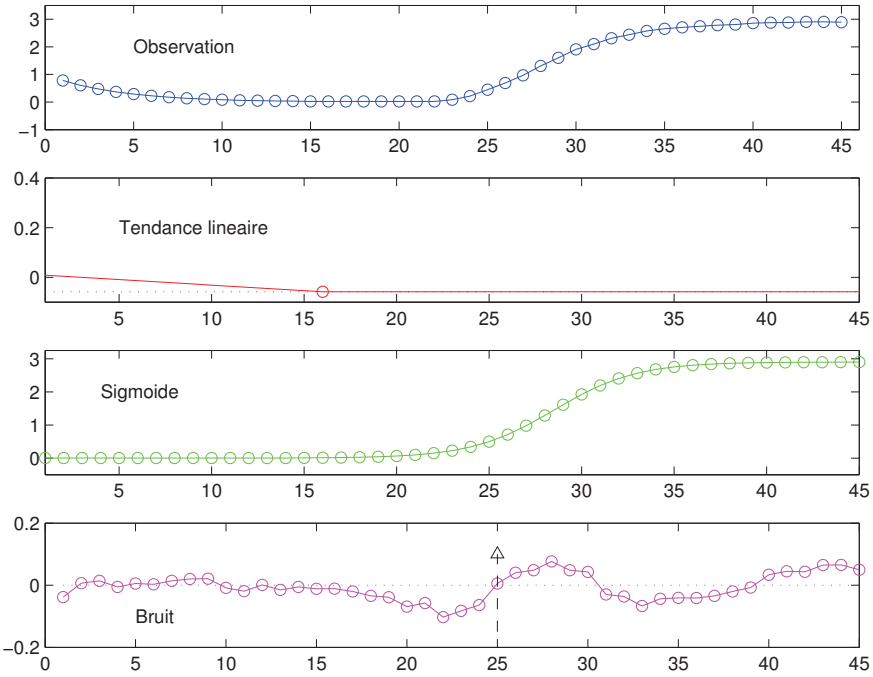


FIGURE 3.4 – Modèle représentatif des signaux de fluorescence : observation = tendance linéaire + sigmoïde + bruit

décroissance et après ce cycle la fluorescence reste constante jusqu'à un instant de rupture (cycle repère).

De ce fait, on propose le modèle suivant, qui reprend le modèle C_{y_0} plus la tendance linéaire :

$$O_c = \nu \min(c, \tau) + F_b + \frac{F\text{max}}{\left(1 + e^{\frac{-c+c_{1/2}}{k}}\right)^d} + \varepsilon_c \tag{3.9}$$

avec ε_c une erreur, $\tau > 0$ fixé et $\nu < 0, F_b, F\text{max}, c_{1/2}, k$ et $d > 0$ des paramètres à déterminer.

Le modèle (3.9) est non linéaire. Ainsi, comme dans le chapitre 2, pour estimer les paramètres de ce modèle, nous allons minimiser la norme euclidienne des résidus. En utilisant la méthode de descente du gradient conjugué sur un exemple de fluorescence, nous obtenons le résultat présenté sur la figure (3.5).

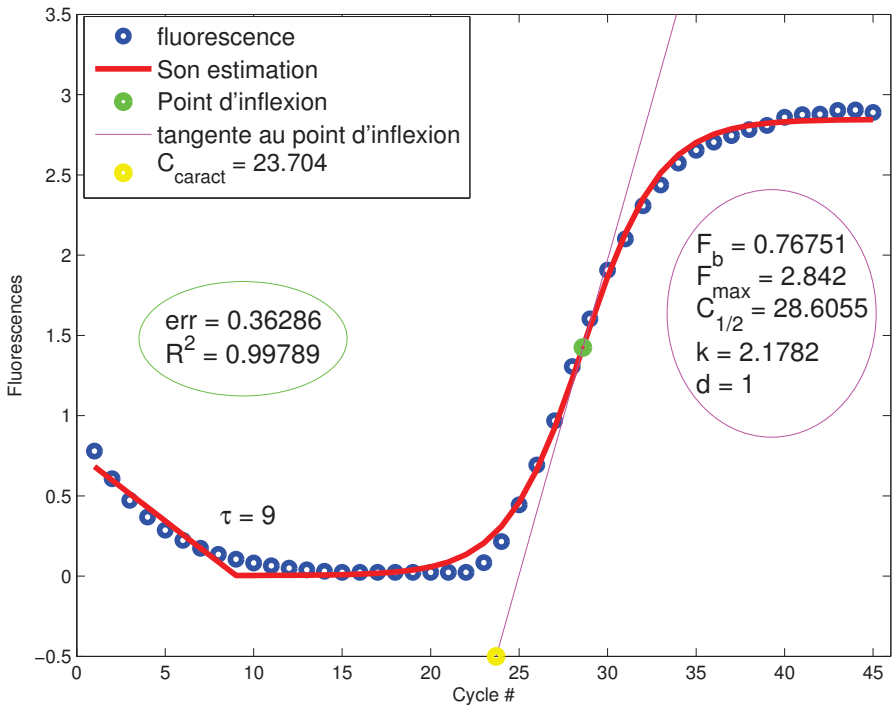


FIGURE 3.5 – Exemple d'une fluorescence avec décroissance au début de la quantification et son estimation par le modèle (3.9).

Par comparaison avec la méthode SCM (eq. 2.10) et de Richards (eq. 2.11) vu dans le chapitre 1, nous obtenons les résultats du tableau (3.1).

	Paramètres							
	τ	F_b	F_{max}	$C_{1/2}$	k	d	err	R^2
SCM	–	0.01	2.83	28.62	2.16	–	2.45	0.90
Richards	–	0.02	2.9	28.01	2.34	1	1.4	0.97
Notre modèle	9	0.73	2.86	28.58	2.2	1	**0.36**	**0.998**

TABLE 3.1 – Comparaison entre les méthodes sigmoïdes de l'état de l'art et notre modèle sur une même fluorescence avec un effet de décroissance au début.

Avec le modèle (3.9), le coefficient de détermination R^2 est meilleur et l'erreur d'estimation de la fluorescence est quatre fois plus petite que celle des modèles sigmoïdes (SCM et C_{y_0}). L'estimation de la fluorescence par ce modèle est meilleure

que celles des deux modèles sigmoïdes de l'état de l'art.

Dans ce qui suit, nous allons présenter les résultats expérimentaux obtenus sur différentes fluorescences avec la présence d'erreurs de mesures en les supprimant par la méthode de la boite à moustaches. Puis, nous allons appliquer et comparer le modèle (3.9) avec les méthodes SCM et C_{y_0} de l'état de l'art sur différentes fluorescences.

3.2 Résultats expérimentaux

3.2.1 Présentation des données

Nous disposons de différentes fluorescences issues du CEA, de la LHVP [1] et du laboratoire de l'IRSEM [2] (université de Rouen, plateforme PRIMACEN). Toutes ces fluorescences sont issues d'une quantification par PCR. En dehors du CEA et de la LHVP qui sont des partenaires sur le projet GENEASE, nous avons solliciter l'IRSEM pour avoir plus de données provenant différentes machine de PCR pour tester notre modèle et tester sa robustesse.

3.2.1.1 Jeux de données du CEA

Nous disposons d'une centaine de fluorescences de la part du CEA, un de nos partenaires sur le projet GENEASE, réalisées à un intervalle de temps différents et sur diverses expériences. Parmi ces fluorescences, nous pouvons observer des erreurs de mesures (fig. 3.1(a)) et la décroissance au début (fig. 3.1(b)).

3.2.1.2 Jeux de données du LHVP

Le LHVP qui est aussi un de nos partenaires sur le projet GENEASE, nous a aussi fourni deux types de fluorescence : avec un instant de rupture et sans instant de rupture (fig. 3.6). Un profil de fluorescence sans instant de rupture, appelé négative, signifie qu'il n'y a pas de présence d'espèce biologique (*cf* chapitre 1).

3.2.1.3 Jeux de données de l'IRSEM

Pour montrer la robustesse de notre méthode, nous avons fait appel à l'IRSEM de la plate-forme PRIMACEN [3]. Comme pour les jeux de données du LHVP, nous pouvons observer deux profils de fluorescences : avec rupture et sans rupture (fig. 3.7).

1. Laboratoire d'Hygiène de la Ville de Paris
2. Institut de Recherche en Systèmes Electroniques Embarqués
3. Plate-forme de Recherche en IMAgerie CEllulaire de haute-Normandie

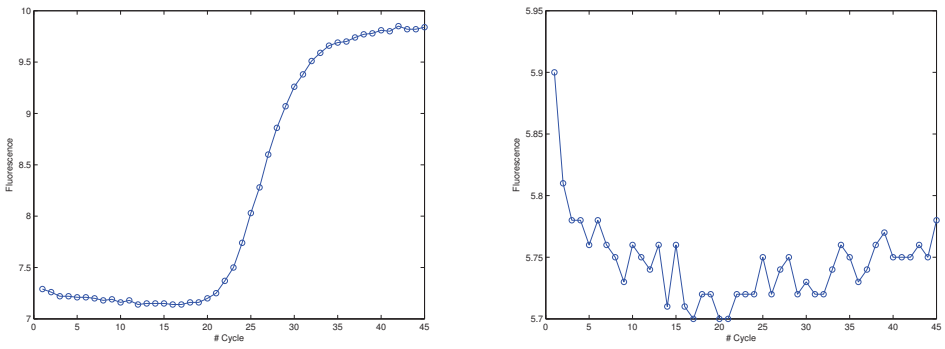


FIGURE 3.6 – Exemples de deux profiles de fluorescence de la LHVP.

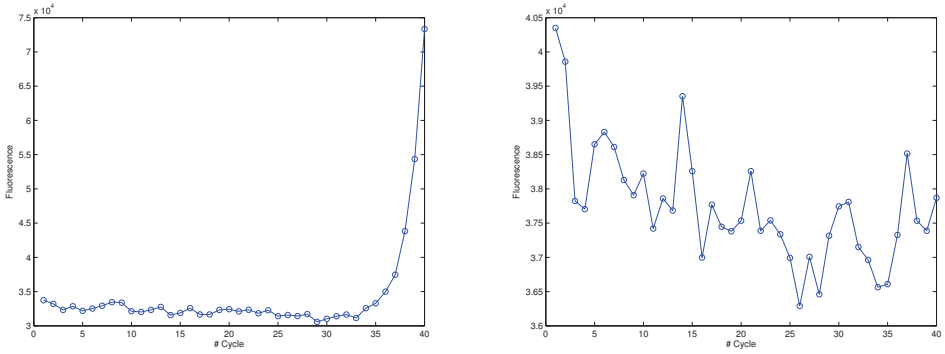


FIGURE 3.7 – Exemples de deux fluorescences de l'IRSEM.

À noter que l'instant de rupture, s'il y en a, peut être détecté très tardivement c'est à dire à partir du 35^{eme} cycle.

3.2.2 Erreurs de mesures

Nous allons appliquer les méthodes de régressions L_1 et L_2 couplées à la méthode de la boite à moustaches pour analyser les résidus des régressions sur 110 courbes expérimentales de fluorescences dont 18 présentant une erreur de mesure et 92 qui n'en ont pas. Comme mentionné précédemment, nous effectuerons le filtrage sur les 15 premiers cycles. Ainsi la performance du filtrage est déterminée sur les $15 \times 110 = 1650$ mesures.

En notant respectivement BD et FD les taux de bonne détection et de fausse détection d'erreurs de mesures, nous obtenons les résultats du tableau (3.2) pour deux valeurs de λ différents (λ est une grandeur pour déterminer la taille des moustaches).

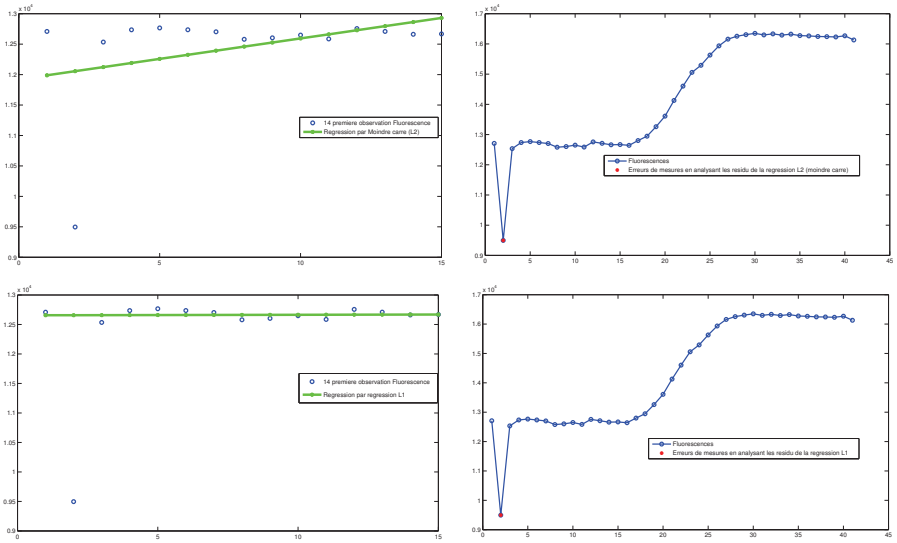
		BD	FD
λ = 1.5	Régression L_2 + boite à moustaches	100 %	1.71 %
	Régression L_1 + boite à moustaches	100 %	0.49 %
λ = 2	Régression L_2 + boite à moustaches	88.33 %	1.04 %
	Régression L_1 + boite à moustaches	**100** %	**0** %

TABLE 3.2 – Analyse des résidus de la régression L_1 et L_2 par la méthode de la boite à moustaches sur l'identification des erreurs de mesures.

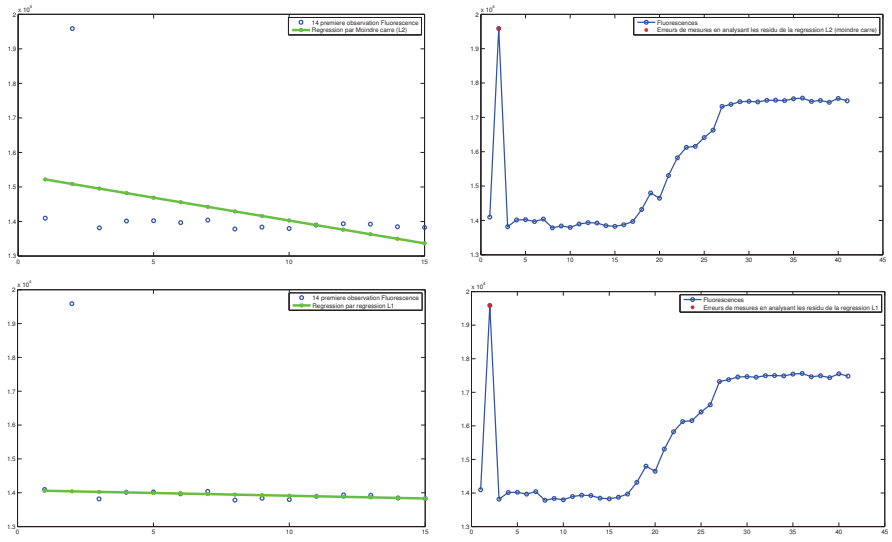
Les résultats sur les taux de bonne détection et de fausses détections sont sensibles à la valeur de λ. Néanmoins, les résultats obtenus en analysant les résidus de la régression L_1 par la méthode de la boite à moustache sont mieux que ceux obtenus en analysant les résidus de la régression L_2. Par exemple pour λ = 2, nous obtenons respectivement 100% et 88.33 % de bonne détection d'erreurs de mesures en analysant les résidus de la régression L_1 et respectivement L_2 par la méthode de la boite à moustaches. En plus, nous obtenons respectivement 0% et 1.04% de fausse détection d'erreurs de mesures en analysant les résidus de la régression L_1 et respectivement L_2 par la méthode de la boite à moustaches. Ainsi, en couplant la régression L_1 et la méthode de la boite à moustaches, nous détectons mieux les mesures aberrantes que lorsque nous couplons la régression des moindres carrés et la méthode de la boite à moustaches. La figure (3.8) illustre les résultats obtenus. Sur la figure (3.8(b)), nous remarquons que la régression L_2 identifie des mesures de mesures là où il n'y en a pas. En conclusion, la régression L_1 est alors plus robuste que la régression L_2 pour l'identification des erreurs de mesures en analysant les résidus. Voir annexe pour observer plus de résultats schématiques.

3.2.3 Décroissance au début de la fluorescence

Nous allons présenter les résultats d'une étude comparative de notre modèle par rapport aux modèles sigmoïdes SCM et C_{y_0} sur les données du CEA, du LHVP et du PRIMACEN. Pour chaque méthode, nous allons calculer le coefficient de corrélation moyen et l'erreur d'estimation moyenne de la fluorescence ($\bar{e}$). Nous allons proposer un intervalle de confiance autour de $\bar{e}$ définie ci-dessous. L'intervalle de confiance a été calculé en supposant que les résidus sont gaussiens, alors



(a)



(b)

FIGURE 3.8 – Illustration d'observations aberrantes sélectionnées par la méthode de la boite à moustaches sur trois fluorescences.

$IC_{95\%} = [\bar{e} - 1.96\frac{s}{\sqrt{n}} \; ; \; \bar{e} + 1.96\frac{s}{\sqrt{n}}]$, où n est la taille de l'échantillon et s est l'écart type estimé en utilisant l'estimateur sans biais de la variance.

En modélisant les fluorescence par les différents modèles énumérés ci-dessus, nous obtenus les résultats récapitulés sur le tableau (3.3).

		R^2 moyen	$\bar{e}$ ($IC_{95\%}$)
CEA	SCM	0.88	0.07 ± 0.027
	C_{y_0}	0.88	0.07 ± 0.025
	Notre modèle	0.94	0.05 ±0.02
LHVP	SCM	0.96	1.11 ± 0.39
	C_{y_0}	0.97	1.04 ± 0.16
	Notre modèle	0.998	0.36 ±0.10
PRIMACEN	SCM	0.95	0.51 ± 0.20
	C_{y_0}	0.97	0.39 ± 0.15
	Notre modèle	0.98	0.32 ±0.09

TABLE 3.3 – Résultats comparatifs des 3 modèles (notre modèle, SCM, C_{y_0}) pour représenter la fluorescence issue de la qPCR.

Sur le tableau (3.3), nous avons calculé le coefficient de détermination (R^2) moyen et l'erreur d'estimation moyenne de la fluorescence pour chaque modèle des différentes données dont nous disposons. Nous déduisons que le modèle (3.9) que nous avons mis en place estime mieux la fluorescence car son coefficient de détermination moyen est mieux et il a une plus petite erreur d'estimation de fluorescence que les méthodes SCM et C_{y_0}. Vous trouverez en annexe des figures illustrant les résultats obtenus sur l'estimation de la fluorescence par les trois modèles présentés ci-dessus.

3.3 Conclusion

Dans ce chapitre, nous avons d'abord présenté la régression L_2 et la régression L_1 pour modéliser les premières observations de la fluorescence. Grâce à la régression, nous déterminons les résidus qui sont la différence entre la droite de régression et les observations. Puis nous analysons ces résidus par la méthode de la boite à moustaches pour identifier les erreurs de mesure. Ensuite, nous avons mis en place un modèle analytique représentant les fluorescences en tenant compte de la décroissance au début de la quantification de la PCR. Nous avons obtenu de meilleurs résultats que les modèles de l'état de l'art étudiés. Dans le chapitre

qui suit, nous allons étudier une méthode statistique robuste pour la détection de rupture sur des données, le CUSUM.

CHAPITRE 4

Méthode statistique de détection de changements

Sommaire

L'OBJECTIF DE CE CHAPITRE est d'étudier et d'appliquer l'algorithme du CUSUM à notre problème. Le CUSUM est une méthode statistique de détection de changement construit à partir du rapport de vraisemblance dynamique entre deux hypothèses. Dans ce qui suit, nous allons d'abord faire une introduction générale sur les techniques de détection de changement, puis présenter en détails la méthode du CUSUM et enfin l'appliquer sur des fluorescences.

4.1 Introduction

La détection d'un agent pathogène est caractérisée par la mesure de sa quantité d'ADN initiale. Cette dernière est proportionnelle à une quantité appelée cycle repère qui est défini comme l'instant de rupture de la fluorescence issue d'une PCR quantitative (cf chapitre 1, fig. 2.1). La détection de cet instant de rupture consiste à repérer, à partir des signaux de fluorescence à notre disposition, le passage du système sous surveillance de la phase initiale, appelée régime $\mathcal{H}_0$, à la phase exponentielle (ou rupture), appelée régime $\mathcal{H}_1$. Ce passage de $\mathcal{H}_0$ et $\mathcal{H}_1$ est caractérisé par un changement de paramètre dans le modèle.

Ainsi, en considérant une suite de variables aléatoires $(X_k)_k$ de densité de probabilité $p_\theta(.)$, supposée connue, paramétrée par θ. La rupture est caractérisée par un changement sur le paramètre θ du modèle. Avant l'instant de rupture, noté c_r, θ vaut θ_0 et après c_r, θ vaut θ_1 (avec $\theta_1 \neq \theta_0$).

Il existe deux façons d'aborder le problème :

— **la détection hors-ligne** : à partir d'un échantillon de taille n, le but est de décider si oui ou non il y a un changement sur la paramètre θ et d'estimer l'instant de rupture c_r.

— **la détection en-ligne** : les observations arrivent en continu et on cherche à détecter le plus rapidement possible un changement sur le paramètre, s'il y en a. Nous allons construire une règle de décision qui, à chaque fois qu'une nouvelle observation arrive, tranchera entre les régimes $\mathcal{H}_0$ et $\mathcal{H}_1$. On cherche aussi à minimiser le retard à la détection. Il faudra alors déterminer l'instant d'arrêt, noté c_a, où l'hypothèse $\mathcal{H}_0$ est rejetée. Néanmoins si l'hypothèse $\mathcal{H}_0$ est retenue, la surveillance continue jusqu'au changement de régime ou jusqu'à la fin des observations.

  Si on a détecté un instant d'alarme c_a sachant qu'on est sous le régime $\mathcal{H}_0$, alors nous parlerons de fausse alarme. Au cas où $\mathcal{H}_0$ est rejetée, on estime l'instant de rupture c_r à l'instant c_a et la quantité $c_a - c_r$ est appelée le *retard à la détection*.

La détection en ligne est le problème le plus compliqué puisqu'on doit prendre une décision en minimisant le retard à la détection. Dans le cadre de cette thèse, pour estimer le cycle repère, nous allons appliquer la méthode du CUSUM sur des données de fluorescence fournies par le LHVP, le CEA et l'IRSEM, avec l'objectif de minimiser le retard à la détection. Pour effectuer ce travail, nous nous sommes basés sur le livre de Basseville and Nikiforov [1993].

Remarque : Tout au long de ce document, nous avons caractérisé une détection de rupture par un changement de paramètre (qui passe de la valeur de θ_0 à θ_1) dans la loi des observations (test paramétrique) plutôt que par un changement total de la loi des observations caractérisée par deux densités de probabilités : p sous $\mathcal{H}_0$ et q sous $\mathcal{H}_1$. En pratique, il est difficile de modéliser la loi des observations avant et après la rupture. Ce type de problème est abordé par exemple dans la thèse de Verdier [2007].

Règles de détection élémentaires : Il existe d'autres méthodes de détection de changement en ligne, comme par exemple la méthode des cartes de contrôle ou règle de Shewhart [1931] et les règles de la moyenne mobile étudiées par Roberts [1959]. Ces algorithmes élémentaires, tous basés sur le rapport de vraisemblance, peuvent être utilisés en pratique mais ne possèdent pas de propriété d'optimalité. Or, il est important de pouvoir montrer l'optimalité d'une méthode, pour garantir sa robustesse. Ainsi, une des toutes premières règles possédant des propriétés d'optimalité, pour la détection de changement sur des données, est le CUSUM que nous allons présenter maintenant.

4.2 Méthode du CUSUM

La règle CUSUM, ou règle des sommes cumulées, est une technique d'analyse séquentielle permettant la détection de changement en ligne. Elle est introduite par Page [1954]. En considérant une distribution de probabilité p_θ, paramétrée par θ (par exemple l'espérance), Page a voulu mettre en place une méthode qui consiste à détecter un changement sur ce paramètre.

4.2.1 Le rapport de vraisemblance

La vraisemblance a été mise en œuvre pour quantifier l'adéquation entre une distribution de probabilité et un échantillon. Plus est grande la vraisemblance de l'échantillon, meilleure est l'adéquation. Par ailleurs, le but des tests est de discriminer entre deux hypothèses ($\mathcal{H}_0$ *vs* $\mathcal{H}_1$) qui affirment chacune contenir la distribution qui a engendré l'échantillon. C'est dans ce sens que le théorème de Neyman-Pearson a été mis en place. Ce théorème indique la meilleure région de rejet de l'hypothèse nulle $\mathcal{H}_0$, voir Lehmann [2005]. Cette région de rejet est caractérisé par un rapport de vraisemblance entre deux distributions.

Nous supposons que les distributions que nous allons rencontrer, appartiennent à une même famille et ne différent que par la valeur d'un paramètre θ.

Définition : Soit X une variable aléatoire continue. La fonction de vraisemblance notée $L(X;\theta)$, est une fonction de probabilités conditionnelles qui décrit les observations x_i ($i = 1,\ldots,n$) de X en fonction du paramètre θ. Elle s'exprime à partir de la fonction de densité $p_\theta(X)$ de X par

$$L(X;\theta) = \prod_{i=1}^{n} p_\theta(x_i) \tag{4.1}$$

Une grande partie des règles de décision sur la détection de changements sur des données est basée sur le logarithme du rapport de vraisemblance :

$$s(X) = \log \frac{L(X;\theta_1)}{L(X;\theta_0)} \tag{4.2}$$

En se basant sur les travaux de Basseville and Nikiforov [1993], nous avons :

$$E_{\theta_0}(s) < 0 \quad \text{et} \quad E_{\theta_1}(s) > 0 \tag{4.3}$$

où $E_{\theta_0}(s)$ et $E_{\theta_1}(s)$ sont respectivement les espérances sous p_{θ_0} et p_{θ_1}.

Ainsi un changement de paramètre entraine un changement de signe pour l'espérance du logarithme du rapport de vraisemblance. C'est à partir de cette propriété que des algorithmes de détection de changement sont construits.

4.2.2 La règle du CUSUM

Les données dont nous disposons sont des signaux de fluorescences issues d'une PCR quantitative (voir chapitre 1). Une courbe fluorescence peut être caractérisée par un cycle repère qui est l'instant du début de la phase exponentielle (fig. 1.6). La présence de ce cycle repère permet d'affirmer qu'il y a présence de l'espèce biologique recherché et nous permet d'estimer la quantité initiale de molécule d'ADN de l'agent biologique. Dans le cas contraire, il n'y a pas présence de l'espèce recherché.

4.2.2.1 Test d'hypothèse

Notons X la variable aléatoire (v.a) représentant la fluorescence et $\{x_t\}_{t=1,\ldots,c_{max}}$, avec c_{max} le cycle maximal de la fluorescence, les réalisations de la v.a X qui admet une densité de probabilité p_θ paramétrée par $\theta \in \mathbb{R}$. Si nous supposons que $p_\theta = N(\mu,\sigma^2)$, où μ et σ^2 représentent respectivement l'espérance et la variance alors $\theta = (\mu,\sigma^2)$.

Deux situations se présentent pour la détection d'un changement sur des données :

— soit toutes les observations de l'échantillon sont des réalisations d'une même v.a suivant la même densité de probabilité, caractérisée par θ_0,

— soit il existe un instant c_r inconnu, $1 \leq c_r \leq c_{max}$, tels que $\theta = \theta_0$ avant c_r et $\theta = \theta_1$ après c_r.

En écrivant ces deux situations sous la forme d'un test d'hypothèses, comme sur les travaux de Keita et al. [2012, 2013], nous obtenons :

$\forall\ j$ tel que $1 \leq j \leq c_{max}$,

$$\begin{cases} \mathcal{H}_0 : & \{X_t\}_{t=1,\dots,j} \rightsquigarrow p_{\theta_0} \\ \mathcal{H}_1 : & \exists\ c_r \text{ tel que} \\ & \quad \{X_t\}_{t=1,\dots,c_r-1} \rightsquigarrow p_{\theta_0} \\ & \quad \{X_t\}_{t=c_r,\dots,j} \rightsquigarrow p_{\theta_1} \end{cases} \tag{4.4}$$

Notons respectivement $L_{\mathcal{H}_1}(X;\theta)$ et $L_{\mathcal{H}_0}(X;\theta)$ les vraisemblances sous les hypothèses $\mathcal{H}_1$ et $\mathcal{H}_0$. Elles sont définies comme suit,

$$L_{\mathcal{H}_1}(X;\theta) = \prod_{t=1}^{c_r-j} p_{\theta_0}(x_t) \prod_{t=c_r}^{j} p_{\theta_1}(x_t) \tag{4.5}$$

$$L_{\mathcal{H}_0}(X;\theta) = \prod_{t=1}^{j} p_{\theta_0}(x_t) \tag{4.6}$$

Si la détection hors-ligne est un test d'hypothèse « à la Neyman-Pearson », puisqu'on peut parler d'erreur de première espèce (probabilité de rejeter à tord $\mathcal{H}_0$) et de puissance (probabilité de rejeter $\mathcal{H}_0$ sachant qu'elle est fausse). Il en va différemment pour la détection en-ligne puisqu'il est difficile d'utiliser le concept de puissance. Pour définir et calibrer le test, on utilisera plutôt la notion de retard à la détection et l'erreur de premier espèce α (la probabilité de fausse alarme).

Nous désirons détecter l'instant de rupture en minimisant le retard à la détection. Le test se fait alors de manière récursive suivant j. Lorsque j est égal à c_{max}, nous pouvons avoir deux cas : soit nous sommes sous $\mathcal{H}_0$, cela signifie qu'il n'y a pas de présence de l'espèce recherchée ; soit $\mathcal{H}_1$ est choisie, nous avons alors la présence de l'espèce recherchée, $c_a = c_{max}$ et on estime c_r.

4.2.2.2 Mécanisme du CUSUM

À l'instant j, le log du rapport de vraisemblance noté $s(X)$ est égal à,

$$s(X) = \log \frac{\prod_{t=1}^{c_r-1} p_{\theta_0}(x_t) \prod_{t=c_r}^{j} p_{\theta_1}(x_t)}{\prod_{t=1}^{j} p_{\theta_0}(x_t)} \tag{4.7}$$

$$= \log \prod_{t=c_r}^{j} \frac{p_{\theta_1}(x_t)}{p_{\theta_0}(x_t)} \tag{4.8}$$

$$= \sum_{t=c_r}^{j} \log \frac{p_{\theta_1}(x_t)}{p_{\theta_0}(x_t)} \tag{4.9}$$

L'instant de rupture c_r étant inconnu, notons alors $\Phi_{c_r}(j)$ l'indicateur du CUSUM, $\forall\ 1 \leq c_r \leq j$

$$\Phi_{c_r}(j) := s(X) = \sum_{t=c_r}^{j} \log \frac{p_{\theta_1}(x_t)}{p_{\theta_0}(x_t)} \tag{4.10}$$

Pour chaque instant j, on calcule l'indicateur Φ, illustré par le schéma suivant :

$$\begin{aligned}
j = 1 &\longrightarrow \Phi_1(1) \\
j = 2 &\longrightarrow \Phi_1(2), \Phi_2(2) \\
j = 3 &\longrightarrow \Phi_1(3), \Phi_2(3), \Phi_3(3) \\
\vdots &\qquad\qquad \vdots \\
j = c_{max} &\longrightarrow \Phi_1(c_{max}), \Phi_2(c_{max}), \Phi_3(c_{max}), \ldots, \Phi_{c_{max}}(c_{max})
\end{aligned}$$

Nous allons déterminer dans ce qui suit le comportement de l'indicateur $\Phi_{c_r}(j)$ pour j fixé. En effet, en exprimant $\Phi_{c_r+1}(j)$ en fonction de $\Phi_{c_r}(j)$, nous obtenons :

$$\Phi_{c_r+1}(j) = \Phi_{c_r}(j) - \log \frac{p_{\theta_1}(x_{c_r})}{p_{\theta_0}(x_{c_r})},\ \forall\ 1 \leq c_r \leq j-1 \tag{4.11}$$

Alors $\forall\ 1 \leq c_r \leq j-1$

$$\Phi_{c_r+1}(j) \begin{cases} > \Phi_{c_r}(j) & \text{si} \quad \log \frac{p_{\theta_1}(x_{c_r})}{p_{\theta_0}(x_{c_r})} < 0 \\ < \Phi_{c_r}(j) & \text{si} \quad \log \frac{p_{\theta_1}(x_{c_r})}{p_{\theta_0}(x_{c_r})} > 0 \\ = \Phi_{c_r}(j) & \text{si} \quad \log \frac{p_{\theta_1}(x_{c_r})}{p_{\theta_0}(x_{c_r})} = 0 \end{cases} \tag{4.12}$$

Ainsi, nous avons une constance ou une croissance ou une décroissance de Φ selon la quantité $\log \frac{p_{\theta_1}(x_{c_r})}{p_{\theta_0}(x_{c_r})}$.

Illustration : Si p_θ est la densité de la loi gaussienne $N(\mu, \sigma^2)$ et en considérant que σ^2 est connu et fixe (changement que sur l'espérance μ, $\mu_0 \neq \mu_1$), nous obtenons (Keita et al. [2012]),

$$\Phi_{c_r}(j) = \frac{\mu_1 - \mu_0}{\sigma^2} \sum_{t=c_r}^{j} \left(x_t - \frac{\mu_1 + \mu_0}{2} \right) \tag{4.13}$$

L'équation (4.11) devient,

$$\Phi_{c_r+1}(j) = \Phi_{c_r}(j) - \frac{\mu_1 - \mu_0}{\sigma^2} \left(x_{c_r} - \frac{\mu_1 + \mu_0}{2} \right), \ \forall \ 1 \leq c_r \leq j-1 \tag{4.14}$$

Alors, $\forall \ 1 \leq c_r \leq j-1$,

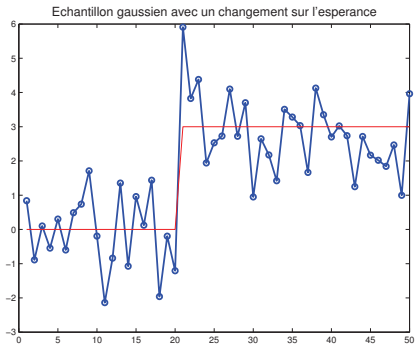
$$\Phi_{c_r+1}(j) \begin{cases} > \Phi_{c_r}(j) & \text{si} \quad x_{c_r} < \frac{\mu_1+\mu_0}{2} \\ < \Phi_{c_r}(j) & \text{si} \quad x_{c_r} > \frac{\mu_1+\mu_0}{2} \\ = \Phi_{c_r}(j) = 0 & \text{si} \quad x_{c_r} = \frac{\mu_1+\mu_0}{2} \quad \text{ou } \mu_1 = \mu_0 \end{cases}$$

Notons $\gamma = \frac{\mu_1+\mu_0}{2}$ qui représente la moyenne des espérances. Si on est sous le régime $\mathcal{H}_1$ ($\mu_0 \neq \mu_1$), on observe deux comportements de l'indicateur du CUSUM Φ, à savoir une croissance pour des observations inférieurs à γ et une décroissance pour des observations supérieurs à γ. Ainsi l'instant de rupture est le moment où on observe un changement de signe de Φ.

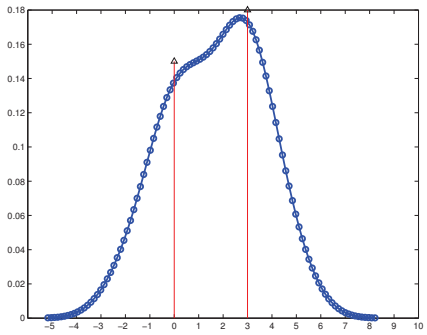
NB : Si $\sigma_0^2 \neq \sigma_1^2$, l'indicateur du CUSUM est égale à (Basseville and Nikiforov [1993]),

$$\Phi_{c_r}(j) = (j - c_r + 1) \ \log \frac{\sigma_0^2}{\sigma_1^2} + \sum_{t=c_r}^{j} \frac{(x_t - \mu_0)^2}{\sigma_0^2} - \frac{(x_t - \mu_1)^2}{\sigma_1^2} \tag{4.15}$$

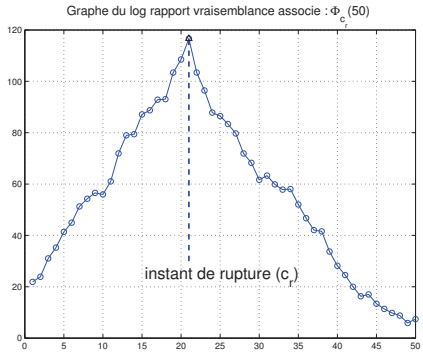
Exemple : Soit X une suite de variables aléatoires gaussiennes de taille $n = 50$ tel que l'espérance avant la rupture $c_r = 20$ vaut $\mu_0 = 0$ et après c_r est égal à $\mu_1 = 3$ et $\sigma_0 = \sigma_1 = 1$ (fig. 4.1(a)). On voit bien la concaténation de deux gaussiens en représentant la densité de probabilité de X (fig. 4.1(b)). La figure (4.1(c)) montre le comportement typique du log rapport de vraisemblance $\Phi_{c_r}(50)$, $c_r = 1, \ldots, 50$. Ce dernier est une courbe en forme de *v* renversé où on observe un changement de comportement après l'instant de rupture.



(a) Échantillon gaussien avec un changement sur la moyenne.



(b) Sa densité de probabilité correspondante.



(c) Inducteur du CUSUM $\Phi_{c_r}(50)$.

FIGURE 4.1 – Comportement de l'inducteur du CUSUM correspondant à un changement sur l'espérance d'une distribution gaussienne avec une variance constante.

En outre, nous observons une croissance jusqu'à la valeur maximale M de l'indicateur du CUSUM Φ et une décroissance après M. Par conséquent, dans le cadre de la détection en ligne d'une rupture, les informations utiles résident dans la différence entre la valeur de Φ et de sa valeur maximale à chaque instant j, $1 \leq j \leq c_{max}$. Ainsi, la règle de décision correspondante consiste, à chaque instant j, à comparer cette différence à un **seuil** h convenablement choisi. Il est choisi par calibrage, ceci est présenté sur la partie 4.3.

Posons,

$$M_j := \max_{1 \leq c_r \leq j} \Phi_{c_r}(j), \tag{4.16}$$

qui est la valeur maximale du $\log$ du rapport de vraisemblance des hypothèses $\mathcal{H}_1$ et $\mathcal{H}_0$ des observations de 1 à j avec $1 \leq j \leq c_{max}$.

Puis nous posons la **statistique de test** suivante, $\forall\ 1 \leq j \leq c_{max}$:

$$g_j = M_j - \Phi_j(j). \tag{4.17}$$

Le comportement de g, sur l'exemple gaussien précédent, est représenté sur la figure (4.2).

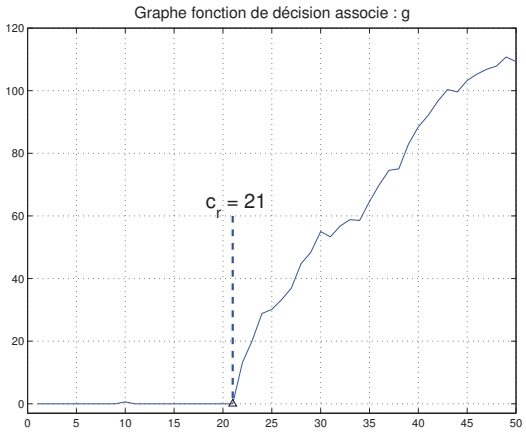


FIGURE 4.2 – Comportement de la fonction de décision g appliqué sur l'échantillon gaussien précédent avec un changement sur l'espérance.

Sur la figure (4.2) où nous avons représenté le comportement de la fonction de décision g, nous remarquons qu'elle est nulle avant l'instant de rupture c_r et croissante après c_r.

Ce qui nous conduit à définir la règle de décision ci-dessous.

1. **Règle de décision :** La règle de décision d, telle que $d = 0$ ou 1 est accordée respectivement à $\mathcal{H}_0$ ou $\mathcal{H}_1$, est donnée par,

$$d = \begin{cases} 1 & \text{si } g_j \geq h : \text{ on rejette } \mathcal{H}_0 \\ 0 & \text{sinon} \end{cases}$$

   où h est un seuil convenablement choisi.

2. **Calcul du temps d'alarme** c_a **:** Nous détectons la rupture ou l'instant d'arrêt noté c_a suivant le paramètre h (c_a est différent de l'instant de rupture c_r). En outre, c_a est l'instant minimum où une rupture est observée, c'est à dire le choix $d = 1$ est pris.

$$c_a = \inf\{j\ : g_j \geq h\}, \quad 1 \leq j \leq c_{max} \tag{4.18}$$

   Une fois que l'instant d'arrêt c_a est déterminé, nous sommes alors sous $\mathcal{H}_1$, il reste à estimer l'instant de rupture c_r.

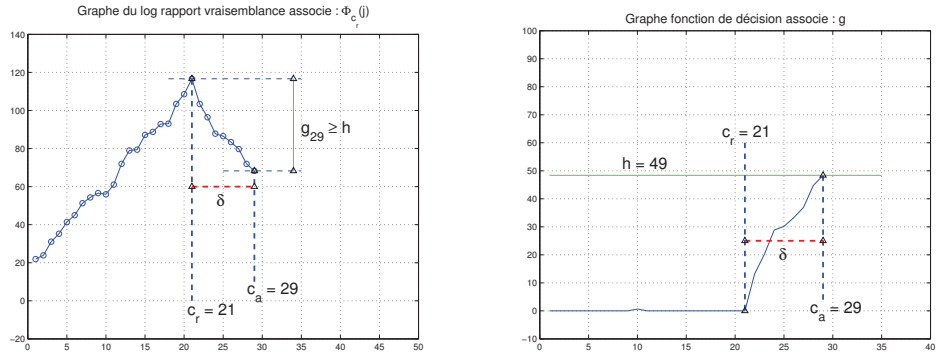


FIGURE 4.3 – Illustration du comportement en ligne de l'indicateur du CUSUM (à gauche) et de de la fonction de décision CUSUM (à droite) appliqué sur l'exemple précédent pour $h = 49$.

3. **Estimation du cycle repère $\hat{c}_r$:** Sous $\mathcal{H}_1$, il existe un instant c_r où il y a une rupture. Le problème revient alors à estimer cet instant. C'est l'instant où le graphe de l'indicateur du CUSUM Φ change de signe (eq. (4.12) et fig 4.1(c)). Ainsi, l'instant de rupture est estimé par :

$$\hat{c}_r = \underset{1 \leq c_r \leq c_a}{\arg\max}\, \Phi_{c_a}(c_r). \tag{4.19}$$

Le retard à la détection δ est alors égal à,

$$\delta = c_a - \hat{c}_r. \tag{4.20}$$

En appliquant la statistique de test g_j sur l'exemple précédent (fig. 4.1), nous obtenons la figure (4.3).

Remarque : Comme nous l'avons vu dans le chapitre 1, la fluorescence est obtenue en calculant à chaque cycle (un cycle dure environ une minute) l'amplification de la réaction. Ainsi la règle du CUSUM donne un instant de rupture discret c_r^{cusum} (fig. 4.4). Cet instant peut être imprécis pour quantifier la concentration. Nous avons alors utilisé une approximation spline cubic de l'indicateur du CUSUM, et l'instant de rupture c_r est estimé par le cycle qui maximise la spline (fig. 4.4). Nous avons utilisé la fonction spline d'ordre 3 de matlab.

Dans ce qui suit, nous allons détailler l'algorithme du CUSUM dans le cas où les paramètres θ_0 et θ_1 sont inconnus.

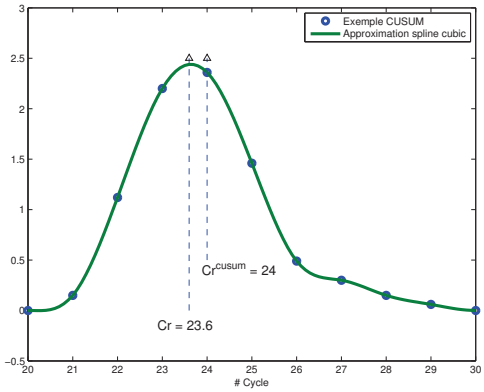
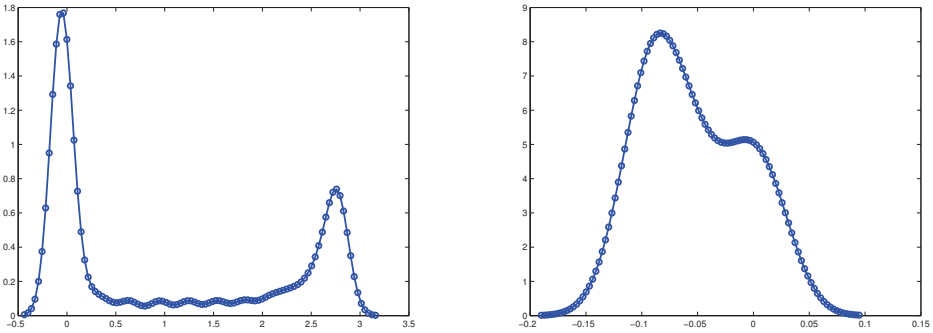


FIGURE 4.4 – Illustration de l'approximation spline cubic sur l'indicateur du CUSUM.

4.2.3 Algorithme du CUSUM

Comme nous l'avons mentionné précédemment, il est difficile de trouver la vraie loi modélisant les observations. Sur la figure (4.5(a)), nous observons trois comportements en représentant la densité de probabilité de la fluorescence : deux qui ressemblent à des gaussiens et correspondant à la phase initiale et phase plateau. Entre les deux, nous observons une autre loi qui est difficile à qualifier. Elle modélise la phase exponentielle.



(a) Comportement de la densité de probabilité de la totalité de la fluorescence (1.6).

(b) Comportement de la densité de probabilité de la fluorescence (1.6) en considérant que les c_a premières observations.

FIGURE 4.5 – Illustration de la densité de probabilité de la fluorescence (1.6) en considérant toute la fluorescence (à gauche) et en considérant qu'une partie de la fluorescence (de la première observation jusqu'au déclenchement de l'alarme c_a).

Le plus difficile est alors de trouver la loi sous $\mathcal{H}_1$.

Nous verrons dans la partie expérimentation qu'avec la méthode du CUSUM, la détection a eu lieu très tôt et l'instant d'alarme est déclenché entre 1 et 2 cycles environ après l'instant de rupture (la montée). De ce fait les observations correspondantes à ces cycles ne figurent pas sur la partie ascendante de la phase exponentielle. En représentant la densité de probabilité de la fluorescence de la première observation jusqu'à l'instant d'alarme, nous observons une concaténation de deux gaussiennes (fig. 4.5(b)).

Ainsi, d'un point de vu pratique, nous allons supposer que « les observations suivent une loi normale » paramétrée par $\theta = \{\mu, \sigma^2\}$, où μ représente l'espérance et σ^2 la variance.

4.2.3.1 Estimation des paramètres θ_0 et θ_1

Les paramètres avant la rupture θ_0 et après la rupture θ_1 ne sont pas nécessairement connus. Dans ce cas, nous utilisons le log du rapport de vraisemblance généralisé, consistant à modifier légèrement l'équation (4.7), voir Basseville and Nikiforov [1993] :

$$\sup_{\theta_1,\theta_0} s(X) = \sup_{\theta_1,\theta_0} \sum_{t=c_r}^{j} \log \frac{p_{\theta_1}(x_t)}{p_{\theta_0}(x_t)} =: \hat{\Phi}_{c_r}(j) \tag{4.21}$$

En pratique l'estimation de θ_0 est facile à réaliser. En effet, nous l'estimons sur les $t_{reg} = 15$ premières observations car on admet que l'instant de rupture sur la fluorescence n'est pas observé avant ces 15 observations. L'estimation de θ_0 est réalisée après avoir supprimer les erreurs de mesures (voir chapitre 3). En effet, nous faisons un pré-traitement des données sur les 15 premiers cycles, puis nous estimons θ_0.

Si par exemple $\theta_0 = \{\mu_0, \sigma_0^2\}$ représentant respectivement l'espérance et la variance avant la rupture, nous prenons alors leurs estimateurs sans biais, définies comme suit

$$\hat{\mu}_0 = \frac{1}{t_{treg}} \sum_{t=1}^{t_{reg}} x_t \tag{4.22}$$

$$\hat{\sigma_0}^2 = \frac{1}{t_{treg} - 1} \sum_{t=1}^{t_{reg}} (x_t - \hat{\mu}_0)^2 \tag{4.23}$$

Il reste à estimer $\theta_1 = \{\mu_1, \sigma_1^2\}$ représentant respectivement l'espérance et la variance après la rupture. Nous supposons que les variances avant et après la rupture ne changent pas, c'est à dire $\sigma_0^2 = \sigma_1^2 = \sigma^2$. Nous avons : $\forall\ t_{reg} \leq j \leq c_{max}$,

$$\hat{\Phi}_{c_r}(j) = \sup_{\mu_1} \frac{\mu_1 - \hat{\mu}_0}{\sigma^2} \sum_{t=c_r}^{j} \left(x_t - \frac{\mu_1 + \hat{\mu}_0}{2} \right) \tag{4.24}$$

Notons $f(\mu_1) := \frac{\mu_1 - \hat{\mu}_0}{\sigma^2} \sum_{t=c_r}^{j} \left(x_t - \frac{\mu_1 + \hat{\mu}_0}{2} \right)$. Ainsi, nous allons d'abord calculer la dérivée de f par rapport à μ_1 puis l'annuler pour évaluer l'estimateur de μ_1. Nous avons,

$$\frac{d\, f(\mu_1)}{d\, \mu_1} = \frac{1}{\sigma^2} \sum_{t=c_r}^{j} \left(x_t - \frac{\mu_1 + \hat{\mu}_0}{2} \right) + \frac{\mu_1 - \hat{\mu}_0}{\sigma^2} \sum_{t=c_r}^{j} \left(\frac{-1}{2} \right) \tag{4.25}$$

$$= \frac{1}{\sigma^2} \sum_{t=c_r}^{j} \left(x_t - \frac{\mu_1 + \hat{\mu}_0}{2} \right) - \frac{j - c_r + 1}{2\, \sigma^2} (\mu_1 - \hat{\mu}_0) \tag{4.26}$$

$$= \frac{1}{\sigma^2} \sum_{t=c_r}^{j} x_t - \frac{j - c_r + 1}{\sigma^2} \left(\frac{\mu_1 + \hat{\mu}_0}{2} \right) - \frac{j - c_r + 1}{2\, \sigma^2} (\mu_1 - \hat{\mu}_0) \tag{4.27}$$

$$= \frac{1}{\sigma^2} \sum_{t=c_r}^{j} (x_t - \mu_1) \tag{4.28}$$

En annulant l'équation (4.28), nous obtenons :

$$\hat{\mu}_1 = \frac{1}{j - c_r + 1} \sum_{t=c_r}^{j} x_t =: \bar{X}_j(c_r), \tag{4.29}$$

qui est la moyenne empirique des observations de c_r à j.

Ainsi l'indicateur du CUSUM Φ est égal à,

$$\hat{\Phi}_{c_r}(j) = \frac{\hat{\mu}_1 - \hat{\mu}_0}{\sigma^2} \sum_{t=c_r}^{j} \left(x_t - \frac{\hat{\mu}_1 + \hat{\mu}_0}{2} \right). \tag{4.30}$$

D'où, $\forall\ 1 \leq j \leq c_{max}$:

$$\hat{g}_j = \hat{M}_j - \hat{\Phi}_j(j). \tag{4.31}$$

avec $\hat{M}_j := \max_{1 \leq c_r \leq j} \hat{\Phi}_{c_r}(j)$.

4.2.3.2 Algorithme

Le schéma (4.6) illustre en détails la procédure de l'estimation de c_r pour une courbe de fluorescence. Soit h un seuil fixé.

$\forall\ 1 \leq j \leq c_{max}$:

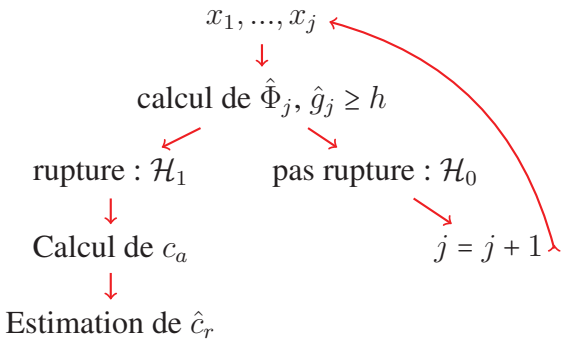


FIGURE 4.6 – Schéma illustrant l'algorithme de détection de rupture

L'algorithme du CUSUM pour la détection de rupture sur une courbe fluorescence issue de qPCR est donné par l'algorithme 1.

Dans la partie expérimentation, nous allons appliquer l'algorithme du CUSUM sur des données de fluorescence dont nous disposons. L'effet du seuil h dans la règle de décision est discuté (voir partie expérimentation 4.4.2).

4.3 Calibration

Avant de présenter les résultats obtenus par la méthode du CUSUM, nous allons montrer comment est fixé le seuil h de la règle de décision et les paramètres de la régression m et b pour l'estimation de la concentration initiale (eq. (2.6)).

4.3.1 Choix de h

Les données dont nous disposons, sont soit :

1. des fluorescences brutes sous $\mathcal{H}_0$ et pour lesquelles nous pouvons calculer la probabilité de rejeter à tord $\mathcal{H}_0$ (α). Ainsi, nous allons déterminer h en fonction de α. Nous simulons le test du CUSUM sur ces fluorescences sous $\mathcal{H}_0$ avec différentes valeurs de h et notons les valeurs de α. Puis selon les contraintes de l'utilisateur en terme de probabilité de fausse alarme, nous fixons h. Nous observerons, dans la partie expérimentation (voir 4.4.2.3), que ce paramètre h influe sur les résultats en terme de taux de bonne détection et de fausse alarme.
2. des fluorescences brutes sous $\mathcal{H}_1$ dont nous connaissons la concentration initiale Q_0 et pour lesquelles nous ne pouvons pas calculer α (aucune ou pas suffisamment de fluorescence sous $\mathcal{H}_0$).

Algorithm 1 Algorithm du CUSUM adopté pour le problème traité

Input : t_{reg}, c_{max}, h
Pré-traitement :
Filtrer les erreurs de mesures s'il n'y en a
Estimer μ_0 et σ_0 (eq. (4.22) et eq. (4.23))
Algorithme principal :
$j \leftarrow t_{reg} + 1$, $decision \leftarrow 0$
while ($decision$ == 0 and $j \leq c_{max}$) **do**
  Estimer μ_1 (eq. (4.29))
  Calculer $\hat{\Phi}_{c_r}(j)$ (eq. (4.30))
  $\hat{M}_j \leftarrow \max_{1 \leq c_r \leq j} \hat{\Phi}_{c_r}(j)$
  $\hat{g}_j \leftarrow \hat{M}_j - \hat{\Phi}_j(j)$ (eq. (4.31))
  if $g_j >= h$ **then**
    $decision \leftarrow 1$, $c_a \leftarrow j$,
    $c_0 \leftarrow \arg\max_{1 \leq t \leq c_a} \Phi_{c_a}(t)$
  else
    $j \leftarrow j + 1$ (Attendre une nouvelle observation)
  end if
end while
Post-traitement :
if $decision$ == 1 **then**
  $S\Phi_{c_a} \leftarrow spline(\Phi_{c_a}(c_0 - 2 : c_0 + 1))$ (fig. 4.4)
  $\hat{c}_r \leftarrow \arg\max S\Phi_{c_a}$
else
  $\hat{c}_r = -\infty$ (Pas de détection)
end if
Output : c_r

Dans ce cas, on choisit h en utilisant la validation croisée par *leave one out* en cherchant le h qui minimise l'erreur de la validation croisée moyenne sur l'estimation de Q_0. La validation croisée est une méthode d'estimation de fiabilité d'un modèle fondé sur une technique d'échantillonnage. Le *leave one out* (loo) est une technique de validation croisée qui en considérant n observations, on apprend sur les $n-1$ observations (apprentissage) puis on valide le modèle sur la énième observation (test). On répète n fois cette opération. Le modèle que nous voulons valider par le loo est la linéarité entre le $\log(Q_0)$ et le c_r. Nous allons le réaliser suivant l'algorithme 2.

Algorithm 2 Algorithme *leave one out cross validation*

Soit $\left(c_r^k, \log(Q_0^k)\right)_{k=1,\ldots,n}$, n observations
for $k = 1$ *to* n **do**
Enlever temporairement la k^{eme} observation $\left(c_r^k, \log(Q_0)^k\right)$ des observations
Faire la régression sur les $n-1$ observations restantes : notons $\hat{y}$ la droite de régression

Calculer erreur d'estimation du k^{eme} observation $\left(c_r^k, \log(Q_0^k)\right)$: $e_k = |\hat{y}_k - \log(Q_0^k)|$
end for
Calculer erreur de la validation crossée moyenne :

$$MSE_{loocv} = \frac{1}{n}\sum_{k=1}^{n} e_k \tag{4.32}$$

Pour le choix du seuil h, nous allons appliquer l'algorithme 2 pour différentes valeurs de h. Pour chaque valeur de h, nous estimons les c_r des fluorescences et le h choisi est celui correspondant à la plus petite erreur de validation croisée.

4.3.2 Fixation des paramètres m et b de la régression

Nous partons des fluorescences brutes dont nous connaissons leurs concentrations initiales Q_0 pour calibrer les paramètres. Ainsi, en connaissant Q_0 des fluorescences brutes, nous déterminons leurs cycles repères (c_r) pour le seuil h choisi. Grâce à la relation log linéaire entre Q_0 et c_r, nous estimons les paramètres m et b de l'équation (2.6) par la régression des moindres carrés. Sur certaines données, à la place de Q_0, nous disposons des dilutions. La dilution est un mécanisme consistant à obtenir une solution finale de concentration inférieure à celle de départ. Elle se caractérise par son taux de dilution noté D tel que $0 \leq D \leq 1$ et qui peut être exprimé par,

$$D = \frac{Q_f}{Q_0} = \frac{concentration\ finale}{concentration\ initiale}. \tag{4.33}$$

D'après l'équation (2.6), en remplaçant Q_0, nous obtenons la relation log linéaire entre le taux de dilution et le cycle repère par,

$$\log(D) = \tilde{m}\, c_r + \tilde{b}, \tag{4.34}$$

où $\tilde{m} = -m$ et $\tilde{b} = b + \log Q_f$. Ainsi, la log linéarité entre la concentration initiale et le cycle repère implique aussi la log linéarité entre la dilution et le cycle repère.

Notons $\hat{y}$ la droite de régression entre les cycles repères et la log dilution, nous avons

$$\hat{y} = \hat{m}\, c_r + \hat{b},$$

où $\hat{m}$ et $\hat{b}$ sont les estimations de $\tilde{m}$ et $\tilde{b}$.

Le résidu e et le résidu moyen du log régression ξ sont définis ci dessous,

$$e = |\hat{y} - log(D)|, \tag{4.35}$$

$$\xi = \frac{1}{n}\sum_{i=1}^{n} e_i. \tag{4.36}$$

où n est le nombre d'observations.

La quantité ξ permet d'évaluer les résultats (voir partie 4.4.2.2).

4.4 Résultats expérimentaux

Nous allons présenter les résultats expérimentaux obtenus sur la détection du cycle repère permettant d'estimer la concentration initiale d'un agent biologique.

4.4.1 Évaluer la précision des mesures

Nous désirons étudier 8 courbes de fluorescences différentes, issues des appareils de mesure de la LHVP (fig. 1.5), dont nous ne connaissons pas Q_0 et c_r. Pour calibrer les différents paramètres (h, m et b), nous avons 12 courbes de fluorescences « brutes » sous $\mathcal{H}_1$ dont nous connaissons la concentration initiale pour chacune. Nous estimons le cycle repère pour chacune de ces 12 fluorescences brutes pour différentes valeurs de h. Puis, nous utilisons la validation croisée pour fixer h (algo. 2). Nous obtenons les résultats du tableau (4.1) pour différentes valeurs de h.

La valeur de h égale à $exp(1,5)$ minimise l'erreur moyenne de la validation croisée par *leave one out* sur l'estimation de Q_0. Nous allons alors fixer le seuil h à cette valeur. Puis, grâce à la relation log linéaire entre Q_0 et c_r, nous estimations les paramètres m et b de la régression de l'équation (2.6). Ainsi, nous allons appliquer la règle du CUSUM à cette valeur de h sur les 8 fluorescences à étudier et comparer nos résultats avec ceux obtenus par le LHVP.

	$\log(h)$	1	1,25	**1,5**	1,75	2	2,25	...	3
Fluorescence brute	Q_0	c_r	c_r	c_r	c_r	c_r	c_r	...	c_r
1	10e6	25	25	25	25	25	25	...	26
2	10e6	25	25	25	26	26	26	...	26
3	10e5	27	27	27	27	27	27	...	28
4	10e5	27	27	27	27	27	27	...	28
5	10e4	31	31	31	31	32	32	...	32
6	10e4	31	31	31	31	32	32	...	32
7	10e3	38	38	38	38	38	38	...	39
8	10e3	26	35	36	36	37	37	...	39
9	10e2	38	38	38	38	38	39	...	39
10	10e2	39	39	39	39	40	40	...	40
11	10	39	39	40	40	40	40	...	40
12	10	38	38	38	39	39	39	...	39
MSE_{loocv} (eq. 4.32)		1,71	0,95	**0,93**	0,96	1,09	1,05	...	1,15
$\begin{pmatrix} m \\ b \end{pmatrix}$				$\begin{pmatrix} -0,67 \\ 30,28 \end{pmatrix}$					

TABLE 4.1 – Résultats calibrage sur les 12 courbes fluorescences brutes de la LHVP.

Par exemple, en appliquant l'algorithme du CUSUM sur une fluorescence, nous obtenons le résultat schématisé sur la figure (4.7). L'instant d'arrêt c_a est déclenché au 30^{ieme} cycle et l'instant de rupture c_r est estimé aux alentours du $28^{ieme} cycle$. Ce qui donne une retard à la détection de 2 cycles. En résumé, nous obtenons les résultats du tableau (4.2) sur les 8 fluorescences.

Comme nous pouvons le voir sur le tableau (4.2), la méthode du CUSUM donne des concentrations initiales et des cycles repères cohérents avec ceux de la méthode standard (la méthode du maximum de la dérivée seconde) utilisée par le LHVP. Sur la figure (4.8(a)), nous remarquons la linéarité entre le log de la concentration initiale obtenu par la méthode du CUSUM et le LHVP. La linéarité entre le log de la concentration initiale et les cycles repères est aussi vérifié (fig. 4.8(b)). La méthode du CUSUM fournit alors des résultats cohérents par rapport à ceux fournis par l'appareil standard du LHVP.

En plus, elle permet de minimiser le retard à la détection avec un retard à la détection moyen égale à 2 cycles qui est largement inférieur aux 15.11 cycles de

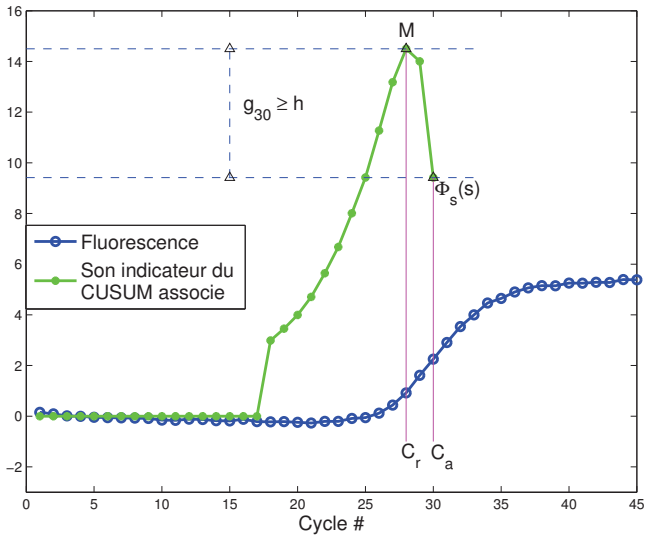
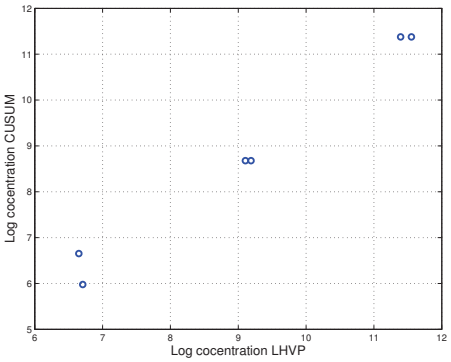


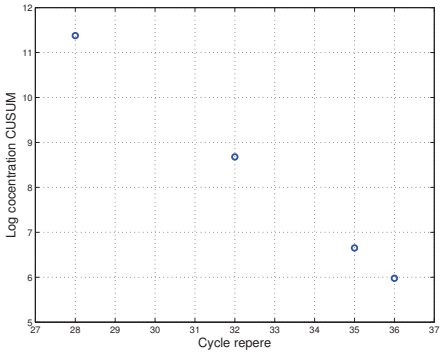
FIGURE 4.7 – Graphe du CUSUM pour une fluorescence donnée.

	LHVP (max dérivée seconde)				Ce travail (CUSUM)			
	c_r	c_a	Retard	Concentration initiale (log)	c_r	c_a	Retard	Concentration initiale (log)
1	26.67	45	18.33	11.40	28	30	2	11.38
2	26.45	45	18.55	11.55	28	30	2	11.38
3	29.85	45	15.15	9.10	32	34	2	8.68
4	29.73	45	15.27	9.19	32	34	2	8.68
5	33.27	45	11.73	6.71	36	38	2	5.98
6	33.36	45	11.64	6.65	35	37	2	6.65
7	0	...	...	0	0	...	...	0
8	0	...	...	0	0	...	...	0
			15.11				**2**	

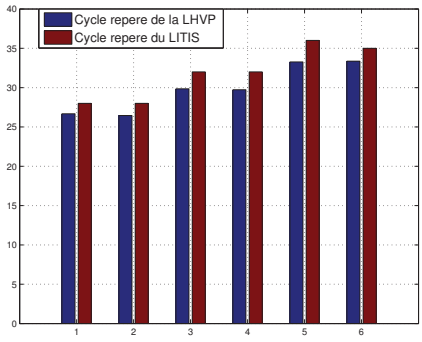
TABLE 4.2 – Résultats obtenus par le LHVP et par le LITIS.



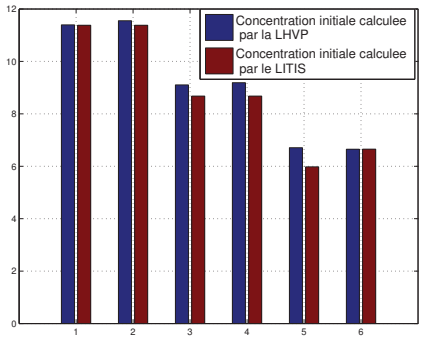
(a) $\log(Q_0)$ LHVP par rapport à $\log(Q_0)$ CUSUM



(b) $\log(Q_0)$ par rapport c_r CUSUM



(c) c_r obtenus par le LHVP et le CUSUM



(d) Q_0 obtenues par le LHVP et le CUSUM

FIGURE 4.8 – Log linéarité entre la concentration initiale Q_0 et les cycles repères et comparaison des résultats obtenus par le LHVP et par le CUSUM.

la méthode standard de l'appareil de mesure de la LHVP (tab. 4.2). Nous pouvons conclure que la méthode du CUSUM est meilleur que la méthode du maximum de la dérivée seconde utilisée par le LHVP.

4.4.2 Robustesse de la règle du CUSUM

Dans cette partie, nous analysons l'effet du seuil en terme de pourcentage de vrai positif, faux positif sur la détection, ainsi que le retard à la détection et le résidu moyen du log régression.

Les fluorescences ont été enregistrées sur un appareil standard, 7500 *Fast Real-Time PCR System of Applied Biosystems* [1]. Cet appareil donne lui-même un cycle repère par la méthode du seuil (*threshold method*, voir chapitre 1) à la fin de l'enregistrement pour chaque fluorescence.

4.4.2.1 Présentation des données

Cet appareil standard est composé d'une plaque disposée en 8 lignes et 6 paires de colonnes de puits (tab. 4.3). Des matériaux de différents types et à différentes concentrations d'ADN initiales sont analysées dans les puits.

1	1	0.33	0.33	0.11	0.11	...	Negative	Negative
1	1	0.33	0.33	0.11	0.11	...	Negative	Negative
...	...	...	...	...	...	...	...	...
1	1	0.33	0.33	0.11	0.11	...	Negative	Negative

TABLE 4.3 – Plan de la plaque pour obtenir les fluorescences.

Chaque ligne de la plaque contient un seul agent biologique testé (une seule combinaison de matériel et de fluorophore). Les colonnes 2 par 2 contiennent une même concentration spécifique et les deux dernières sont négatives, c'est à dire qu'il n'y a pas présence d'espèce biologique recherché. Sur le tableau (4.3), les chiffres 1, 0.33, 0.11, ..., de la plaque correspondent aux dilutions et chaque dilution au $1/3$ est testée en duplicata.

À noter qu'à la place des concentrations, nous disposons des valeurs de la dilution. Un ensemble de fluorescence provenant de deux plaques est traité, c'est à dire

192 signaux (160 positifs et 32 négatifs). Ainsi nous avons 32 fluorescences sous $\mathcal{H}_0$. Nous allons alors calculer la probabilité de rejeter à tord $\mathcal{H}_0$ (α) et déterminer h

1. Données fournies par Colas CALBRIX de la plateforme PRIMACEN de l'université de Rouen, F-76821 Mont Saint-Aignan, France, colas.calbrix@univ-rouen.fr

en fonction de α. La robustesse de la méthode du CUSUM sera étudiée sur les 160 fluorescences positives (fluorescences avec instant de rupture) en fonction de h.

4.4.2.2 Quantification des résultats

Nous allons définir deux concepts que nous allons utiliser dans ce qui suit pour interpréter nos résultats :

— On appelle taux de bonne détection (TP), la proportion de décider qu'il y a détection d'instant de rupture sachant que $\mathcal{H}_1$ est vraie.

— On appelle taux de fausse alarme (TF), la proportion de décider qu'il y a détection d'instant de rupture sachant que $\mathcal{H}_0$ est vraie.

Les résultats trouvés par la méthode du CUSUM seront comparés à ceux fournis par la méthode du seuil de cet appareil standard. On dira que nous avons de bons résultats si nous avons un TP élevé, un faible taux de TF et que la linéarité (l'alignement) entre le log dilution et les cycles repères soit vérifiée. Cette linéarité est qualifiée par la mesure de ξ.

Pour différentes valeurs de h en fonction de α, nous allons d'abord examiner les résultats en terme de détection par la méthode du CUSUM (vrai positif par rapport à faux positif, retard à la détection et résidu moyen du log-régression entre c_r et D). Enfin nous allons chercher la précision de la méthode du CUSUM (log-linéarité de D par rapport à c_r) et mesurer le retard à la détection uniquement sur les vrais positifs.

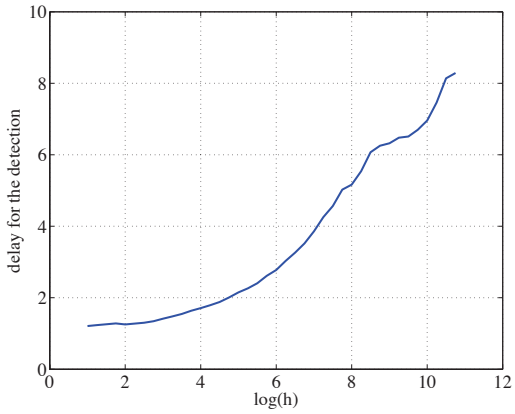
La précision de la méthode est évaluée sur la moyenne des résidus (ξ) de la log-régression entre D et c_r effectuée ligne par ligne de la plaque. Un compromis entre ces deux indicateurs (précision et retard à la détection) peut être faite à travers le paramètre h dans la méthode du CUSUM.

4.4.2.3 Résultats

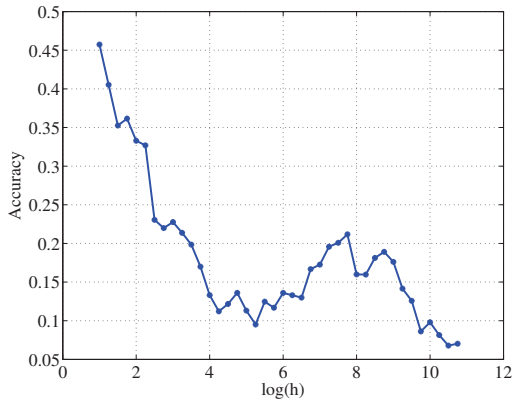
Comme mentionné précédemment, nous avons déterminé différentes valeurs du seuil h par calibrage en fonction de α (tab. 4.4). Sur ces différentes valeurs de h, nous allons analyser les taux de vrais positifs, de faux positifs et de retard à la détection moyenne. Nous obtenons les résultats représentés sur le tableau (4.4).

α	$\log(h)$	TP (%)	FP (%)	Résidu moyen du log régression (ξ)	Retard à la détection moyen
9 %	1,0	94,37	9,38	0,45	1,20
6 %	1,25	94,37	6,25	0,40	1,23
6 %	**1,5**	**94,37**	**3,12**	**0, 35**	**1,26**
6 %	1,75	94,37	3,12	0,36	1,28
6 %	2,0	93,13	3,12	0,33	1,25
3 %	2,25	92,50	3,12	0,32	1,27
3 %	2,5	91,87	3,12	0,23	1,30
3 %	2,75	91,25	3,12	0,22	1,34
3 %	3,0	90	3,12	0,23	1,41
3 %	3,25	90	3,12	0,21	1,47
3 %	3,5	90	3,12	0,20	1,54
0 %	3,75	89,38	0	0,17	1,63
0 %	4,0	88,75	0	0,13	1,71

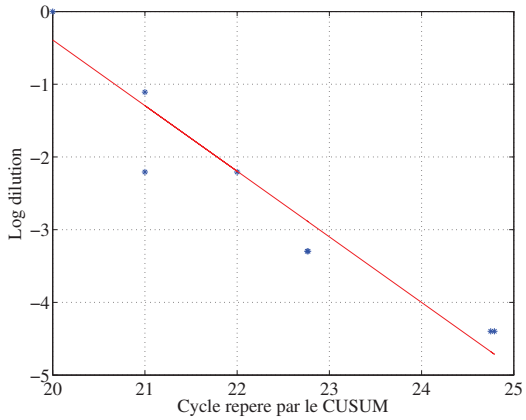
TABLE 4.4 – Robustesse obtenue par le CUSUM avec différentes valeurs de h.



(a) Retard à la détection par rapport à h



(b) Précision



(c) Cycle repère par CUSUM par rapport au log dilution pour $h = exp(1.5)$.

FIGURE 4.9 – Retard à la détection par rapport à h avec la méthode du CUSUM et résidu moyen du log-régression des cycles repères c_r/dilution par rapport à h avec la méthode CUSUM (comparé au 0.77 de l'appareil).

Le pourcentage de bonne détection augmente avec le pourcentage de fausse alarme. Lorsque le seuil h augmente, le retard à la détection devient important et la précision diminue (tab. 4.4 et fig. 4.9(b)).

L'appareil standard qui utilise la méthode du seuil, donne $93,13\%$ de taux de vrais positifs, $3,12\%$ de fausses alarmes, $13,88$ cycles de retard à la détection moyenne et $0,77$ de résidu moyen du log-régression entre les cycles repères et les concentrations initiales (tab. 4.5). Le meilleur point de fonctionnement est obtenu par la méthode du CUSUM au seuil $h = \exp(1.5)$ et de probabilité de fausse alarme $\alpha = 6\%$ [Keita et al., 2013]. À ce point on obtient $94,37\%$ de taux de vrais positifs et 3.12% de fausses alarmes. On obtient aussi $0,35$ de moyenne résiduelle du log-régression entre les cycles repères et les dilutions, et $1,26$ cycles de retard à la détection (tab. 4.5).

	Taux de vrai positif	Taux de faux positif	Retard à la détection	ξ
Référence	93,13 %	**3,12 %**	13,88 ± 0,69 cycle	0,77 ± 0,65
CUSUM	**94,37 %**	**3,12 %**	**1,26** ± 0,09 cycle	**0,35** ± 0,05

TABLE 4.5 – Comparaison des résultats obtenus par la méthode du CUSUM en ligne par rapport à la méthode du seuil de l'appareil standard (7500 fats Real-time PCR Systems of Applied Biosystems), pour $h = exp(1.5)$.

Nous présenterons schématiquement en annexe d'autres résultats sur la détermination du cycle repère par la méthode du CUSUM, notamment sur les fluorescences traitées dans le chapitre 3.

4.5 Conclusion

Dans ce chapitre, nous avons présenté la méthode du CUSUM qui est une technique statistique utilisée pour la détection de rupture (ou de changement) sur des données. C'est une méthode en ligne qui permet de minimiser le retard à la détection tout en étant précis sur la détection. En plus, le CUSUM introduit un paramètre réglable qui permet d'arbitrer entre le retard à la détection et la précision de la mesure (log-linéarité entre la concentration initiale et le cycle repère). Ce paramètre peut s'adapter alors aux différents cas d'utilisation et peut être choisi de deux façons selon la nature des fluorescences brutes dont nous disposons. La méthode du CUSUM est mieux que celles de l'état de l'art présentées sur le chapitre (1). Néanmoins, pour améliorer la méthode du CUSUM sur les fluorescences issues

de la qPCR, nous pouvons orienter la recherche en cherchant à trouver la vrai loi modélisant les observations.

Dans le chapitre 4, nous allons présenter les résultas que nous avons obtenus sur le recalage d'images. En effet, nous avons étudié la position des gouttes sur la puce par recalage d'images pour obtenir directement la fluorescence d'une goutte grâce aux images prises par une caméra à chaque cycle PCR.

Conclusion et Perspectives

CETTE THÈSE est réalisée dans le cadre du projet GENEASE. L'idée principale de ce projet est de mettre en place un système sous surveillance en continu de l'environnement qui permet de détecter la présence d'un espèce biologique recherché.

Conclusion

Nous avons présenté la réaction en chaine par polymérase (PCR). C'est une méthode biochimique qui consiste à la multiplication en chaine d'un molécule d'ADN ou de l'ARN. Il existe différentes méthodes associées à la PCR : *Reverse Transcriptase* PCR, PCR multiplex, PCR en point final et PCR quantitative. La *Reverse Transcriptase* PCR est une méthode réalisée sur un ADN complémentaire (ADNc). Elle peut être utilisée pour la construction de plusieurs ADNc et pour la construction de sondes d'ADN. La PCR multiplex permet d'amplifier plus d'un ADN avec l'utilisation de 3 amorces au moins. Elle peut être utilisée pour l'analyse des microsatellites et les polymorphismes d'un seul nucléotide. La PCR en point final et la PCR quantitative (qPCR) sont basées sur la quantification de l'agent biologique recherché. La différence est que la mesure se fait à la fin de tous les cycles pour la première. Ce qui n'est pas le cas pour la PCR quantitative appelée aussi PCR en temps réel. Un suivi à chaque cycle de la qPCR donne un signal d'émission appelé fluorescence. Cette dernière est caractérisée par un point de rupture (cycle repère) qui indique la présence ou non d'un agent biologique recherché. La position du cycle repère est fonction de la concentration. Dans le cadre de cette thèse, la qPCR sera réalisée sur une puce microfluidique pour réduire le volume de la manipulation. Les expériences biologiques seront effectuées sur cette puce en forme de gouttes qui contiennent les espèces biologiques. La puce ne fournit pas directement la fluorescence mais des images prises à chaque cycle. Ainsi, nous avons effectué un recalage d'images pour trouver les gouttes. Une fois ces dernières trouvées, nous avons obtenu la fluorescence en utilisant des indicateurs sur la goutte à chaque cycle de réaction tels que la moyenne, la médiane, l'écart type et les quartiles.

Il existe différentes techniques de l'état de l'art pour le calcul du cycle repère de la fluorescence : méthode du seuil, maximum de la dérivée seconde, méthode des points ajustées, *sigmoïde cure fitting method* et méthode de Richards. Ces différentes méthodes présentent des avantages et des inconvénients. La fluorescence peut présenter deux problématiques : présence d'erreurs de mesure et décroissance

au début de la quantification. Ces méthodes de l'état de l'art ne sont pas en général robuste à ces problématiques. Pour palier à cette décroissance que l'on peut observer au début de la fluorescence, nous avons mis en place un modèle analytique qui estime la fluorescence avec cette problématique. Ce modèle est déduit des modèles de l'état de l'art, il permet alors de s'affranchir directement du calcul du cycle repère. Puis, nous avons fait un pré-traitement pour corriger les erreurs de mesures de la fluorescence. Pour ce faire, nous avons analysé les résidus de la régression L_1 par la méthode de la boite à moustache. Ceci, nous a permis de sélectionner et supprimer ces points aberrants.

Notre modèle et les méthodes de l'état de l'art étudiés ne permettent pas de minimiser le retard à la détection. Ainsi, nous avons mis en place une méthode statistique robuste en-ligne pour déterminer un cycle repère permettant d'estimer la concentration initiale d'un espèce biologique recherché dans l'environnement. C'est la méthode des sommes cumulées, le CUSUM, qui est une technique de détection de rupture sur des données.

La méthode CUSUM proposée, alliée au pré-traitement des données, nous a permis d'obtenir un cycle repère qui permet d'estimer la quantité initiale de l'information génétique recherchée. En plus de minimiser le retard à la détection et d'être robuste aux erreurs de mesures, cette méthode introduit un paramètre réglable qui arbitre entre le délais d'alarmes et la précision de la mesure. Nous avons choisi ce paramètre par calibrage selon deux possibilités : par rapport à la probabilité de rejeter à tord l'hypothèse selon laquelle il n'y a pas de présence de l'espèce biologique recherché ou par rapport à la plus petite erreur de validation croisée par *leave one out* sur l'estimation de la concentration initiale. La méthode du CUSUM proposée, en plus de bien déterminer le cycle repère, est meilleur et plus robuste que les méthodes de l'état l'art en terme de bonne détection, de fausses alarmes et de retard à la détection.

Perspectives

Au regard de nos objectifs sur le projet GENEASE, nous avons apporté, à divers niveaux, des éléments de réponses. Dans la continuité de ce travail, plusieurs perspectives de recherche passionnantes sont envisageables. En particulier, ça serait intéressant de s'atteler sur la méthode du CUSUM en l'appliquant sur d'autres données réelles. En plus de cela, il serait intéressant de trouver la vrai loi de probabilité modélisant la fluorescence si possible précisément la vrai loi sous $\mathcal{H}_1$. Enfin, une étude plus approfondie peut être faite pour améliorer la variante du modèle analytique proposé pour analyser l'ensemble du signal de fluorescence.

Bibliographie

F. Bach, R. Jenatton, J. Mairal, and G. Obozinski. Convex optimization with sparsity-inducing norms. *Optimization for Machine Learning*, pages 19–54, 2011.

C. Barillot, J.C. Gee, L. Le Briquer, and G. Le Goualher. Fusion intra et inter individus en imagerie médicale appliquée à la modélisation anatomique du cerveau humain. *TS. Traitement du signal*, 11(6) :513–523, 1994.

M. Basseville and I.V. Nikiforov. *Detection of abrupt changes : theory and application*, volume 10. Prentice-Hall, 1993.

A. Boisbunon. *Model selection : a decision-theoretic approach.* PhD thesis, Université de Rouen, January 2013. URL `http://aurelie.boisbunon.free.fr/downloads/these.pdf`.

L.G. Brown. A survey of image registration techniques. *ACM computing surveys (CSUR)*, 24(4) :325–376, 1992.

J.D. Durtschi, J. Stevenson, W. Hymas, and K.V. Voelkerding. Evaluation of quantification methods for real-time PCR minor groove binding hybridization probe assays. *Analytical biochemistry*, 361(1) :55–64, 2007. ISSN 0003-2697.

B. Efron, T. Hastie, I. Johnstone, and R. Tibshirani. Least angle regression. *The Annals of statistics*, 32(2) :407–499, 2004.

H.A. Erlich. Polymerase chain reaction. *Journal of clinical immunology*, 9(6) : 437–447, 1989.

D. Euvrard. *Résolution numérique des équations aux dérivées partielles*, volume 988. Masson, 1994.

J. Fan and R. Li. Variable selection via nonconcave penalized likelihood and its oracle properties. *Journal of the American Statistical Association*, 96(456) : 1348–1360, 2001.

D. Fourdrinier and M.T. Wells. Comparaisons de procedures de selection d'un modele de regression : une approche decisionnelle. *Comptes rendus de l'Académie des sciences. Série 1, Mathématique*, 319(8) :865–870, 1994.

D.G. Ginzinger. Gene quantification using real-time quantitative pcr : an emerging technology hits the mainstream. *Experimental hematology*, 30(6) :503–512, 2002.

M. Guescini, D. Sisti, M.B.L. Rocchi, L. Stocchi, and V. Stocchi. A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC bioinformatics*, 9(1) :326, 2008. ISSN 1471-2105.

V. Guigue, A. Rakotomamonjy, and S. Canu. Kernel basis pursuit. In *Machine Learning : ECML 2005*, pages 146–157. Springer, 2005.

M.J. Hayden, T.M. Nguyen, A. Waterman, G.L. McMichael, and K.J. Chalmers. Application of multiplex-ready pcr for fluorescence-based ssr genotyping in barley and wheat. *Molecular breeding*, 21(3) :271–281, 2008.

A. Keita, R. Herault, S. Canu, and others. Estimation de la concentration d'un agent biologique par détection de rupture sur vidéos de fluorescences issues de pcr. 2012.

A. Keita, R. Hérault, C. Calbrix, and S. Canu. Detection and quantification in real-time polymerase chain reaction. In *European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning, Bruges : Belgium (2013)*, number 21, 2013.

M. Kubista, J.M. Andrade, M. Bengtsson, A. Forootan, J. Jonák, K. Lind, B. Sjögreen L. Strömbom A. Stahlberg Sindelka, R. Sjöback, and N. Zoric. The real-time polymerase chain reaction. *Molecular aspects of medicine*, 27(2-3) :95–125, 2006. ISSN 0098-2997.

E.L. Lehmann. *Testing statistical hypotheses*. Springer, 2005.

W. Liu and D.A. Saint. A new quantitative method of real time reverse transcription polymerase chain reaction assay based on simulation of polymerase chain reaction kinetics. *Analytical biochemistry*, 302(1) :52–59, 2002.

K.J. Livak and T.D. Schmittgen. Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods*, 25(4) :402–408, 2001. ISSN 1046-2023.

P. Markoulatos, N. Siafakas, and M. Moncany. Multiplex polymerase chain reaction : a practical approach. *Journal of clinical laboratory analysis*, 16(1) :47–51, 2002.

K.B. Mullis and F.A. Faloona. Specific synthesis of DNA in vitro via a polymerase-catalyzed chain reaction. *Methods in enzymology*, 155 :335, 1987. ISSN 0076-6879.

G. Muyzer, W. De, C. Ellen, and A.G. Uitterlinden. Profiling of complex microbial populations by denaturing gradient gel electrophoresis analysis of polymerase chain reaction-amplified genes coding for 16s rrna. *Applied and environmental microbiology*, 59(3) :695–700, 1993.

V. Noblet. *Recalage non rigide d'images cérébrales 3D avec contrainte de conservation de la topologie*. PhD thesis, Ph. D. thesis, Université Louis Pasteur-Strasbourg, 2006.

M.R. Osborne, B. Presnell, and B.A. Turlach. On the lasso and its dual. *Journal of Computational and Graphical statistics*, pages 319–337, 2000a.

M.R. Osborne, B. Presnell, and B.A. Turlach. A new approach to variable selection in least squares problems. *IMA journal of numerical analysis*, 20(3) :389–403, 2000b.

E.S. Page. Continuous inspection schemes. *Biometrika*, 41(1-2) :100–115, 1954. ISSN 0006-3444.

S.N. Peirson, J.N. Butler, and R.G. Foster. Experimental validation of novel and conventional approaches to quantitative real-time PCR data analysis. *Nucleic Acids Research*, 31(14) :e73, 2003. ISSN 0305-1048.

M.W. Pfaffl. A new mathematical model for relative quantification in real-time rt–pcr. *Nucleic acids research*, 29(9) :e45–e45, 2001.

C Ramakers, J.M Ruijter, R.H.L Deprez, and A.F.M. Moorman. Assumption-free analysis of quantitative real-time polymerase chain reaction (pcr) data. *Neuroscience letters*, 339(1) :62–66, 2003.

B.S. Reddy and B.N. Chatterji. An fft-based technique for translation, rotation, and scale-invariant image registration. *Image Processing, IEEE Transactions on*, 5 (8) :1266–1271, 1996.

SW. Roberts. Control chart tests based on geometric moving averages. *Technometrics*, 1(3) :239–250, 1959.

R.T. Rockafellar. On the maximal monotonicity of subdifferential mappings. *Pacific J. Math*, 33(1) :209–216, 1970.

S. Rosset and J. Zhu. Piecewise linear regularized solution paths. *The Annals of Statistics*, 35(3) :1012–1030, 2007.

P.J. Rousseeuw and A.M. Leroy. *Robust regression and outlier detection*, volume 589. Wiley. com, 2005.

R.G. Rutledge. Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. *Nucleic acids research*, 32(22) :e178, 2004. ISSN 0305-1048.

J.N. Sarvaiya, S. Patnaik, and S. Bombaywala. Image registration by template matching using normalized cross-correlation. In *Advances in Computing, Control, & Telecommunication Technologies, 2009. ACT'09. International Conference on*, pages 819–822. IEEE, 2009.

G. Schochetman, C.Y Ou, and W.K. Jones. Polymerase chain reaction. *The Journal of infectious diseases*, 158(6) :1154–1157, 1988.

G. Schwarz. Estimating the dimension of a model. *The annals of statistics*, 6(2) : 461–464, 1978.

W.A. Shewhart. Economic control of quality of manufactured product. *New York*, 501, 1931.

C.M. Stein. Estimation of the mean of a multivariate normal distribution. *The annals of Statistics*, pages 1135–1151, 1981.

S.M. Stigler. Gauss and the invention of least squares. *The Annals of Statistics*, 9 (3) :465–474, 1981.

C. Tse and J. Capeau. Quantification des acides nucléiques par PCR quantitative en temps réel. *Ann Biol Clin*, 61(3) :279–293, 2003.

J.W. Tukey. Exploratory data analysis. *Reading, Ma*, 231, 1977.

G. Verdier. *Détection Statistique de Rupture de Modèle dans les Systèmes Dynamiques-Application à la Supervision de Procédés de Dépollution Biologique*. PhD thesis, Université Montpellier II-Sciences et Techniques du Languedoc, 2007.

G. Wolberg and S. Zokai. Robust image registration using log-polar transform. In *Image Processing, 2000. Proceedings. 2000 International Conference on*, volume 1, pages 493–496. IEEE, 2000.

Annexes

ANNEXE A

Exemple de recalage d'images

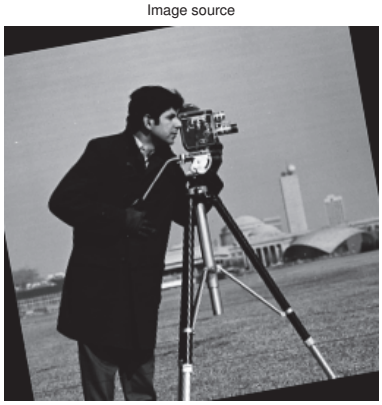
Dans ce qui suit, nous allons illustrer la technique de recalage d'images ci-dessus sur un exemples et l'étudier sur les images de la puce afin de déduire directement la fluorescence.

Exemple jouet : Nous considérons une image source d'un caméraman que nous souhaitons recaler avec une image cible.

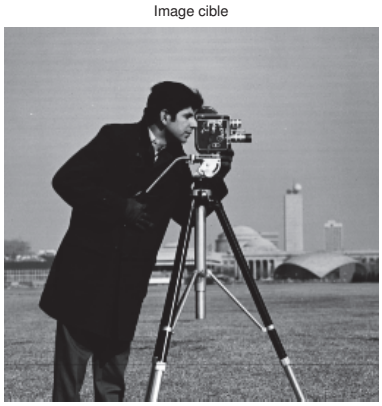
a) Avec une déformation angulaire : Voici ce qui est typiquement obtenu pour une déformation angulaire (fig. A.1), les points de correspondance (fig. A.2(a)), ainsi que le recalage obtenu (fig. A.2(c)).

b) Avec un effet d'agrandissement : même travail que précédemment (fig. A.2).

c) Avec un effet de rotation et d'agrandissement : même travail sur une combinaison de déformation (fig. A.4).

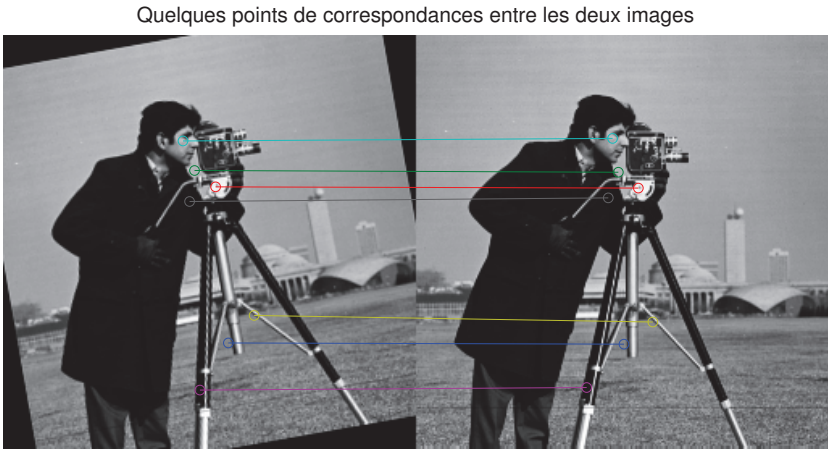


(a) Image source



(b) Image cible

FIGURE A.1 – Recalage d'une image source avec un effet de rotation par rapport à l'image cible.



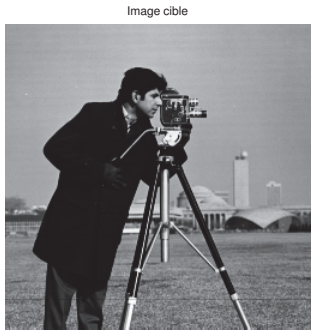
(a) Points de correspondance

(b) Superposition avant recalage

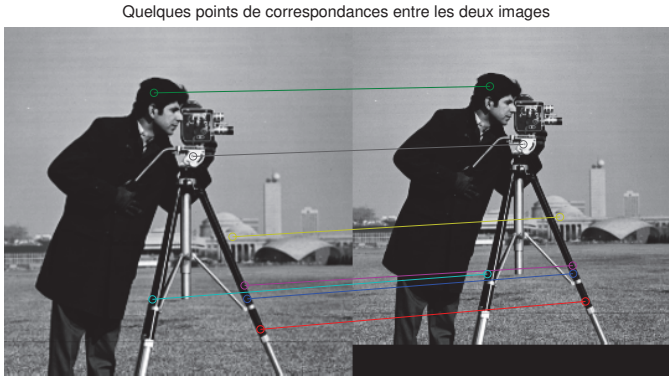
(c) Superposition après recalage

FIGURE A.2 – Résultats obtenus après le recalage.

(a) Image source



(b) Image cible

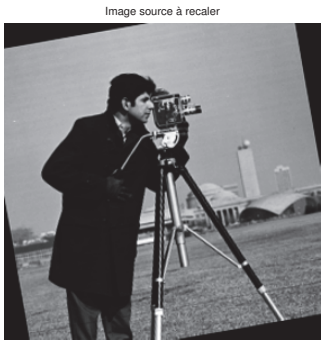


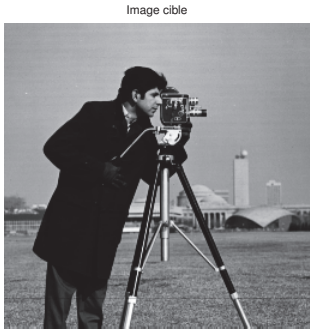
(c) Points de correspondance

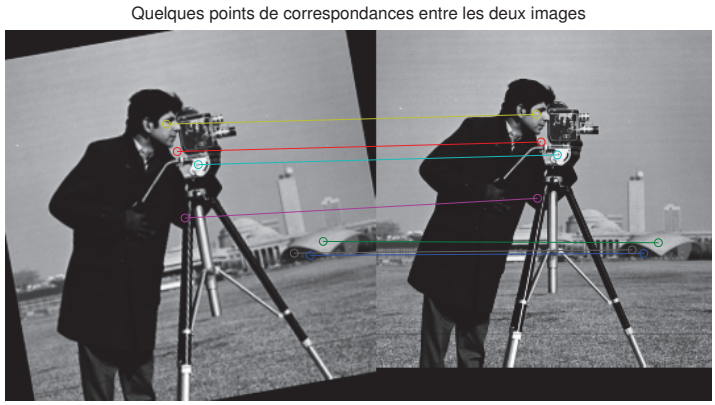
(d) Superposition avant recalage

(e) Superposition après recalage

FIGURE A.3 – Points de correspondance entre les deux images, superposition des deux images avant le recalage et résultats obtenus après le recalage.







(a) Points de correspondance

(b) Superposition avant recalage

(c) Superposition après recalage

FIGURE A.4 – Recalage d'une image source avec un effet d'agrandissement et de rotation par rapport à l'image cible et résultats obtenus après le recalage.

ANNEXE B

Détails des calculs

Régression linéaire par la méthode des moindres carrés

Soient $(x_i, y_i) \in \mathbb{R} \times \mathbb{R}$, $i = 1, \ldots, n$, où n est la taille des données. On cherche une fonction f telle que $y = f(x)$. Dans ce cas, le modèle linéaire s'écrit :

$$y_i = a_0 + a_1 x_i + \varepsilon_i \quad \forall\ i = 1, \ldots, n \tag{B.1}$$

où $\varepsilon = (\varepsilon_i)_{i=1,\ldots,n}$ est un bruit gaussien centré ; a_0 et a_1 sont des paramètres réels inconnus.

En posant $a^t = (a_0, a_1)$, l'estimateur des moindres $\hat{a}_{MC}$ de a donne le résultat suivant, Pour l'estimation de ces paramètres, on utilisera la méthode des moindres carrés (MMC) (ordinary least squares). Cette méthode consiste à minimiser l'erreur quadratique $\|\varepsilon\|_2^2 = \sum_{i=1}^{n} \varepsilon_i^2$. C'est à dire,

$$\min_{a_0,a_1} \sum_{i=1}^{n} \varepsilon_i^2 = \min_{a_0,a_1} \left[\sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2 \right] \tag{B.2}$$

Ce qui est équivalent à :

$$\min_{a_0,a_1} \frac{1}{2} \left[\sum_{i=1}^{n} (y_i - (a_0 + a_1 x_i))^2 \right] := C(a) \tag{B.3}$$

On pose $a^t = (a_0, a_1)$; $X = \begin{pmatrix} 1 & x_1 \\ \vdots & \vdots \\ 1 & x_n \end{pmatrix}$ et $y^t = (y_1, \ldots, y_n)$

En écriture simple, nous avons :

$$C(a) = \min_a \frac{1}{2}(y - Xa)^t (y - Xa) = \min_a \frac{1}{2}\|y - Xa\|^2. \tag{B.4}$$

En calculant le gradient par rapport à a, nous obtenons :

$$\nabla_a C(a) = \frac{1}{2}\left[2X^t X a - 2X^t y\right]$$

puis en l'annulant au point $\hat{a}$, qui est la solution des moindres carrés, nous obtenons :

$$\hat{a} = (X^t X)^{-1} X^t y \tag{B.5}$$

Détails du calcule de C_{y_0}

Nous calculons d'abord la dérivée première de F_c par rapport au cycle c, nous obtenons

$$\frac{dF_c}{dc} = \frac{\frac{d}{k}\, F\text{max}\; e^{-\frac{c-c_{1/2}}{k}}}{\left(1+e^{-\frac{c-c_{1/2}}{k}}\right)^{d+1}} \tag{B.6}$$

Ensuite, nous calculons la dérivée seconde,

$$\frac{d^2F_c}{dc^2} = \frac{\frac{-d}{k^2}\, F\text{max}\; e^{-\frac{c-c_{1/2}}{k}}\left(1+e^{-\frac{c-c_{1/2}}{k}}\right)^{d+1} + \frac{d(d+1)}{k^2}\, F\text{max}\left(e^{-\frac{c-c_{1/2}}{k}}\right)^2\left(1+e^{-\frac{c-c_{1/2}}{k}}\right)^{d}}{\left(1+e^{-\frac{c-c_{1/2}}{k}}\right)^{2d+2}}$$

Après simplification, nous obtenons

$$\frac{d^2F_c}{dc^2} = \frac{-d}{k^2} F\text{max}\; e^{-\frac{c-c_{1/2}}{k}}\left(\frac{1-d\, e^{-\frac{c-c_{1/2}}{k}}}{\left(1+e^{-\frac{c-c_{1/2}}{k}}\right)^{d+2}}\right) \tag{B.7}$$

$$\frac{d^2F_c}{dc^2} = 0 \Longleftrightarrow e^{-\frac{c-c_{1/2}}{k}} = \frac{1}{d}$$

En appliquant le logarithmique à gauche et à droite de l'égalité, nous obtenons

$$c = k\log(d) + c_{1/2}$$

Le point d'amplification F_{inflex} associé à ce cycle est alors,

$$F_{inflex} = F_b + \frac{F\text{max}}{\left(1+e^{-\frac{k\log(d)+c_{1/2}-c_{1/2}}{k}}\right)^{d}} = F_b + F\text{max}\left(\frac{d}{d+1}\right)^{d}$$

Les coordonnées du point d'inflexion, noté I, sont :

$$I = \left(k\log d + c_{1/2} \quad ; \quad F\text{max}\left(\frac{d}{d+1}\right)^{d} + F_b\right) \tag{B.8}$$

Le coefficient directeur m ou pente de la tangente passant par le point d'inflexion est donné par la première dérivée évaluée à la valeur de l'abscisse de ce point. Nous obtenons d'après l'équation ((B.6)) :

$$m = \frac{\frac{d}{k}\, F\text{max}\; e^{-\frac{k\log d+c_{1/2}-c_{1/2}}{k}}}{\left(1+e^{-\frac{k\log(d)+c_{1/2}-c_{1/2}}{k}}\right)^{d+1}}$$

$$m = \frac{\frac{F\text{max}}{k}}{\left(\frac{d+1}{d}\right)^{d+1}}$$

Enfin, nous avons

$$m = \frac{F\text{max}}{k}\left(\frac{d}{d+1}\right)^{d+1} \tag{B.9}$$

La droite de la tangente passant par le point d'inflexion et de pente m est,

$$(Tg) : z_c = m\ c + p,\ \forall\ c$$

où p est obtenue en remplaçant z par l'ordonnée du point d'inflexion, c par son abscisse et m par sa valeur.

Nous avons

$$F\text{max}\left(\frac{d}{d+1}\right)^{d} + F_b = \frac{F\text{max}}{k}\left(\frac{d}{d+1}\right)^{d+1}\ (k\log d + c_{1/2}) + p$$

d'où

$$p = F_b + F\text{max}\left(\frac{d}{d+1}\right)^{d}\left(1 - \frac{1}{k}\left(\frac{d}{d+1}\right)(k\log d + c_{1/2})\right) \tag{B.10}$$

Le point C_{y_0} défini ci-dessus, qui est l'intersection entre (Tg) et l'axe des abscisses, est alors égal à,

$$C_{y_0} = \frac{-p}{c}$$

$$C_{y_0} = \frac{-F_b - F\text{max}\left(\frac{d}{d+1}\right)^{d}\left(1 - \frac{1}{k}\left(\frac{d}{d+1}\right)(k\log d + c_{1/2})\right)}{\frac{F\text{max}}{k}\left(\frac{d}{d+1}\right)^{d+1}}$$

Après simplification, nous obtenons :

$$C_{y_0} = k\log(d) + c_{1/2} - k\left(\frac{d+1}{d}\right)\left(1 - \frac{F_b}{F\text{max}}\left(\frac{d+1}{d}\right)^{d}\right)$$

ANNEXE C

Régression *LASSO* à noyau

Présentation

La régression *LASSO* a été proposée par Tibshirani (1996), où l'acronyme *LASSO* signifie *Least Absolute Shrinkage and Selection Operator*. *LASSO* régularise la régression des moindres carrées avec une pénalité de la norme L_1 (voir chapitre 3). Il s'agit d'une technique de régression linéaire permettant la sélection de variable (*feature selection*). Elle peut conduire à des solutions parcimonieuses en réduisant les coefficients des variables intitules et redondantes à 0. Dans ce sens, on l'a étudié pour sélectionner les observations aberrantes. Néanmoins, on verra dans la partie application qu'elle n'est pas robuste. Nous allons d'abord présenter la méthode dans ce qui suit.

Formulation sans noyau

Dans un problème de régression, nous avons un ensemble de données sous forme de paires $\{(\mathbf{x}_i, y_i)\}_{i=1}^n \subset \mathbb{R}^d \times \mathbb{R}$, où x_i et y_i sont respectivement l'entrée et la sortie de la i^{eme} paire. La variable à expliquer $\mathbf{y} = \{y_i\}_{i=1,\dots,n}$ est mise en relation avec d variables explicatives $\mathbf{x}^1, \dots, \mathbf{x}^d$ où $\mathbf{x}^j = \{x_i^j\}_{i=1,\dots,n}$, $\forall j = 1, \dots, d$.

Une représentation du modèle linéaire d'une fonction f quelconque est la suivante :

$$f(\mathbf{x}_i) = \sum_{j=1}^{d} \beta_j x_i^j + \beta_0 \tag{C.1}$$

où $\boldsymbol{\beta} = (\beta_j)_{i=0}^d$ sont les coefficients de la fonction de régression.

L'écriture du modèle linéaire dans cette situation est alors,

$$y_i = \beta_0 + \beta_1 x_i^1 + \beta_2 x_i^2 + \cdots + \beta_d x_i^d + \varepsilon_i \quad i = 1, \dots, n \tag{C.2}$$

où $\varepsilon = (\varepsilon_i)_{i=1}^n$ un bruit centré.

Les données peuvent être rangées dans une matrice X de terme général x_i^j, dont la première colonne contient le vecteur $\mathbf{1}$ ($x_i^0 = 1$), et dans un vecteur $\mathbf{y}$ de terme

général y_i. Le modèle s'écrit alors,

$$\mathbf{y} = X\boldsymbol{\beta} + \varepsilon \tag{C.3}$$

Le problème de minimisation du *LASSO* est défini par :

$$\tilde{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta}\in R^{d+1}}{\operatorname{argmin}} \quad \frac{1}{2}\|X\boldsymbol{\beta} - \mathbf{y}\|_2^2 + \lambda\|\boldsymbol{\beta}\|_1 \tag{C.4}$$

où $\|.\|_2$ et $\|.\|_1$ sont respectivement la norme $\mathcal{L}_2$ et la norme $\mathcal{L}_1$ d'un vecteur. $\lambda > 0$ est le paramètre de régularisation : plus il est important, plus les $\tilde{\boldsymbol{\beta}}_j$, $j = 0, \ldots, d$, vont ses rapprocher de 0.

La méthode, dite *LASSO*, consiste à trouver le $\tilde{\boldsymbol{\beta}}$ défini ci dessus. Ce paramètre $\tilde{\boldsymbol{\beta}}(\lambda)$ ne peut pas être déterminé à travers d'une formule explicite en général. De ce fait, des algorithmes pour déterminer $\tilde{\boldsymbol{\beta}}(\lambda)$ existent déjà dans ce sens, notamment ceux du LARS (*Least Angle Regression*) Efron et al. [2004], de Rosset et Zhu Rosset and Zhu [2007], de Osborne, Presnell et Turlach Osborne et al. [2000a,b] et de Françis Bach Bach et al. [2011].

Formulation avec noyau

La régression *LASSO* à noyau (*Kernelized LASSO Regression*) est une méthode pour la représentation de modèles linéaires Guigue et al. [2005]. La régression *LASSO* à noyau utilise un dictionnaire $D = \{k(x, x_1), \ldots, k(x, x_n)\}$ de noyaux k.

Définition (Noyaux définis positifs) : Un noyau k défini positif est une fonction à deux variables définie,

$$k : \quad \mathcal{X} \times \mathcal{X} \longmapsto \mathbb{R} \tag{C.5}$$

$$(x, x') \longmapsto k(x, x') \tag{C.6}$$

où $\mathcal{X}$ est l'espace d'entré, vérifiant les deux propriétés suivantes :

a) symétrique :

$$\forall x, x' \in \mathcal{X}, k(x, x') = k(x', x)$$

b) positivité : $\forall n > 0$, $\forall x_1, \ldots, x_n \in \mathcal{X}$, $\forall \alpha_1, \ldots, \alpha_n$ non simultanément tous nuls,

$$\sum_{i,j=1}^{n} \alpha_i \alpha_i k(x_i, x_j) > 0$$

Examples de noyaux

a) Linéaire : $\mathcal{X} = \mathbb{R}^d$,
$$k(x, x') = x^T x'$$

b) Générique : soit une fonction caractéristique $\Phi : \mathcal{X} \longmapsto \mathcal{F}$, où $\mathcal{F}$ est un espace de Hilbert,
$$k(x, x') = \langle \Phi(x), \Phi(x') \rangle_{\mathcal{F}}$$
où $\langle ., . \rangle_{\mathcal{F}}$ est un produit scalaire défini dans l'espace $\mathcal{F}$.

Nous utilisons la régression *LASSO* à noyau pour sélectionner les erreurs de mesures qui apparaissent sur les fluorescences avant le début de la phase exponentielle lors de la qPCR. Nous allons utiliser un **noyau linéaire** et prendre $\mathcal{X} = \mathbb{R}^d$.

À partir d'un noyau k et d'un ensemble d'observation x_i, on construit la fonction suivante :
$$f(x) = \sum_{i=1}^{n} \beta_i k(x_i, x) + \beta_0 \tag{C.7}$$

Si l'on note $K = [k(x_i, x_j)]_{i=1,\ldots,n;j=1,\ldots,n}$ la matrice noyau. Le modèle s'écrit comme dans le cas linéaire,
$$\mathbf{y} = X\boldsymbol{\beta} + \varepsilon \tag{C.8}$$
où $X = [\mathbf{1} \quad K]$ une matrice avec $\mathbf{1}^T = (1, \ldots, 1)$; $\varepsilon = (\varepsilon_i)_{i=1}^n$ un bruit. Soit composant par composant :

$$\begin{bmatrix} y_1 \\ y_2 \\ \vdots \\ y_n \end{bmatrix} = \begin{bmatrix} 1 & k_{11} & k_{12} & \ldots & k_{1n} \\ 1 & k_{21} & k_{22} & \ldots & k_{2n} \\ \vdots & \vdots & \vdots & & \\ 1 & k_{n1} & k_{n2} & \ldots & k_{nn} \end{bmatrix} \times \begin{bmatrix} \beta_0 \\ \beta_1 \\ \beta_2 \\ \vdots \\ \beta_n \end{bmatrix} + \begin{bmatrix} \varepsilon_1 \\ \varepsilon_2 \\ \vdots \\ \varepsilon_n \end{bmatrix}$$

où $k_{ij} = k(x_i; x_j) \quad \forall i, j = 1, \ldots, n$.

En posant
$$Q(\mathbf{y}, X\boldsymbol{\beta}) := \frac{1}{2} \|X\boldsymbol{\beta} - \mathbf{y}\|^2$$
et
$$J(\boldsymbol{\beta}) := \|\boldsymbol{\beta}\|_1 = \sum_{j=0}^{n} |\beta_j|$$
Le problème de minimisation du *LASSO* devient :
$$\tilde{\boldsymbol{\beta}}(\lambda) = \underset{\boldsymbol{\beta} \in R^{n+1}}{\operatorname{argmin}} \quad Q(\mathbf{y}, X\boldsymbol{\beta}) + J(\boldsymbol{\beta}) \tag{C.9}$$

Vu la non dérivabilité de la fonction de pénalité $J(\boldsymbol{\beta})$, nous étudierons une technique basée sur les sous différentielles pour traiter le problème d'optimisation de la régression *LASS0* à noyau.

Résolution avec les sous différentielles.

De part la nature du problème, la solution est parcimonieuse en $\tilde{\boldsymbol{\beta}}$ (un certain nombre de $\tilde{\beta}_j$ vont être égal à $0, j = 0, \ldots, n$) du fait de la norme L_1 Efron et al. [2004] (fig. C.1).

Définition : De par la formulation du problème *LASSO*, il faudra trouver pour chaque valeur de λ son $\tilde{\boldsymbol{\beta}}(\lambda)$ associé. En outre l'approche consiste à calculer l'ensemble des solutions possibles pour toutes les valeurs de λ. On l'appelle le chemin de régularisation : $\{\tilde{\boldsymbol{\beta}}(\lambda) : 0 < \lambda < \infty\}$ (voir fig. C.1).

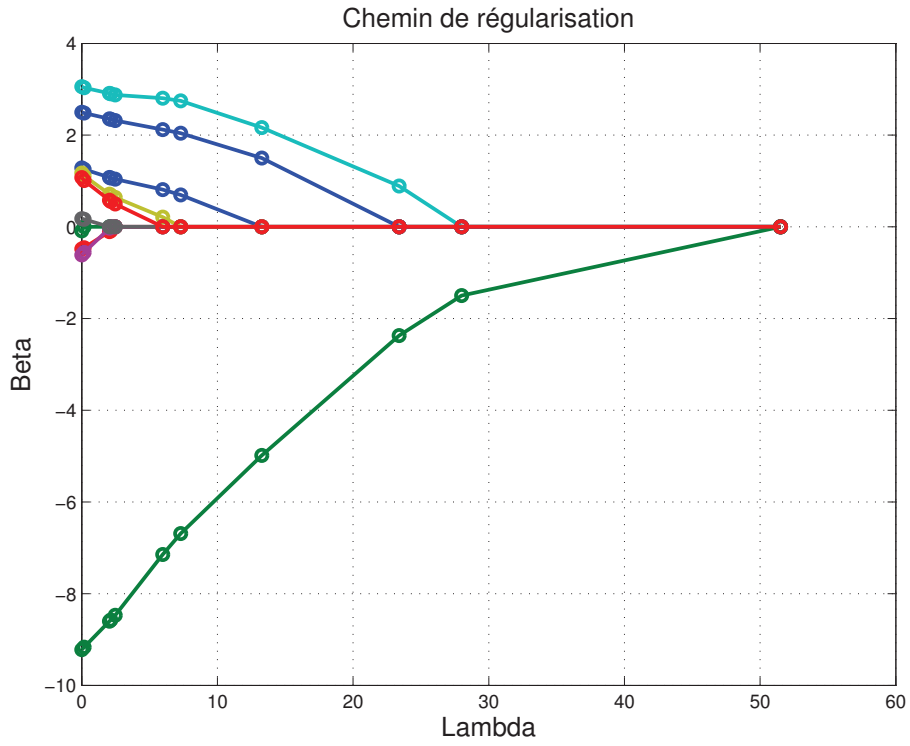


FIGURE C.1 – Exemple d'un chemin de régularisation, les paramètres $\boldsymbol{\beta}$ estimés par rapport aux λ, sur 10 variables.

Comme nous pouvons le voir sur la figure C.1, pour chaque valeur λ, seulement un certain nombre de β_j est non nul. Par exemple,

— pour $\lambda > 52$, aucune variables n'est sélectionnée

— pour $\lambda \in [28, 52]$, 2 variables sont sélectionnées

— pour $\lambda = 0$, nous avons 10 variables sélectionnées.

Pour résoudre ce problème, nous partons d'un λ grand, par exemple en l'initialisant à $\lambda_{[0]}$ au dessus duquel tous les β sont nuls ($\forall \lambda > \lambda_{[0]}$, $\tilde{\beta}_j(\lambda) = 0$). À l'étape suivante il faudra trouver $\lambda_{[1]} < \lambda_{[0]}$, où il y a un changement dans la parcimonie et son $\beta(\lambda_{[1]})$ associé ; ainsi de suite.

Optimisation : La fonction de pénalité $J(\boldsymbol{\beta})$ est non dérivable, la notion de sous différentielle sera alors utilisée pour résoudre ce problème d'optimisation Rockafellar [1970].

Définition (sous-gradient et sous-différentielle) : On dit qu'un vecteur $g \in \mathbb{R}^p$ est un sous-gradient d'une fonction convexe $f : \mathbb{R}^p \longrightarrow \mathbb{R}$ au point $\mathbf{x}_0 \in \mathbb{R}^p$ si $\forall \mathbf{x} \in \mathbb{R}^p$, l'inégalité suivante est satisfaite,

$$f(\mathbf{x}) \geq f(\mathbf{x}_0) + g^t(\mathbf{x} - \mathbf{x}_0) \tag{C.10}$$

L'ensemble de tous les sous gradients de f en $\mathbf{x}_0$ est appelé le sous-différentielle et est noté $\partial f(\mathbf{x}_0)$. Sur la figure C.2, nous avons représenté un exemple d'une fonction f convexe et non dérivable au point $x_0 = 0$. La fonction f est différentiable en $x_0 = -4$ et a un unique sous-gradient (ligne rouge). Elle n'est pas dérivable au point $x_0 = 0$ et a une infinité de sous-gradients, par exemple g_1 (ligne verte) et g_2 (ligne magenta).

Le sous gradient de la fonction à minimiser est :

$$\nabla_{\boldsymbol{\beta}}(Q(\mathbf{y}, X\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta})) = X^t(X\boldsymbol{\beta} - \mathbf{y}) + \lambda \mathbf{V}. \tag{C.11}$$

Où $\mathbf{V} = (v_j)_{j=0,\dots,d}$ tel que Boisbunon [2013]

$$v_j = \nabla_{\beta_j} J(\boldsymbol{\beta}) = \begin{cases} 1, & \text{si } \beta_j > 0 \\ -1, & \text{si } \beta_j < 0 \\ \alpha_j, & \text{si } \beta_j = 0 \text{ avec } \alpha_j \in [-1, 1] \end{cases}$$

Nous allons définir deux ensembles :

— $I_0(\lambda) := \{j / \beta_j = 0\}$, qui est l'ensemble des indices j tel que les β_j sont nuls.

— $I_{\boldsymbol{\beta}}(\lambda) := \{j / \beta_j \neq 0\}$, qui est l'ensemble des indices j tel que les β_j sont non nuls.

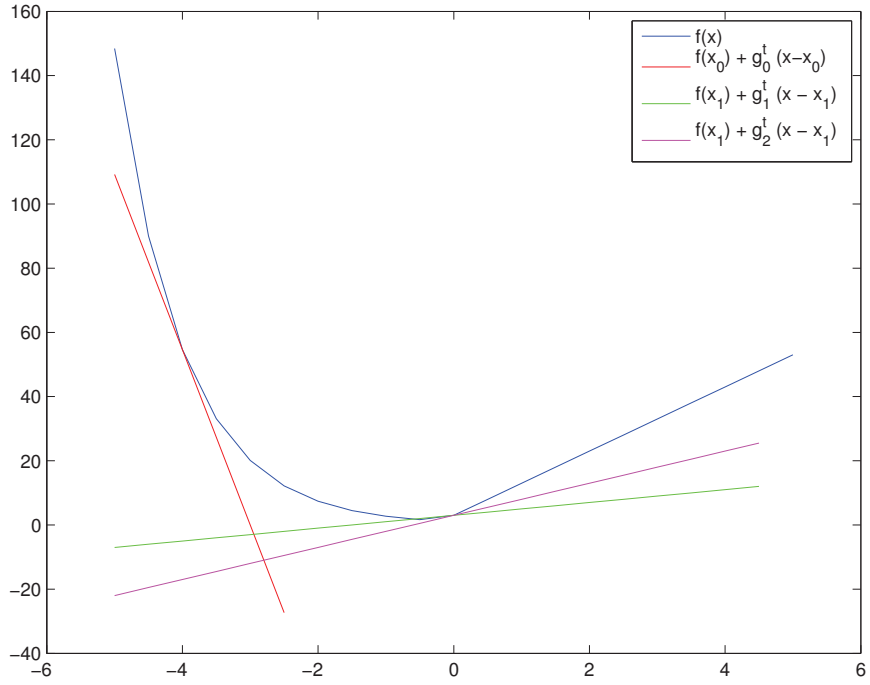


FIGURE C.2 – Exemple de sous-gradient de $f(\mathbf{x})$ pour $x_0 = -4$ et $x_0 = 0$.

L'ensemble des indices $I = \{j = 0, ..., n\}$ est alors égal à la réunion de $I_0(\lambda)$ et de I_β ($I = I_0(\lambda) \cup I_{\boldsymbol{\beta}}(\lambda)$).

A l'optimalité $0 \in \partial_{\boldsymbol{\beta}}(Q(\mathbf{y}, X\boldsymbol{\beta}) + \lambda J(\boldsymbol{\beta}))$, ce qui équivaut au système à deux équations suivantes :

$$(X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})\boldsymbol{\beta}_\beta - X_{\boldsymbol{\beta}}^t \mathbf{y} + \lambda sign(\boldsymbol{\beta}_\beta) = \vec{\mathbf{0}} \tag{C.12}$$

$$(X_0^t X_{\boldsymbol{\beta}})\boldsymbol{\beta}_\beta - X_0^t \mathbf{y} + \lambda \boldsymbol{\alpha} = \vec{\mathbf{0}} \tag{C.13}$$

Où

— $X_{\boldsymbol{\beta}} = (X_{ij})_{i=1,...,n;j \in I_{\boldsymbol{\beta}}(\lambda)}$ est la matrice X associée au $\beta_j \neq 0$, $j \in I_{\boldsymbol{\beta}}(\lambda)$. En outre $X_{\boldsymbol{\beta}}$ est une matrice de n lignes et n_β colonnes avec $n_\beta = card(I_\beta)$;

— $X_0 = (X_{ij})_{i=1,...,n;j \in I_0(\lambda)}$ est la matrice X associée au $\beta_j = 0$, $j \in I_0(\lambda)$. En outre X_0 est une matrice de n lignes et n_0 colonnes avec $n_0 = card(I_0(\lambda))$;

— $\boldsymbol{\beta}_\beta$ est un vecteur de R^{n_β} et $\boldsymbol{\alpha} \in [-1, 1]^{n_0}$.

Le système est vrai pour tout $\lambda \in (\lambda_{[N]}, \lambda_{[N+1]})$ tel que les ensembles $I_0(\lambda)$ et $I_{\boldsymbol{\beta}}(\lambda)$ sont inchangés. L'idée est de trouver la valeur du prochain λ telle que les

ensembles $I_0(\lambda)$ et $I_{\boldsymbol{\beta}}(\lambda)$ changent, ce qui se produit lorsque l'une des équations atteint sa limite. L'équation (C.13) atteint sa limite si un composant de $\boldsymbol{\alpha}$ atteint la valeur ± 1, ce qui signifie que sa composante β_j correspondante est au départ à 0 et son indice devrait entrer à l'ensemble $I_{\boldsymbol{\beta}}(\lambda)$. Dans un autre sens, l'équation (C.12) atteint sa limite quand une composante de $\boldsymbol{\beta}_\beta$ atteint la valeur 0 et par conséquent l'indice correspond va dans l'ensemble $I_0(\lambda)$. L'algorithme permet ainsi d'ajouter et de supprimer des variables selon celle de ces équations qui atteint en premier sa limite.

Illustration

Soit $\tilde{\lambda}$ proche de λ, $\tilde{\lambda}$ plus petit que λ, tel que $I_0(\lambda)$ et $I_{\boldsymbol{\beta}}(\lambda)$ restent inchangés.

L'équation (C.12) pour λ et $\tilde{\lambda}$ est

$$(X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})\boldsymbol{\beta}_\beta(\lambda) - X_{\boldsymbol{\beta}}^t \mathbf{y} + \lambda sign(\boldsymbol{\beta}_\beta) = \vec{\mathbf{0}} \tag{C.14}$$

$$(X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})\boldsymbol{\beta}_\beta(\tilde{\lambda}) - X_{\boldsymbol{\beta}}^t \mathbf{y} + \tilde{\lambda} sign(\boldsymbol{\beta}_\beta) = \vec{\mathbf{0}} \tag{C.15}$$

qui en considérant leur différence :

$$\boldsymbol{\beta}_\beta(\tilde{\lambda}) = \boldsymbol{\beta}_\beta(\lambda) - (\tilde{\lambda} - \lambda)\mathbf{w} \tag{C.16}$$

avec $\mathbf{w} := (X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})^{-1} sign(\boldsymbol{\beta}_\beta)$ qui est un vecteur de R^{n_β}.

Comment trouver $\lambda_{[N+1]}$, le paramètre de régularisation à l'étape $N+1$?

Nous supposons connaître le λ à l'étape N, on la note $\lambda_{[N]}$. Nous partons d'un λ grand au dessus duquel tous les $\boldsymbol{\beta}$ sont nuls. De ce fait $\lambda_{[N+1]} < \lambda_{[N]}$.

Soit $\tilde{\lambda}_j$ le plus petit λ possible pour la composante j tel que β_j reste dans le même état (nul ou non nul). Nous allons calculer $\tilde{\lambda}_j$ pour chaque j dans l'ensemble $I_0(\lambda)$ et $I_{\boldsymbol{\beta}}(\lambda)$.

a) Pour $j \in I_\beta$, c'est à dire β_j non nul.
D'après l'équation C.16, nous avons :

$$\beta_{\beta_j}(\lambda) = \beta_{\beta_j}(\lambda_{[N]}) - (\lambda - \lambda_{[N]})w_j \tag{C.17}$$

Soit $\hat{\lambda}_j$, le λ qui résout $\beta_{\beta_j}(\lambda) = 0$. Donc d'après l'équation(C.17), il s'écrit : $\hat{\lambda}_j = \frac{\beta_{\beta_j}(\lambda_{[N]})}{w_j} + \lambda_{[N]}$.

Et alors en respectant la condition $\lambda_{[N+1]} < \lambda_{[N]}$, nous posons

$$\tilde{\lambda}_j = \begin{cases} \hat{\lambda}_j & \text{si } \hat{\lambda}_j < \lambda_{[N]} \\ 0 & \text{sinon} \end{cases} \tag{C.18}$$

b) Pour $j \in I_0(\lambda)$, β_j est nul.

Nous avons deux cas limites possibles, $\alpha_j(\tilde{\lambda}) = 1$ ou $\alpha_j(\tilde{\lambda}) = -1$.

Soit $\hat{\lambda}_{j,1}$, le λ qui résout $\alpha_j(\lambda) = 1$ et $\hat{\lambda}_{j,-1}$, le λ qui résout $\alpha_j(\lambda) = -1$.
D'après (C.13),

$$(X_0^t X_{\boldsymbol{\beta}})_j \beta_{\beta_j}(\lambda) - (X_0^t \mathbf{y})_j + \lambda \alpha_j(\lambda) = 0 \ , \tag{C.19}$$

$$(X_0^t X_{\boldsymbol{\beta}})_j \beta_{\beta_j}(\lambda_{[N]}) - (X_0^t \mathbf{y})_j + \lambda_{[N]} \alpha_j(\lambda_{[N]}) = 0 \ . \tag{C.20}$$

(C.19) − (C.20) donne

$$(X_0^t X_{\boldsymbol{\beta}})_j (\beta_{\beta_j}(\lambda) - \beta_{\beta_j}(\lambda_{[N]})) + \lambda \alpha_j(\lambda) - \lambda_{[N]} \alpha_j(\lambda_{[N]}) = 0 \ , \tag{C.21}$$

et d'après l'équation (C.16), nous avons

$$(X_0^t X_{\boldsymbol{\beta}})_j (\lambda_{[N]} - \lambda) w_j + \lambda \alpha_j(\lambda) - \lambda_{[N]} \alpha_j(\lambda_{[N]}) = 0 \ , \tag{C.22}$$

$$\alpha_j(\lambda) = \frac{\lambda_{[N]}}{\lambda} \alpha_j(\lambda_{[N]}) - \left(\frac{\lambda_{[N]}}{\lambda} - 1 \right) (X_0^t X_{\boldsymbol{\beta}})_j w_j \ . \tag{C.23}$$

En ajoutant et en retranchant $\alpha_j(\lambda_{[N]})$, nous obtenons

$$\alpha_j(\lambda) = \alpha_j(\lambda_{[N]}) + \left(\frac{\lambda_{[N]}}{\lambda} - 1 \right) \left(\alpha_j(\lambda_{[N]}) - (X_0^t X_{\boldsymbol{\beta}})_j w_j \right) \ , \tag{C.24}$$

et en posant $z_j = \alpha_j(\lambda_{[N]}) - (X_0^t X_{\boldsymbol{\beta}})_j w_j$,

$$\alpha_j(\lambda) = \alpha_j(\lambda_{[N]}) + \left(\frac{\lambda_{[N]}}{\lambda} - 1 \right) z_j \ . \tag{C.25}$$

Au cas limites, $\alpha(\lambda) = \pm 1$, nous obtenons

$$\hat{\lambda}_{j,1} = \frac{\lambda_{[N]} z_j}{1 - \alpha_j(\lambda_{[N]}) + z_j} \ . \tag{C.26}$$

et

$$\hat{\lambda}_{j,-1} = \frac{\lambda_{[N]} z_j}{-1 - \alpha_j(\lambda_{[N]}) + z_j} \ . \tag{C.27}$$

Nous posons

$$\hat{\lambda}_j := \min(\hat{\lambda}_{j,1}, \hat{\lambda}_{j,-1}) \ , \tag{C.28}$$

et alors, comme précédemment

$$\tilde{\lambda}_j = \begin{cases} \hat{\lambda}_j & \text{si } \hat{\lambda}_j < \lambda_{[N]} \\ 0 & \text{sinon} \end{cases} \ . \tag{C.29}$$

Au final, une fois que les $\tilde{\lambda}_j$ calculés pour tous les j (C.18,C.29), il faut choisir

$$\lambda_{[N+1]} = \max_j \tilde{\lambda}_j. \tag{C.30}$$

et donc

$$\tilde{j} = \underset{j}{\text{argmax}} \quad \tilde{\lambda}_j \tag{C.31}$$

NB : Une indice $j \in I_0(\lambda)$ passe dans l'ensemble I_β équivaut à $\alpha_j = \pm 1$, donc $sign(\boldsymbol{\beta}_\beta)_j = \alpha_j$.

— Si $\tilde{j} \in I_0(\lambda)$ alors pour $\lambda \in [\lambda_{[N+2]}; \lambda_{[N+1]}]$, $\beta_{\tilde{j}}$ devient non nul, alors $\tilde{j}$ passe dans l'ensemble I_β $(\tilde{j} \to I_\beta)$.

— Si $\tilde{j} \in I_\beta$ alors pour $\lambda \in [\lambda_{[N+2]}; \lambda_{[N+1]}]$, $\beta_{\tilde{j}}$ devient nul, alors $\tilde{j}$ passe dans l'ensemble I_0 $(\tilde{j} \to I_0(\lambda))$.

Proposition C.0.1. *Il existe $\varepsilon > 0$, tel qu'à l'étape 1, avec $\lambda_{[1]} = \lambda_{[0]} - \varepsilon$ où $\lambda_{[0]} = \max_j |(X^t\mathbf{y})_j|$, la solution admet $I_\beta^{[1]} = \{\tilde{j}\}$ avec $\tilde{j} = \underset{j}{\text{argmax}} |(X^t\mathbf{y})_j|$.*

Démonstration. Soit $k \in I_\beta^{[1]}$, quelque soit un paramètre de régularisation λ tel que $I_\beta^{[1]}$ reste inchangé, l'équation (C.14) donne :

$$(X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})_k \boldsymbol{\beta}_{\beta_k} - (X_{\boldsymbol{\beta}}^t \mathbf{y})_k + \lambda sign(\boldsymbol{\beta}_{\beta_k}) = 0$$

Soit

$$(X_{\boldsymbol{\beta}}^t \mathbf{y})_k = (X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})_k \boldsymbol{\beta}_{\beta_k} + \lambda sign(\boldsymbol{\beta}_{\beta_k})$$

En mettant les valeurs absolue, nous obtenons

$$|(X_{\boldsymbol{\beta}}^t \mathbf{y})_k| = (X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})_k |\boldsymbol{\beta}_{\beta_k}| + \lambda |sign(\boldsymbol{\beta}_{\beta_k})|$$

car $\lambda \geq 0$, $X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}} = \|X_{\boldsymbol{\beta}}\|^2 \geq 0$ et β_{β_k} est soit positif ou négatif, donc les deux termes après l'égalité sont de même signe. D'où,

$$|(X_{\boldsymbol{\beta}}^t \mathbf{y})_k| = (X_{\boldsymbol{\beta}}^t X_{\boldsymbol{\beta}})_k |\boldsymbol{\beta}_{\beta_k}| + \lambda$$

alors,

$$|(X^t_{\boldsymbol{\beta}}\mathbf{y})_k| \geq \lambda$$

or $\lambda_{[0]} = \max_j |(X^t\mathbf{y})_j|$ et $\infty > \lambda_{[0]} > \lambda_{[1]} > ... > \lambda_{[d]} > 0$, $d \geq n$, donc

$$|(X^t_{\boldsymbol{\beta}}\mathbf{y})_k| = \lambda_{[0]} \text{ et } k = \tilde{j}$$

□

Initialisation

Nous partons d'une valeur de λ grande, $\lambda_{[0]} \leftarrow \max \|X^t\mathbf{y}\|_\infty$, en mettant tous les indices dans l'ensemble $I_0(\lambda)^{[0]}$ et alors $I_\beta^{[0]}$ est vide. En outre les β sont nuls au début, $\beta^t_{[0]} = (0, ..., 0)$. Notons $\tilde{j} := \underset{j}{\operatorname{argmax}} |(X^t\mathbf{y})_j|$.

La proposition ci-dessous, nous permet de connaître les indices des ensembles $I_0(\lambda)$ et $I_{\boldsymbol{\beta}}(\lambda)$ à l'étape 1 et mettre à jour les matrices X_0 et $X_{\boldsymbol{\beta}}$.

Remarque

D'après la proposition, nous avons : $I_\beta^{[1]} = \{\tilde{j}\}$, $I_0^{[1]} = \{1, \ldots, n\} - \{\tilde{j}\}$. Les paramètres $\boldsymbol{\beta}$, à l'initialisation, sont toujours nuls donc $\beta_{\tilde{j}} = 0$. L'indice $\tilde{j}$ quitte l'ensemble $I_0(\lambda)$ pour appartenir à l'ensemble I_β donc $\alpha_{\tilde{j}} = +1$ ou $\alpha_{\tilde{j}} = -1$ (voir ci dessus).

Algorithme

Ce qui nous permet de déduire l'algorithme suivant pour trouver le chemin de régularisation.

1. $I_0(\lambda) := \{j = 1, \ldots, n\}$.
   $I_\beta = I_0(\lambda)^c = \varnothing$, nous mettons tous les $\boldsymbol{\beta}$ à zéro.
2. Initialisation : étape $N \longleftarrow 0$
   $\lambda_{[N]} \leftarrow \max \|X^t\mathbf{y}\|_\infty$.
   $\tilde{j} \leftarrow \underset{j}{\operatorname{argmax}} |(X^t\mathbf{y})_j|$.
   $\boldsymbol{\alpha} = \frac{1}{\lambda_{[N]}} X^t y$.
   $sign(\beta_{\tilde{j}}) = \alpha_{\tilde{j}}$.
   $I_\beta = \{\tilde{j}\}$ et $I_0(\lambda) = \{1, \ldots, n\} - \{\tilde{j}\}$.
   Mettre à jour X_β et X_0.
   $\mathbf{w} \leftarrow (X^t_\beta X_\beta)^{-1} sign(\beta_{\tilde{j}})$.

$\boldsymbol{\alpha}(\lambda_{[N]}) \leftarrow \frac{1}{\lambda_{[N]}}(X_0^t \mathbf{y})$.
$\mathbf{z} = \boldsymbol{\alpha}(\lambda_{[N]}) - (X_0^t X_{\boldsymbol{\beta}})\mathbf{w}$.

3. Tant que $I_0(\lambda)$ **est non vide** Faire

   — Trouver l'étape suivante du chemin de régularisation $\lambda_{[N+1]}$: se servir des équations de (C.18) à (C.30).

   — Mettre à jour $I_0(\lambda)$ et $I_{\boldsymbol{\beta}}(\lambda)$.

   — Mettre à jour X_β et X_0 ; résoudre avec ce lambda λ à l'étape $N+1$ pour trouver les $\boldsymbol{\beta}_\beta$
   $\boldsymbol{\beta}_\beta(\lambda_{[N+1]}) \leftarrow (X_\beta^t X_\beta)^{-1}(X_\beta^t \mathbf{y} - \lambda_{[N+1]} sign(\boldsymbol{\beta}_\beta))$
   $\mathbf{w} \leftarrow (X_\beta^t X_\beta)^{-1} sign(\boldsymbol{\beta}_\beta))$
   $\boldsymbol{\alpha}(\lambda_{[N+1]}) \leftarrow \frac{1}{\lambda_{[N+1]}}(X_0^t \mathbf{y} - X_0^t X_\beta \boldsymbol{\beta}_\beta(\lambda_{[N+1]}))$
   $\mathbf{z} = \boldsymbol{\alpha}(\lambda_{[N+1]}) - (X_0^t X_{\boldsymbol{\beta}})\mathbf{w}$
   $N \longleftarrow N+1$

4. Fin Tant que

Nous souhaitons sélectionner les variables les plus influentes sur la variable d'étude $\mathbf{y}$. Ceci équivaut à la recherche du sous ensemble J inclus dans $I = \{1, \ldots, n\}$ tels que l'on puisse réduire le modèle (C.8) à :

$$\mathbf{y} = X_J \boldsymbol{\beta}_J + \varepsilon \tag{C.32}$$

où X_J est la matrice de X réduite aux variables de J.

Sélection de modèle : Evaluer l'estimateur du *LASSO* À chaque paramètre de régularisation λ, est associé un vecteur de coefficients $\boldsymbol{\beta}(\lambda)$. Il est alors nécessaire de trouver le λ optimal ($i.e$ trouver le bon $\hat{\boldsymbol{\beta}}$ estimateur de $\boldsymbol{\beta}$) pour bien représenter le modèle, par :

$$\lambda_{opt} = \underset{\lambda}{\operatorname{argmin}} \quad critere(\lambda) \tag{C.33}$$

où $critere$ est une fonction coût qui dépend de λ.

Pour ce faire, nous utilisons une fonction de coût quadratique définie comme suit,

$$critere(\lambda) := Q_\lambda(X\boldsymbol{\beta}, X\hat{\boldsymbol{\beta}}) = \|X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}}(\lambda)\|^2 \tag{C.34}$$

D'où

$$\lambda_{opt} = \underset{\lambda}{\operatorname{argmin}} \quad \|X\boldsymbol{\beta} - X\hat{\boldsymbol{\beta}}(\lambda)\|^2 \tag{C.35}$$

Le problème qui se pose est que l'on ne connaît pas à priori le paramètre $\boldsymbol{\beta}$. De ce fait, il faudra estimer le coût pour un estimateur $\hat{\boldsymbol{\beta}}$ spécifié. Cette procédure est

semblable à l'estimation sans biais du risque, introduite par Stein en 1981 Stein [1981] dans le cadre de l'estimation du paramètre de position de la loi normale. L'estimation de coût comme sélecteur de variables a été développée par Fourdrinier and Wells [1994]. D'autres critères pour évaluer la sélection d'un modèle existe aussi, par exemple le C_p Mallow's, l'*Akaike Information Criterion* (AIC) Fan and Li [2001] ou le critère d'information bayésien (BIC) Schwarz [1978]. Ces critères ont fait leurs preuves lorsque la distributions des erreurs est définie, mais il est plus difficile de les estimer dès que l'on relâche cette hypothèse. Par ailleurs, l'AIC a tendance à sélectionner des modèles trop complexe et le BIC des modèles trop simples Boisbunon [2013].

Définition (estimateur non biaisé) : Soit Z un vecteur de variables aléatoires dans $\mathbb{R}^d$ de moyenne $\theta \in \mathbb{R}^d$ et soit $\hat{\theta} \in \mathbb{R}^d$ un estimateur de θ. Un estimateur $\hat{Q}_0(\theta)$ du coût $Q(\theta, \hat{\theta})$ est dit non biaisé si est seulement si pour tout $\theta \in \mathbb{R}^d$, il satisfait la condition suivante

$$\mathbb{E}_\theta[\hat{Q}_0(\theta)] = \mathbb{E}_\theta[Q(\theta, \hat{\theta})] =: R(\theta, \hat{\theta}) \tag{C.36}$$

où $\mathbb{E}_\theta$ représente l'espérance par rapport à la distribution de Z. R défini le risque de $\hat{\theta}$ à θ.

Théorème C.0.1. *(Estimateur non biaisé du coût quadratique sous les hypothèses gaussiennes). Soit* $\mathbf{y} \sim \mathcal{N}_n(X\beta, \sigma^2 I_n)$. *Soit* $\hat{\boldsymbol{\beta}}$ *un estimateur de* $\boldsymbol{\beta}$ *tels que* $X\hat{\boldsymbol{\beta}}$ *soit faiblement différentiable par rapport à* $\mathbf{y}$ *et soit* $\hat{\sigma}^2$ *un estimateur sans biais de* σ^2 *indépendant de* $div_{\mathbf{y}}(X\hat{\boldsymbol{\beta}})$ *(divergence de* $X\hat{\boldsymbol{\beta}}$ *par rapport à* $\mathbf{y}$*). Alors*

$$\hat{Q}_0(\boldsymbol{\beta}) = \|\mathbf{y} - X\hat{\boldsymbol{\beta}}\|^2 + (2div_{\mathbf{y}}(X\hat{\boldsymbol{\beta}}) - n)\hat{\sigma}^2 \tag{C.37}$$

où $div_{\mathbf{y}}(X\hat{\boldsymbol{\beta}})$ est la fonction divergence de $X\hat{\boldsymbol{\beta}}$ par rapport à $\mathbf{y}$ défini par,

$$div_{\mathbf{y}}\ (X\hat{\boldsymbol{\beta}}) = \sum_{i=1}^{n} \frac{\partial (X\hat{\beta})_i}{\partial y_i} \tag{C.38}$$

Démonstration. La preuve du thèorème C.0.1 se trouve dans Boisbunon [2013] page 64. □

Exemples sur des données jouets et sur 3 fluorescences avec erreurs de mesures

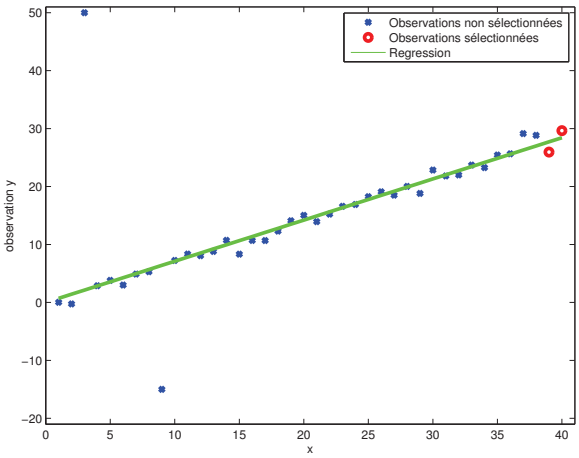
Nous allons appliquer la méthode de la boîte à moustaches et du *LASSO* sur des données artificielles avec des erreurs de mesures.

Générer les données jouets (linéaires) : Soient $\mathbf{x}$ ($x_i = i, \forall i = 1, \ldots, n$) et $\mathbf{y}$ deux vecteurs à n composants tels que,

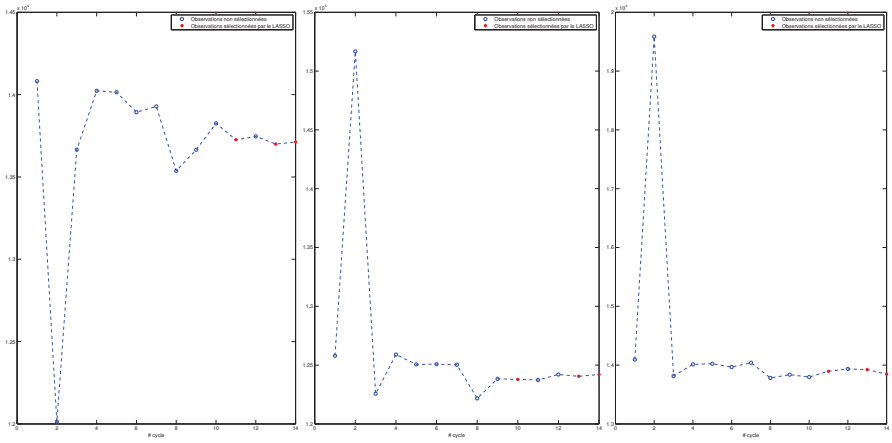
$$\mathbf{y} = a\mathbf{x} + b + \varepsilon \tag{C.39}$$

où a et b sont deux coefficients réels et $\varepsilon \sim \mathcal{N}_n(0, \sigma^2 I_n)$ un bruit gaussien.

Nous obtenons par exemples des données représentées sur la figure C.3(a). Nous voulons détecter les observations (variables) non influentes de $\mathbf{y}$ en appliquant la méthode du *LASSO* à noyau. En appliquant la méthode du *LASSO* à noyau (fig. C.3(a)), deux variables (39 et 40) sont sélectionnées pour faire la régression. Donc, les observations 39 et 40 sont plus pertinentes que les autres. Les autres observations ne sont pas sélectionnées par le *LASSO*. On ne peut pas conclure que les observations 3 et 9 sont des erreurs de mesures. On peut faire la même analyse pour les 3 fluorescences avec la présences d'erreurs de mesures (fig. C.3(b)). Cette méthode n'est pas alors robuste d'où l'analyse des résidus de la régression L_1 et L_2 par la méthode de la boite à moustache vu au chapitre 3.



(a) Résultats avec la méthode du *LASSO* à noyau sur des innées jouets.



(b) Résultats avec la méthode du *LASSO* à noyau sur des 3 fluorescences.

FIGURE C.3 – Résultats

ANNEXE D

Modèle auto-regressif

La phase initiale de la fluorescence peut être considérée comme un modèle auto-régressif. Ainsi, nous allons l'appliquer sur les premières observations de la fluorescence. Avant de montrer le résultat obtenu, nous allons faire un rappel sur le auto-régressif.

Considérons le modèle auto-régressif (AR(1)) suivant :

$$y_{t+1} = y_t.A + B \qquad \forall\ 1 \leq t \leq n \tag{D.1}$$

Où y_t est un vecteur ligne de n_c caractéristiques ($y_t \in \mathbb{R}^{n_c}$), A est une matrice carrée de taille n_c ($A \in \mathcal{M}_{n_c,n_c}(\mathbb{R})$) et $B \in \mathbb{R}^{n_c}$.

Nous allons aborder ce problème en traitant deux cas : le cas où $A \neq I$ et $A = I$ ou I est la matrice identité.

1. Cas où $A \neq I$:
   Si $(I - A)$ est inversible et en posant $R = B.(I - A)^{-1}$, nous avons :

$$y_t = (y_0 - R).A^t + R \qquad \forall\ t \tag{D.2}$$

Démonstration :

Soit un vecteur $R \in \mathbb{R}^{n_c}$ tel que

$$R = R.A + B$$

Nous avons alors

$$R - R.A = B$$

ie

$$R.(I - A) = B$$

Si $(I - A)$ est inversible, nous avons :

$$R = B.(I - A)^{-1} \tag{D.3}$$

Soit x_t un nouveau vecteur caractéristique tel que

$$x_{t+1} = y_{t+1} - R \qquad \forall\ t \tag{D.4}$$

L'équation (eq. (D.1)) devient :

$$x_{t+1} + R = (x_t + R).A + B$$

ie (en remplaçant B)

$$x_{t+1} = (x_t + R).A + R.(I - A) - R$$

Après développement et simplification, nous savons

$$x_{t+1} = x_t.A$$

Nous avons alors une suite géométrique de raison x_0, où x_0 est la première observation des X. Nous avons alors,

$$x_t = x_0.A^t \qquad \forall\ 1 \leq t \leq n \tag{D.5}$$

Nous allons maintenant retrouver l'équation en Y. Nous avons d'après l'équation (D.2),

$$x_0.A^t = y_t - R$$

ie

$$(y_0 - R).A^t = y_t - R$$

D'où

$$y_t = (y_0 - R).A^t + R \qquad \forall\ t \tag{D.6}$$

2. Cas où $A = I$:
   Dans cette situation, l'équation (D.1) devient :

$$y_{t+1} = y_t + B \qquad \forall\ 1 \leq t \leq n \tag{D.7}$$

Nous avons alors une suite arithmétique. Nous obtenons,

$$y_t = y_0 + tB \qquad \forall\ 1 \leq t \leq n \tag{D.8}$$

L'étape qui suit est d'estimer les paramètres inconnus A et B du modèle. D'abord, nous allons modifié légèrement le modèle (D.1) en ajoutant un bruit blanc ε_t par,

$$y_{t+1} = y_t.A + B + \varepsilon_t \qquad \forall\ 1 \leq t \leq n \tag{D.9}$$

Estimation de A et B :

Pour estimer A et B, nous allons utiliser une descente de gradient en minimisant le bruit comme suit,

$$\min_{A,B} \quad \frac{1}{2}\|\varepsilon\|^2 \coloneqq \min_{A,B} \quad \frac{1}{2}\sum_{t=1}^{n} \varepsilon_t.\varepsilon_t^T \tag{D.10}$$

Ce qui revient à

$$\min_{A,B} \quad \frac{1}{2}\sum_{t=1}^{n}(y_{t+1} - y_t.A - B).(y_{t+1} - y_t.A - B)^T \tag{D.11}$$

Posons

$$J(A,B) \coloneqq \sum_{t=1}^{n} J_t(A,B)$$

Où

$$J_t(A,B) \coloneqq \frac{1}{2}(y_{t+1} - y_t.A - B).(y_{t+1} - y_t.A - B)^T$$

En résumé, nous avons le problème d'optimisation sans contrainte suivant,

$$\min_{A,B} J(A,B) \tag{D.12}$$

Nous calculons les dérivées partielles de J par rapport aux inconnus A et B. Nous avons

$$\frac{\partial J(A,B)}{\partial A} = \sum_{t=1}^{n} \frac{\partial J_t(A,B)}{\partial A} \quad \forall\, t$$
$$\frac{\partial J(A,B)}{\partial B} = \sum_{t=1}^{n} \frac{\partial J_t(A,B)}{\partial B} \quad \forall\, t$$

Ces dérivées partielles ne sont pas facile à calculer surtout la première car A est une matrice. Pour ce faire, nous allons faire la dérivée par rapport aux coefficients de la matrice A et du vecteur B.
Nous avons

$$\frac{\partial J_t(A,B)}{\partial A} = \left[\frac{\partial J_t(a_{ij},b_j)}{\partial a_{ij}}\right]_{i,j=1}^{n_c} \quad \forall\, t \tag{D.13}$$

$$\frac{\partial J_t(A,B)}{\partial B} = \left\{\frac{\partial J_t(a_{ij},b_j)}{\partial b_j}\right\}_{j=1,\ldots,n_c} \quad \forall\, t \tag{D.14}$$

où $A = \left[a_{ij}\right]_{i,j=1}^{n_c}$ et $B = \left\{b_j\right\}_{j=1,\ldots,n_c}$.
Avant d'entamer les calculs, détaillons la quantité $J_t(A,B)$. Sur cette dernière, la matrice A est multipliée par le vecteur y_t, $\forall t$. De ce fait, nous avons

$$y_t.A = \left\{\sum_{i=1}^{n_c} y_t^i.a_{ij}\right\}_{j=1,\ldots,n_c} = \left(\sum_{k=1}^{n_c} y_t^k.a_{k1},\ldots,\sum_{k=1}^{n_c} y_t^k.a_{kj},\ldots,\sum_{k=1}^{n_c} y_t^k.a_{kn_c}\right) \quad \forall\, t \tag{D.15}$$

et

$$y_{t+1} - y_t.A - B = \left\{ y^j_{t+1} - \sum_{i=1}^{n_c} y^i_t.a_{ij} - b_j \right\}_{j=1,\dots,n_c} \quad \forall\ t \tag{D.16}$$

où $y_t = (y^1_t, \dots, y^{n_c}_t)$.

Les coefficients b_j ne sont multipliés que par la valeur -1, nous avons alors, $\forall\ j = 1, \dots, n_c$

$$\frac{\partial J_t(a_{ij}, b_j)}{\partial b_j} = \left(y^j_{t+1} - \sum_{k=1}^{n_c} y^k_t.a_{kj} - b_j \right) := \Gamma_j \quad \forall\ t \tag{D.17}$$

En se servant de l'équation suivante :

$$\frac{\partial a_{kl}}{\partial a_{ij}} = \delta_{ik}.\delta_{jl} \tag{D.18}$$

où

$$\delta_{ik} = \begin{cases} 1 & \text{si } i = k \\ 0 & \text{sinon.} \end{cases}$$

Alors,

$$\frac{\partial a_{kl}}{\partial a_{ij}} = \begin{cases} 1 & \text{si } i = k \quad \text{et} \quad j = l \\ 0 & \text{sinon.} \end{cases} \tag{D.19}$$

Nous avons,

$$\frac{\partial \left(y_{t+1} - y_t.A - B \right)}{\partial a_{ij}} = \frac{\partial \left(y^1_{t+1} - \sum_{k=1}^{n_c} y^k_t.a_{k1} - b_1, \dots, y^j_{t+1} - \sum_{k=1}^{n_c} y^k_t.a_{kj} - b_j, \dots \right)}{\partial a_{ij}} \tag{D.20}$$

$$\frac{\partial \left(y_{t+1} - y_t.A - B \right)}{\partial a_{ij}} = \left(0, \dots, -y^i_t, \dots, 0 \right) \tag{D.21}$$

D'où

$$\frac{\partial J_t(a_{ij}, b_j)}{\partial a_{ij}} = -y^i_t \left(y^j_{t+1} - \sum_{k=1}^{n_c} y^k_t.a_{kj} - b_j \right) \quad \forall\ t \tag{D.22}$$

$$= -y^i_t \left(y^j_{t+1} - \sum_{k=1}^{n_c} y^k_t.a_{kj} - b_j \right) = -y^i_t \Gamma^t_j := \Upsilon^t_{ij} \quad \forall\ t \tag{D.23}$$

Finalement, nous obtenons

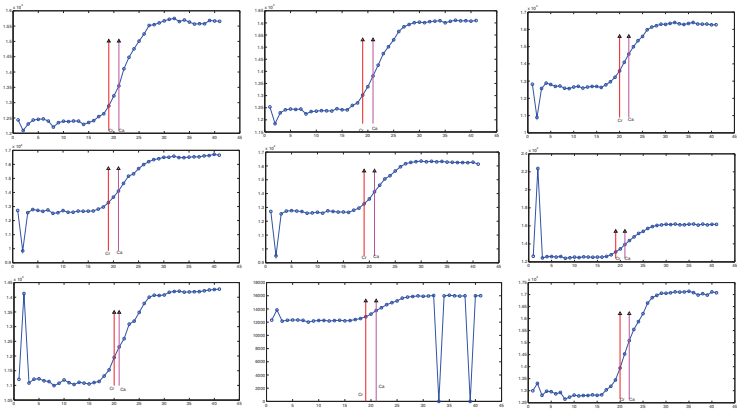
$$\frac{\partial J(A, B)}{\partial A} = \sum_{t=1}^{n} \left[\Upsilon^t_{ij} \right]^{n_c}_{i,j=1} = \left[\sum_{t=1}^{n} \Upsilon^t_{ij} \right]^{n_c}_{i,j=1} \quad \forall\ t \tag{D.24}$$

$$\text{et} \quad \frac{\partial J(A, B)}{\partial B} = \sum_{t=1}^{n} \left\{ \Gamma^t_j \right\}_{j=1,\dots,n_c} = \left\{ \sum_{t=1}^{n} \Gamma^t_j \right\}_{j=1,\dots,n_c} \quad \forall\ t \tag{D.25}$$

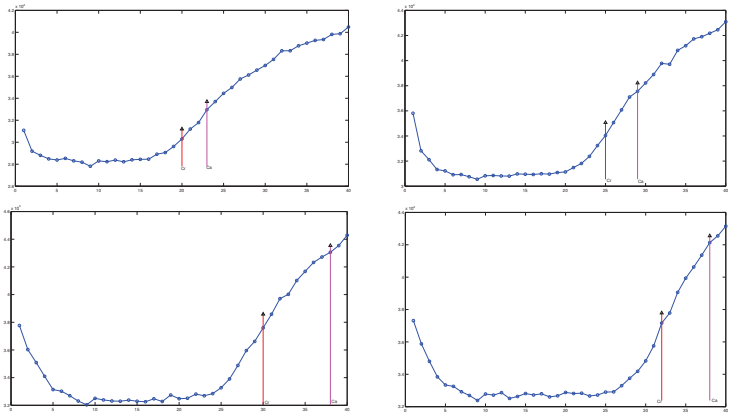
ANNEXE E

Figures complémentaires

Nous allons présenter quelques figures permettant d'illustrer les résultats obtenus par les méthodes que nous avons utiliser dans le cadre de cette thèse. La figure E.1 montre des résultats de l'algorithme CUSUM appliqué aux fluorescences avec des difficultés complémentaires.



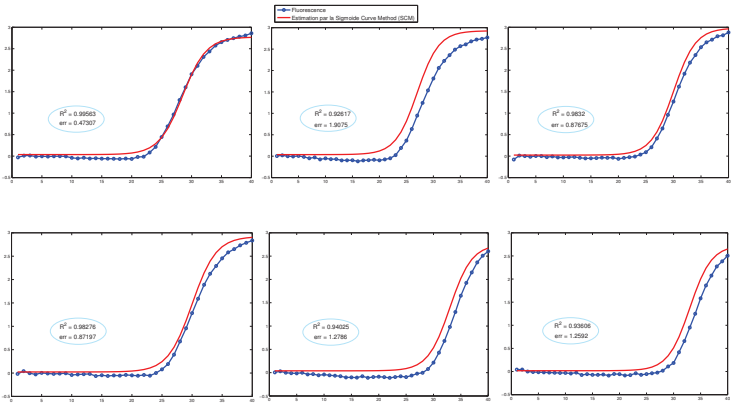
(a) CUSUM appliqué à 9 fluorescescences avec des erreur de mesures



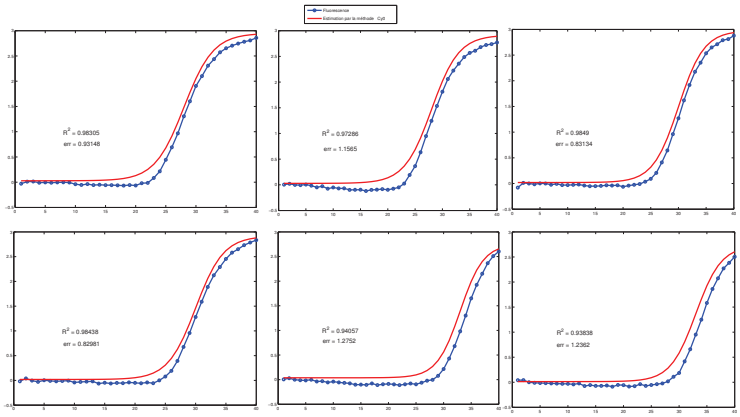
(b) CUSUM appliqué à 4 fluorescescences avec une décroissance au début de l'amplification

FIGURE E.1 – Illustration de la méthode du CSUUM appliquée sur des fluorescences avec erreurs de mesures et une décroissance au début de la quantification PCR.

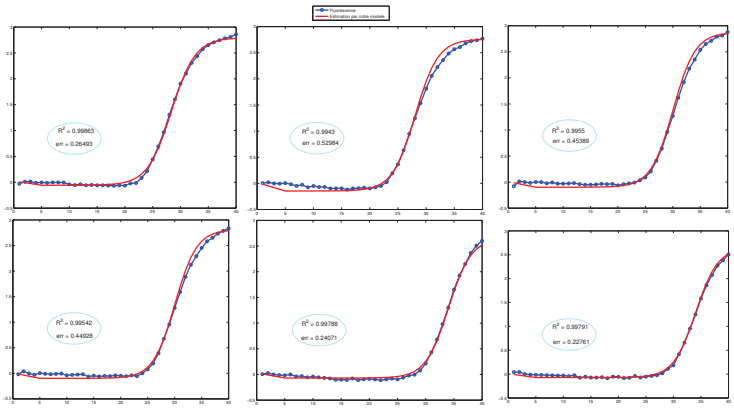
Les figures E.2, E.3 et E.4 montrent des résultats de l'estimation de la fluorescence par la méthode SCM, de Richards et par notre modèle.



(a) Résultats par la méthode SCM

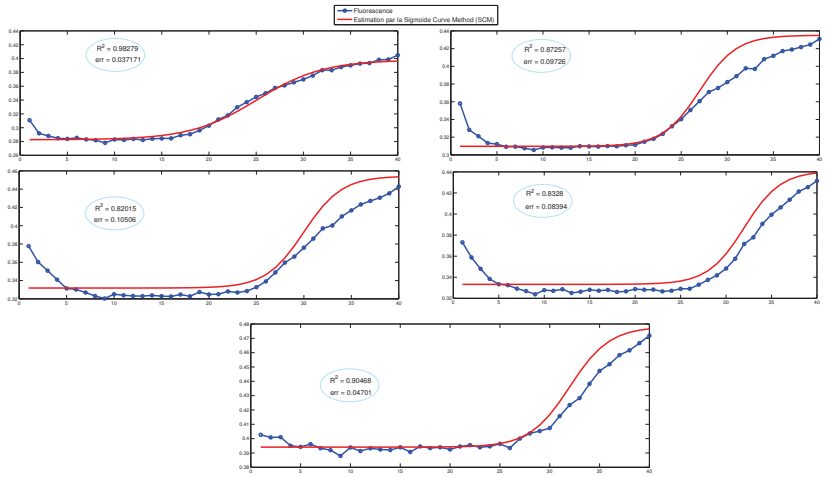


(b) Résultats par la méthode C_{y_0}

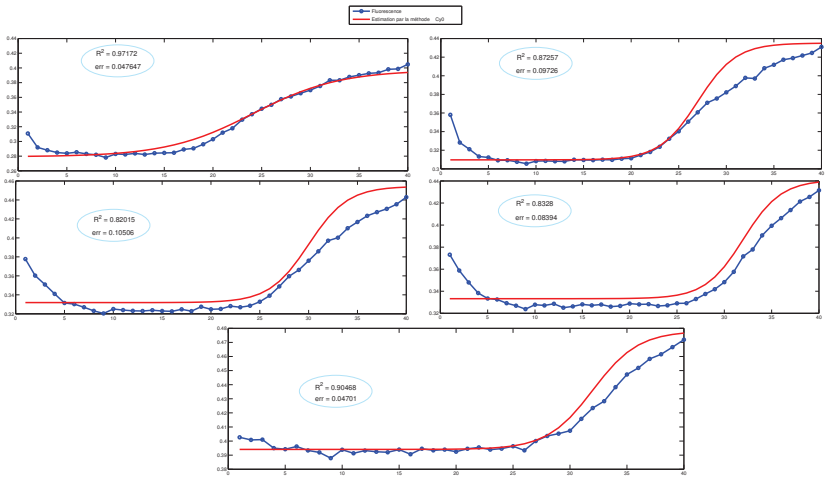


(c) Résultats par notre modèle

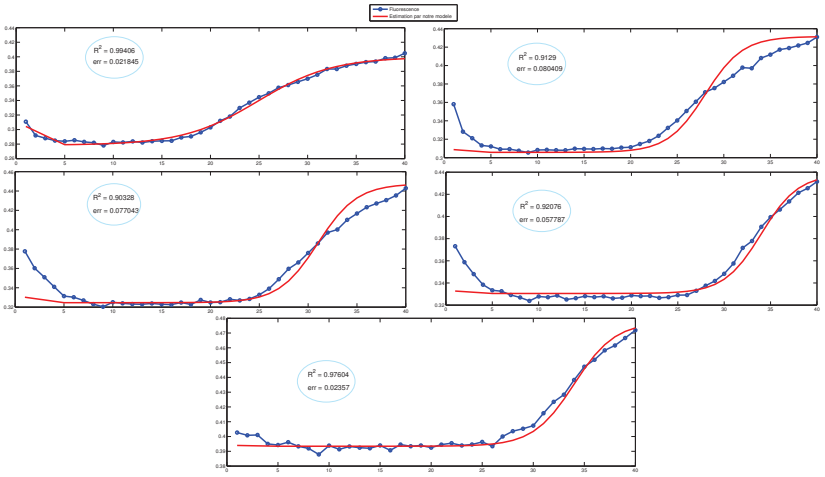
FIGURE E.2 – Illustration sur 6 fluorescences du LHP



(a) Résultats par la méthode SCM

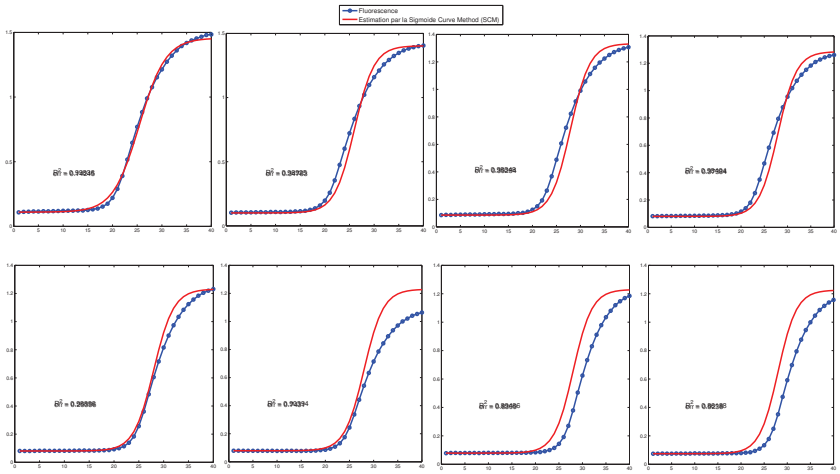


(b) Résultats par la méthode C_{y_0}

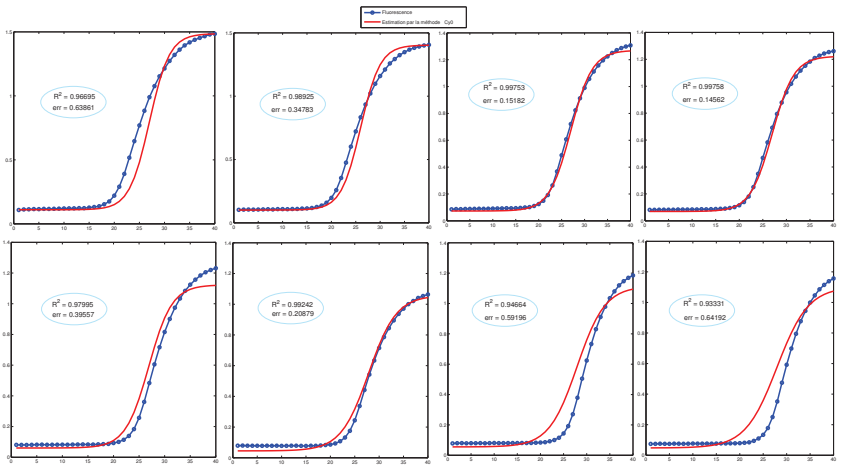


(c) Résultats par notre modèle

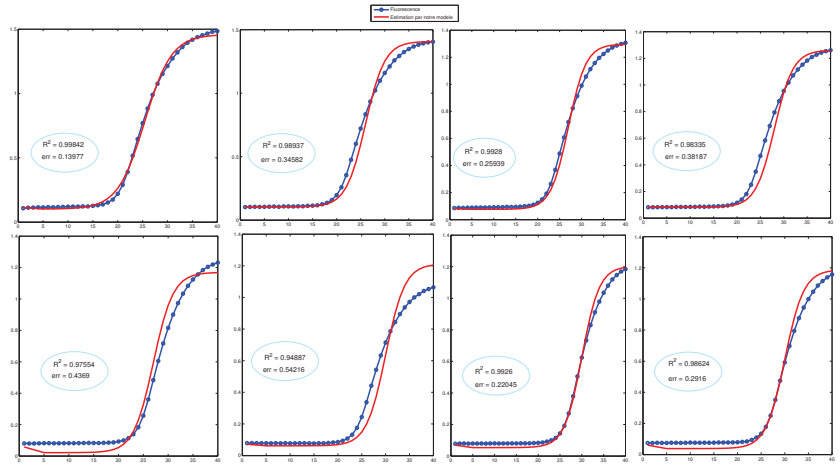
FIGURE E.3 – Illustration sur 5 fluorescences du CEA



(a) Résultats par la méthode SCM



(b) Résultats par la méthode C_{y_0}



(c) Résultats par notre modèle

FIGURE E.4 – Illustration sur 8 fluorescences du Primacen

ANNEXE F

Publications

Voici ci-dessous deux articles, en rapport avec la thèse, qui ont été publiés à la conférence de l'*European Symposium on Artificial Neural Networks, Computational Intelligence and Machine Learning* (ESAAN) 2013 à Bruges-Belgique et de Reconnaissance des Formes et Intelligence Artificielle (RFIA) 2012 de Lyon-France.

Detection and quantification in real-time polymerase chain reaction

Abou Keita[1,2], Romain Hérault[1,2], Colas Calbrix[1,3] and Stéphane Canu[1,2] *

1 - Normandie Univ, France

2 - INSARouen, LITIS, F-76801 Saint Etienne du Rouvray, France
{abou.keita,romain.herault,stephane.canu}@insa-rouen.fr

3 - Université de Rouen, PRIMACEN, F-76821 Mont Saint-Aignan, France
colas.calbrix@univ-rouen.fr

Abstract. The estimation of the concentration of an infectious agent in the environment is a key step to trigger an alert when there is a biological threat. This concentration can be obtained trough a quantitative polymerase chain reaction (qPCR). Nevertheless, standard real-time procedure do not address detection delay which is a main concern in alert triggering. Therefore, we propose a method based on Lasso regression and CUSUM change detection to accurately estimate the concentration while minimizing the detection delay. The trade-off between accuracy and delay can be managed through a parameter. We compare our results with those found by a standard method (threshold method) and promising results are obtained.

Keywords: Real-time PCR, quantitative PCR (qPCR), Change detection, Detection delay, CUSUM, Lasso, Biological threat.

1 Context

A sample of interest is taken from environment in order to be tested by *Polymerase Chain Reaction* (PCR). At each reaction cycle, genetic information supports are doubled [1, 2]. When a fluorophore matching specific genetic information from an infectious agent reaches its target, it emits light. Thus at each cycle of PCR, fluorescent light signal increases as genetic information increases (fig. 1(a)). This signal has three steps: baseline, exponential and plateau. The break time between baseline and exponential steps is log-linear to the initial concentration of targeted agent [3, 1]. Thus if we know accurately this break time (or specific cycle), we can compute the initial concentration of the agent in the sample. Moreover, in the context of alert triggering, the time between this break and the actual detection, that is detection delay, should be short.

Among methods described for specific cycle (change) detection in Biology works [2, 3], we can cite: a) The **threshold method** in which the specific cycle corresponds to the last time the fluorescence curve intersects a threshold.

*Authors would like to thanks the Agence Nationale de la Rercherche (ANR) for its support through the Genetic EquipemeNt for biothrEat environmental Analysis and SurveillancE (GENEASE) project and our partners Bertin technologies, CEA LETI and CEA SBTN.

b) The **second derivative method** in which the specific cycle is defined by the maximum of the fluorescence curve second derivative. c) The **sigmoid curve fitting method** in which the specific cycle is linked to a parameter of a curve model. None of them study the alert delay. Methods described for change (specific cycle) detection in Signal Processing [4, 5] can be split into two families: a) **Off-line methods**. For example, derivation method studied in [4] or the sigmoid curve fitting method (see above). Here, the full knowledge of the signal is assumed before the decision is taken. b) **On-line methods**. For example, the Shewhart rule, the moving average rule, the CUSUM method [4]. Here, observations arrive continuously, alert can be triggered as soon as possible. Under this latter setting, a trade off between precision on change detection and alert delay can be made accordingly to the application context.

Our proposal: To address detection delay meanwhile taking into account accuracy in quantitative real-time PCR, we use an on-line CUSUM method for fluorescence change detection. A kernelized-Lasso regression is done as a preprocessing step to overcome signal drift and to get rid of outlier samples.

Results will be compared with the threshold method included in the apparatus from which our data are recorded.

2 CUSUM

2.1 Introduction

To detect a potential break point t_0 from a time series in a sequential manner Page [6] introduced the CUSUM method that optimizes the detection delay. It is a statistical test between two hypothesis: no break point ($\mathcal{H}_0$) or break point ($\mathcal{H}_1$) occurs. Under ($\mathcal{H}_0$) observed data is assumed to be i.i.d. from some distribution P_{θ_0} where θ_0 is a parameter while under ($\mathcal{H}_1$) observed data underlying distribution parameter change from θ_0 to θ_1 at break time r that is

$$(\mathcal{H}_0):\ X = \{X_t\}_{t=1,\dots,s} \rightsquigarrow P_{\theta_0} \quad \text{vs} \quad (\mathcal{H}_1):\ \begin{array}{l} \exists\ r \text{ such that} \\ \{X_t\}_{t=1,\dots,r-1} \rightsquigarrow P_{\theta_0} \\ \{X_t\}_{t=r,\dots,s} \rightsquigarrow P_{\theta_1} \end{array}$$

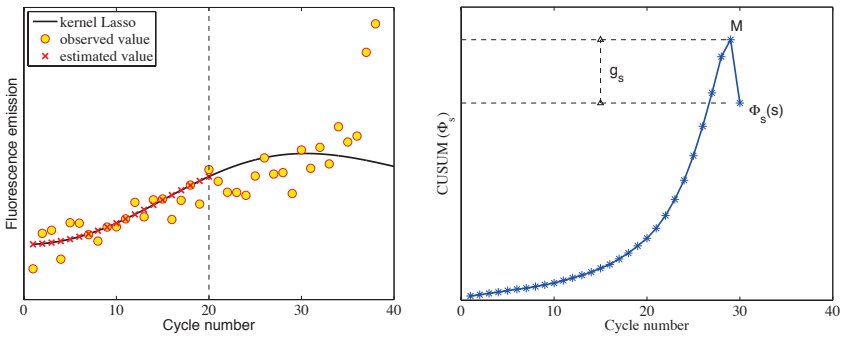
where s is the length of the signal. Thus, under null and alternative hypothesis, data distributions are

$$L(X|\mathcal{H}_0) = \prod_{t=1}^{s} P_{\theta_0}(X_t),\ \ L(X|\mathcal{H}_1) = \prod_{t=1}^{r-1} P_{\theta_0}(X_t) \prod_{t=r}^{s} P_{\theta_1}(X_t)\ , \tag{1}$$

leading to the following log-ratio [4, 5]:

$$S(X) = \log \frac{L(X|\mathcal{H}_1)}{L(X|\mathcal{H}_0)} = \log \frac{\prod_{t=1}^{r-1} P_{\theta_0}(X_t) \prod_{t=r}^{s} P_{\theta_1}(X_t)}{\prod_{t=1}^{s} P_{\theta_0}(X_t)} = \sum_{t=r}^{s} \log \frac{P_{\theta_1}(X_t)}{P_{\theta_0}(X_t)}\ . \tag{2}$$

$S(x)$ as a function of unknown break time r is denoted by $\Phi_s(r)$, the CUSUM indicator.



(a) Kernelized-lasso regression (b) CUSUM indicator Φ_s

Fig. 1: Preprocessing and CUSUM steps

P_{θ_0} and P_{θ_1} are assumed to follow Gaussian distributions respectively parametrized by (μ_0, σ_0) and (μ_1, σ_1). In our application context, providing by experts that breaks do not occurs until $s_{min} = 20$, $(\hat{\mu}_0, \hat{\sigma}_0)$ are estimated on $\{x_t\}_{t=1,\ldots,s_{min}}$, $\hat{\sigma}_1$ is set to $\hat{\sigma}_0$, $\hat{\mu}_1$ is estimated on $\{x_t\}_{t=r,\ldots,s}$.

As r is unknown, detection is based on M, the *log* maximum of the generalized likelihood ratio,

$$M = \max_{1 \leq r \leq s} \Phi_s(r) \ . \tag{3}$$

As we see from the figure 1(b), the indicator increases until it reachs M and then decreases. We perform the following statistic test $g_s = M - \Phi_s(s)$. If g_s is below a chosen threshold h, the signal is in $(\mathcal{H}_0)$ state; otherwise, we have switched to $(\mathcal{H}_1)$ state. Here, an alert is triggered (i.e. the alarm time t_a equals s) and the break point, t_0, is given by the position of M, that is, $t_0 = \arg \max_{1 \leq r \leq s} \Phi_s(r)$.

2.2 Application to fluorescence signals obtained by PCR

CUSUM can not be applied directly to a fluorescence signal. In our application context, the fluorescence is not i.i.d. and outliers are encountered at the beginning of the signal, disturbing the estimation of (μ_0,σ_0). That is why a kernelized-lasso regression [7] is estimated on the twenty first observations by solving the following linear program

$$\underset{\boldsymbol{\alpha}}{argmin} \ ||\mathbf{y} - \boldsymbol{\alpha} K||_1 + \lambda ||\boldsymbol{\alpha}||_1 \ , \tag{4}$$

where $\boldsymbol{\alpha}$ is coefficient vector of size s_{min}, λ the regularization parameter, K the Gaussian kernel matrix and $\mathbf{y}$ the s_{min} first samples where $s_{min} = 20$. Figure 1(a) illustrates on an example the way kernel lasso regression is used.

Moreover, the DNA duplication reaction is continuous but fluorescence is not: fluorophore need light excitation in order to, at there turn, emit light. This

excitation process takes one minute (a cycle) which is far from the precision needed to quantify the concentration. We need a better time precision than the sampling rate! This is overcome by using a spline approximation of the CUSUM indicator, then, the break time (or specific cycle) t_0 is redefined to the spline maximum.

Algorithm 1 CUSUM algorithm for real-time qPCR

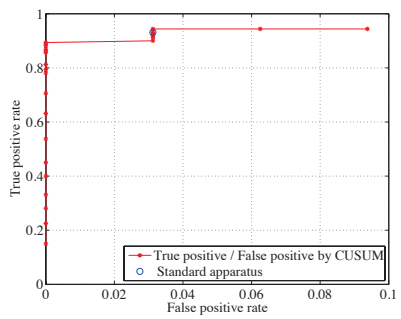
Pre-processing :
Wait for s_{min} samples
Compute kernel regression on s_{min} samples then (μ_0,σ_0) on residuals
Main algorithm:
$s \leftarrow s_{min} + 1$, $decision \leftarrow 0$
while $decision == 0$ **do**
  Compute Φ_s
  $[M, position] \leftarrow \max(\Phi_s)$
  $g_s \leftarrow M - \Phi_s(s)$
  if $g_s >= h$ **then**
    $decision \leftarrow 1$, $t_a \leftarrow s$, $t_0 \leftarrow position$
  else
    $s \leftarrow s + 1$ (Wait for a new sample)
  end if
end while
Post-processing:
$S\Phi_s \leftarrow spline(\Phi_s(t_0 - 2 : t_0 + 1))$
$t_0 \leftarrow arg\max S\Phi_s$

3 Experimentation

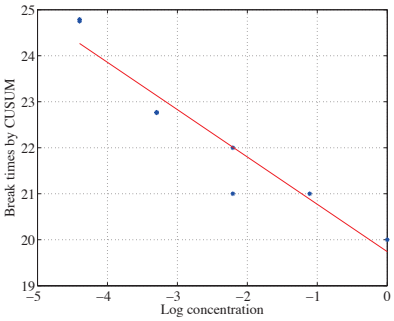
3.1 Set-up

Fluorescence signals have been recorded on a standard apparatus, 7500 Fast Real-Time PCR System of Applied Biosystems, which itself gives the specific cycle by a simple threshold method at the end of the record. It can process multiple samples in plate consisting in 8 lines by 12 columns of wells. Materials of different kinds and at different concentrations are analysed in the wells: Each line containing only one combination of tested material and fluorophore; Columns (2 by 2) containing a specific concentration. The two last columns are negative. The record of two plates have been processed, that is 192 signals (160 positive and 32 negative).

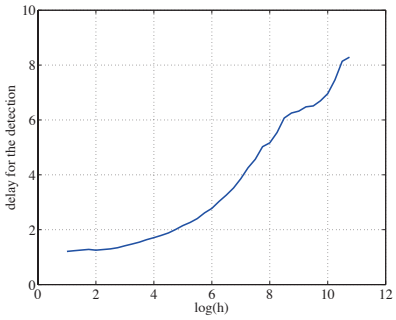
As the threshold method and the CUSUM method do not look for the same event in the signal, results can not be compared directly. Nevertheless, line by line (material wisely), they should be linear together. Eventually, line by line, CUSUM results and threshold results should both be log-linear to the concentration. We will first consider results in term of detection (true positive versus false positive) and, in a second step, look for the accuracy of the detection (log-linearity to concentration) and detection delay only on true positive examples. Accuracy is evaluated on the mean of residuals of log-regression performed line



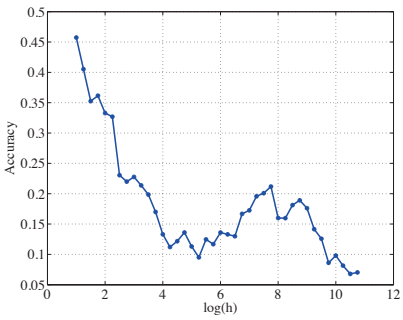
(a) Curve operational



(b) Break times by the CUSUM with respect to the log concentration on online for $h = exp(1.5)$.



(c) Delay detection in cycles with respect to h with CUSUM method (compared to 13.88 cycles for apparatus)



(d) Accuracy with respect to h with CUSUM method (compared to 0.77 for apparatus)

Fig. 2: False alarm rate compared to the rate of correct detection, delay, accuracy

by line. A trade-off between these two indicators can be made trough the h parameter in the CUSUM method.

3.2 Results

Standard apparatus has a fixed point operation that gives 93.13% of true positive rate with 3.12% of false alarm. On the true positive examples, the mean residual of the log-regression between break time and log-concentration is 0.77 and the detection delay 13.88 cycles.

The best operation point is obtained by CUSUM method at $h = 1.5$. It gives 94.37% of true positive rate with 3.12% of false alarm. At this operation point, on the true positive examples, the mean residual of the log-regression between break time and log-concentration is 0.35 and the detection delay 1.26 cycles. The CUSUM method is on all indicators better than the standard apparatus.

Especially, it reduces by 10 the detection delay. An example of log-linearity on a specific plate line is shown on figure 2(b).

By changing h, we can draw an operational curve (fig. 2(a)) in term of true positive rate and false alarm rate. Moreover h influences the detection delay (fig. 2(c)) and accuracy (fig. 2(d)). From those two last curves we can induced two possible strategies: A) Taking h between $exp(4)$ and $exp(6)$ that leads to good accuracy with a delay inferior to 2 cycles; B) Or taking h superior to $exp(10)$ that leads to an even better accuracy at the price of a higher delay.

4 Conclusion

The determination of a characteristic cycle of the fluorescence from qPCR can be done in continuous time and without prior knowledge of the total signal through a CUSUM method. Care must be taken to estimate statistical models at the beginning of the signal: a kernelized lasso regression is performed to avoid outliers. We obtain similar results in term of true positive and false alarm rates as standard apparatus but with better performance in term of concentration estimation accuracy and detection delay. Moreover, the method is easily tunable to application context (delay vs accuracy) trough a trade-off parameter.

References

[1] J.D. Durtschi, J. Stevenson, W. Hymas, and K.V. Voelkerding. Evaluation of quantification methods for real-time PCR minor groove binding hybridization probe assays. *Analytical biochemistry*, 361(1):55–64, 2007.

[2] M. Guescini, D. Sisti, M.B.L. Rocchi, L. Stocchi, and V. Stocchi. A new real-time PCR method to overcome significant quantitative inaccuracy due to slight amplification inhibition. *BMC bioinformatics*, 9(1):326, 2008.

[3] R.G Rutledge. Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. *Nucleic acids research*, 32(22):e178, 2004.

[4] M. Basseville and I.V. Nikiforov. *Detection of abrupt changes: theory and application*, volume 10. Prentice-Hall, 1993.

[5] G. Verdier, N. Hilgert, and J.P. Vila. Optimality of cusum rule approximations in change-point detection problems: application to nonlinear state–space systems. *Information Theory, IEEE Transactions on*, 54(11):5102–5112, 2008.

[6] E.S Page. Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954.

[7] E.J. Candes, M.B. Wakin, and S.P. Boyd. Enhancing sparsity by reweighted ℓ_1 minimization. *Journal of Fourier Analysis and Applications*, 14(5):877–905, 2008.

Estimation de la concentration d'un agent biologique par détection de rupture sur vidéos de fluorescences issues de PCR.

M. KEITA Abou[1] M. HERAULT Romain[1] M. CANU Stephane[1]

[1]LITIS - INSA ROUEN

LITIS EA 4108, INSA de Rouen, Avenue de l'université BP 8
76801 Saint Etienne du Rouvray
abou.keita@insa-rouen.fr / romain.herault@insa-rouen.fr / stephane.canu@insa-rouen.fr

Résumé

L'estimation de la concentration d'un agent pathologique est une étape primordiale pour déclencher une alerte en cas de menace biologique. En appliquant une PCR associée à un marqueur fluorescent sur un échantillon à analyser, nous observons une émission lumineuse qui est fonction de la concentration de l'agent. Cette concentration est alors déduite du nombre de cycles de réaction à partir duquel la fluorescence change de comportement. Dans le cadre du projet GENEASE, nous proposons d'utiliser une méthode de détection de rupture, le CUSUM, qui estime ce nombre de cycles. Nous comparons ces résultats à ceux trouvés par la méthode du maximum de la dérivée seconde, méthode qui est utilisée dans les appareils industriels et des laboratoires. Nous obtenons des résultats comparables, mais contrairement à cette dernière, sans disposer du signal complet et en minimisant le retard à la détection.

Mots Clef

Détection de rupture, Retard à la détection, Signal de fluorescence, PCR, Reconnaissance d'agents biologiques.

Abstract

Estimating concentration of a pathological agent is an essential step for triggering an alert in case of biological threat. By applying a PCR associated with a fluorescent marker, we can observe that emitted light depends on the concentration of the agent. This quantity can be deduced from the number of reaction cycles at which the fluorescence changes in behavior. As part of the GENEASE project, we propose to use the CUSUM method in order to detect the rupture. This method allows to estimate when the rupture occurs in term of cycle number. We compare these results with those found by the method of the second derivative maximum, method that is used in industrial appliances and laboratories. We obtain similar results ; however, compare to the latter, we do not need a complete signal which brings lower detection delay.

Keywords

Rupture detection, Detection delay, Fluorescence signal, PCR, Recognition of biological agents.

1 Introduction

L'émergence de le menace biologique est devenue une réalité, avec notamment les attaques à l'Anthrax en 2001, qui a conduit à la mise en place du plan Biotox complétant le dispositif de l'Organisation de la Réponse de SEcurité Civile (ORSEC) de gestion des risques et de la Crise face à une menace biologique. Dans ce cadre, la détection au plus tôt d'une agression et sa caractérisation de manière fiable permet une réponse optimale en terme d'organisation de secours et de dispositions à adopter afin d'en limiter les conséquences. A ce jour, il n'existe pas de système de surveillance et d'analyse biologique en continu de l'environnement.

Le projet GENEASE [1], auquel nous participons, porte sur l'étude d'un système compact et sensible de surveillance et d'analyse biologique en semi-continu de l'environnement permettant de détecter et d'identifier plusieurs agents biologiques simultanément par biologie moléculaire.

Une *Polymerase Chain Reaction* (PCR) va être utilisée à cette fin. La quantité de matériel génétique est doublée à chaque cycle de la réaction. Si un marqueur fluorescent est lié à ce matériel, nous observons une émission lumineuse croissant exponentiellement. Nous pouvons alors déduire du point de montée de fluorescence la concentration initiale du matériel génétique.

Diverses méthodes de calcul de la concentration sont utilisées en laboratoires et dans l'industrie. Elles sont principalement basées sur le calcul d'un point caractéristique du signal de fluorescence, appelé C_t, point exprimé en nombre de cycles de réaction. Ce C_t est une fonction $\log$ linéaire de la concentration notée C_0. Après étalonnage, nous pouvons donc déduire de C_t la concentration C_0. Cependant

1. Projet financé par l'ANR et composé de Bertin Technologies, du CEA LETI, du CEA SBTN et du LITIS, cf. Remerciements

ces méthodes nécessitent la connaissance complète du signal de fluorescence.
Pour notre application, la détection doit s'effectuer le plus rapidement possible après que le signal de fluorescence ait atteint le point caractéristique, C_t, pour éventuellement déclencher une alerte. Ainsi, on doit à la fois être confiant dans le nombre de cycles retenus mais aussi minimiser le retard à la détection. Deux objectifs *a priori* contradictoires. De plus, nous souhaitons mettre en place une méthode robuste sur un équipement de terrain car les signaux de fluorescence peuvent être très bruités.
De ce fait, nous proposons une technique statistique robuste de détection de rupture, basée sur le CUSUM [1]. Déduit d'un rapport de vraisemblance dynamique entre deux populations, le nombre de cycles C_t est obtenu avec une confiance satisfaisante.
Nous allons d'abord faire un état de l'art en présentant le signal de fluorescence avec deux techniques de détermination du C_t, puis l'approche CUSUM et enfin son application sur les signaux de fluorescences.

2 Présentation d'un signal de fluorescence

Les données dont nous disposons sont obtenues par le suivi en temps réel d'une *Polymerase Chain Reaction* (PCR) [5, 6, 3, 10]. Il s'agit d'une technique découverte par Kary Mullis en 1983 [7]. C'est un outil universel dans le domaine de la biologie pour la détection et les quantifications des acides nucléiques.
Le suivi en temps-réel d'une réaction de PCR peut être représenté sous la forme d'une courbe de fluorescence (fig. 2). Ce signal correspond à la moyenne sur une zone d'intérêt de la lumière fluorescente captée par caméra à la fin de chaque cycle de réaction, chaque cycle durant approximativement 1 minute.
Les signaux de fluorescences sur lesquels nous travaillons sont issus de l'immuno-analyse d'agents biologiques, à partir d'appareils de laboratoire (CEA-LETI et LHVP [2]). Le CEA-LETI fabrique une puce « Smart Drop » sur laquelle se fait cette analyse biologique [2].
La cinétique de cette réaction met en jeu trois phases : une phase d'initiation, une phase exponentielle et une phase plateau. La position de la montée du signal, lorsque l'on passe de la phase d'initiation à la phase exponentielle est fonction de la concentration de l'espèce à qualifier dans la goutte analysée. Cette position est repérée par un nombre de cycles appelé C_t.
Il existe différentes méthodes de détermination du C_t, par exemple la méthode du maximum de la dérivée seconde (fig. 3(b)) et la méthode dite du seuil direct (fig. 3(a)) [6, 12]. La concentration C_0 est log linéaire par rapport au cycle C_t. Un étalonnage préalable pour calculer les paramètres de la fonction linéaire, paramètres différents pour chaque méthode et condition d'expérimentation, permet de déduire C_0 de C_t [9]. Il est ici à noter que ces deux méthodes sont consistantes si les deux séries de C_t trouvées montrent uniquement un biais linéaire.
Cependant ces méthodes requièrent la connaissance de l'intégralité du signal de fluorescence. L'un des buts de notre étude est de déterminer le point de montée avec un minimum de retard et sans la connaissance complète du signal. De ce fait nous allons utiliser la technique du CUSUM qui est une des méthodes de détection de changement sur un ensemble de données.
Sur l'ensemble des signaux de fluorescences (fig. 1(a) par exemple), nous observons deux types de comportements, des signaux avec un effet de saut, signifiant la présence de l'agent, et d'autres sans saut, signifiant son absence (fig. 1(b)). L'utilisation de cette technique nous permettra de distinguer ces deux cas et d'estimer l'instant de rupture.

Remarque : Nous comparerons nos résultas à ceux obtenus par la méthode de la dérivée seconde qui est actuellement la méthode utilisée par le CEA-LETI et le LHVP.

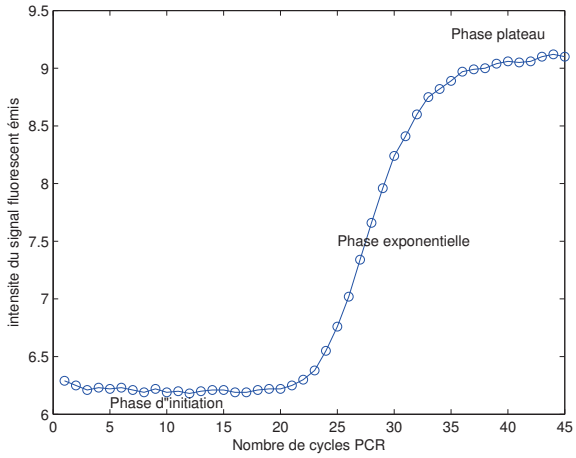


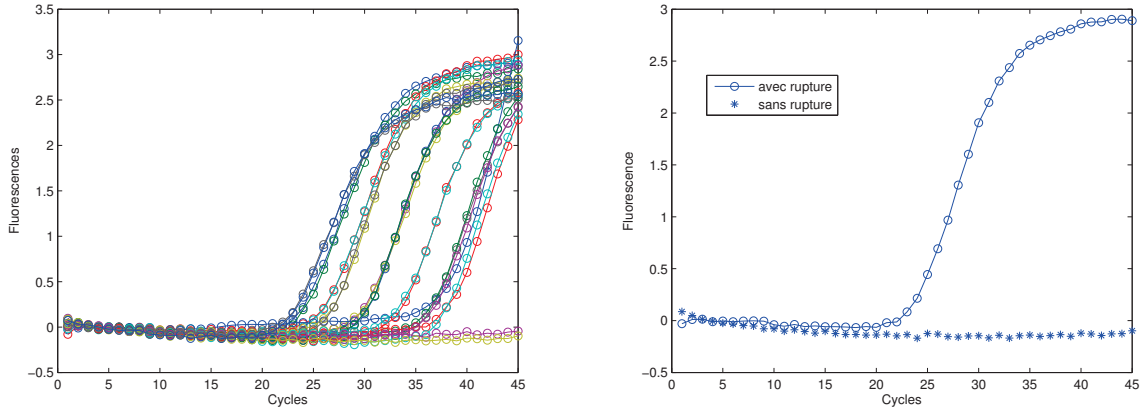
FIGURE 2 – Suivi en temps réel d'une réaction de fluorescence qui correspond à la moyenne de la fluorescence sur une région d'intérêt.

3 Présentation du CUSUM

3.1 Approche

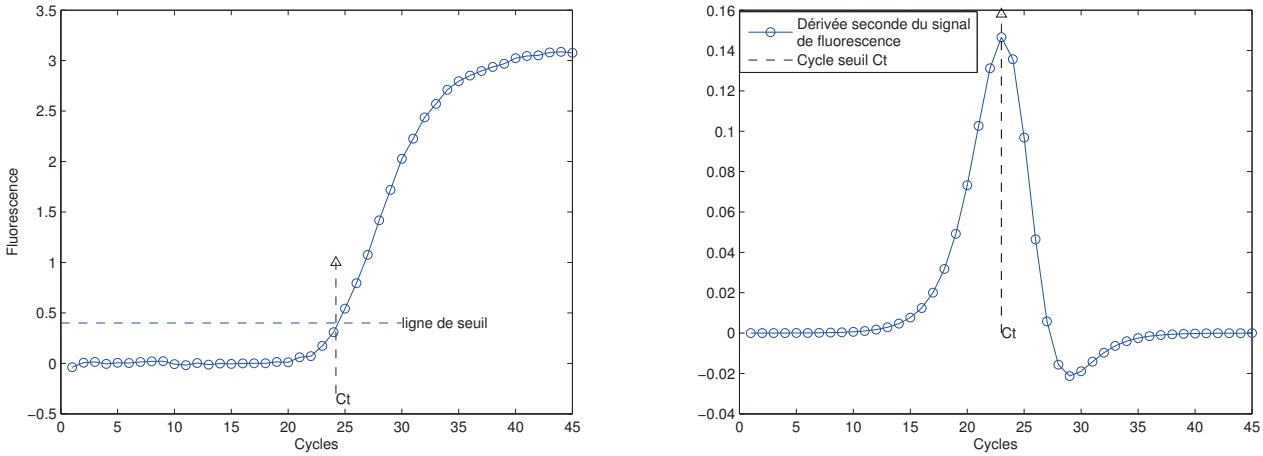
Le *CUmulative SUM* (CUSUM) [1, 8] est l'une des techniques statistiques utilisées pour la détection de rupture sur des données. Cette méthode est déduite d'après le rapport de vraisemblance dynamique entre deux populations. Il s'agit d'une technique d'analyse séquentielle. En considérant une distribution de probabilité P_θ, paramétrée par θ, (l'espérance par exemple), cette méthode consiste à détecter un changement sur ce paramètre.
La figure 4(a) présente un échantillon gaussien de variance constante et deux valeurs distinctes de la moyenne. Il existe un instant de rupture r tel que l'espérance de cet échan-

2. Cf. Remerciements



(a) Données. (b) Deux types de comportements.

FIGURE 1 – Représentation graphique des données de fluorescence (a) et ses deux types de comportements (b).



(a) Méthode du seuil. Le C_t est défini comme étant égal à l'intersection d'une ligne de seuil fixée avec la courbe de fluorescence.

(b) Maximum de la dérivée seconde. Le C_t est défini comme étant égal au maximum de la dérivée seconde de la courbe de fluorescence.

FIGURE 3 – Deux différentes méthodes de détermination d'un cycle seuil.

tillon est égale à un réel θ_0 avant r et un réel θ_1 après r. Le CUSUM permet de détecter ce changement (fig. 4(b)) [1].

3.2 Cadre de travail

Le cœur de ce travail est basé sur un modèle statistique de test paramétrique. Il revient alors à bien poser le test d'hypothèse (éq. 1) et définir sa région critique (éq. 2). Dans la suite nous utilisons les notations suivantes :

Notations

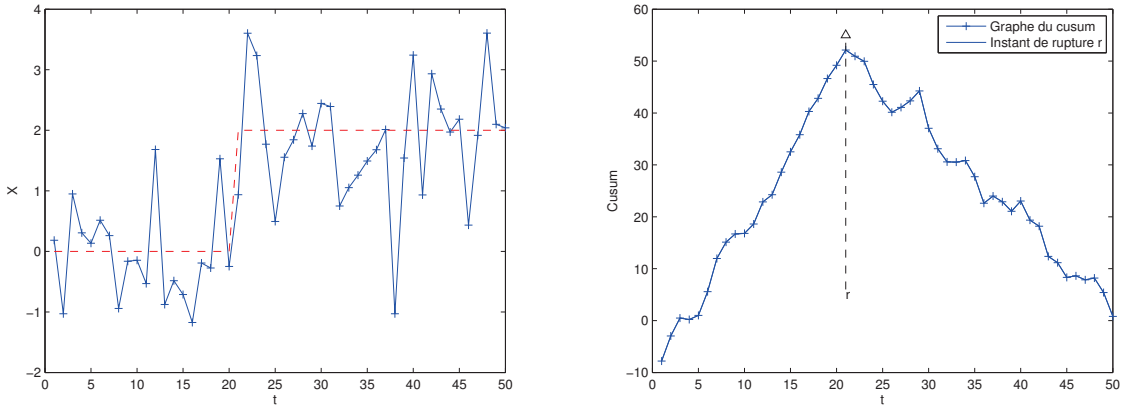
– n : nombre de fluorescences,
– C_{max} : nombre de cycle maximal,
– i, j et k indices associés aux signaux, $1 \leq i, j, k \leq n$,
– r, s et t indices associés aux cycles, $1 \leq r, s, t \leq C_{max}$.

Notons $\mathcal{X} = \{\mathbf{x}^i\}_{i=1,\ldots,n}$ comme étant l'ensemble des fluorescences. Supposons que la fluorescence $\mathbf{x}^i = \{x_t^i\}_{t=1,\ldots,C_{max}}$, $\forall\ i \in [1, n]$, admet une densité de probabilité $\mathcal{P}_\theta$ paramétrée par $\theta \in \Omega$ où Ω est appelé espace des paramètres réels.

Deux situations se présentent :

– soit toutes les observations de l'échantillon ont la même densité de probabilité, caractérisée par θ_0,
– soit il existe r inconnu, $1 \leq r \leq C_{max}$, tels que $\theta = \theta_0$ avant r et $\theta = \theta_1 (\neq \theta_0)$ après r.

Nous allons représenter les deux situations (saut ou sans saut) sous la forme d'un test d'hypothèse.



(a) Échantillon gaussien

(b) Graphe du CUSUM associé l'échantillon gaussien (a)

FIGURE 4 – Un échantillon gaussien avec son graphe de CUSUM correspondant [1]. L'algorithme du CUSUM estime l'instant de rupture à $t = 21$.

Test d'hypothèse de niveau α (voir ci après pour la définition de α) : $\forall\ s$ tel que $1 \leq s \leq C_{max}$

$$\begin{cases} \mathcal{H}_0 : & \mathbf{x}^i = \{x_t^i\}_{t=1,\dots,s} \rightsquigarrow \mathcal{P}_{\theta_0} \\ \mathcal{H}_1 : & \exists\ r \text{ tel que} \\ & \mathbf{x}^i = \{x_t^i\}_{t=1,\dots,r-1} \rightsquigarrow \mathcal{P}_{\theta_0} \\ & \mathbf{x}^i = \{x_t^i\}_{t=r,\dots,s} \rightsquigarrow \mathcal{P}_{\theta_1} \end{cases} \quad (1)$$

Le principal problème est de comparer l'hypothèse nulle $(\mathcal{H}_0)$ à l'hypothèse alternative $(\mathcal{H}_1)$.

Remarque : Nous désirons détecter la rupture le plus tôt possible, le test se fait alors de manière récursive suivant s.

Définition : α est appelé erreur de premier espèce, c'est à dire la probabilité de refuser l'hypothèse nulle $\mathcal{H}_0$ à tort, c'est à dire :

$$\alpha = P\left(\mathcal{H}_1 | \mathcal{H}_0\right).$$

Nous allons maintenant définir la région critique ou région d'acceptation de l'hypothèse alternative [1, 8].

Règle de décision ou région critique [11] : Pour un signal i, $(1 \leq i \leq n)$,

$$W^i = \left\{ (x_1^i, ..., x_s^i) / \log \frac{L(x_1^i, ..., x_s^i | \mathcal{H}_1)}{L(x_1^i, ..., x_s^i | \mathcal{H}_0)} \geq h_\alpha \right\}, \quad (2)$$

avec h_α un seuil décision convenablement choisi [1], et L représente la vraisemblance [1, 4].
W^i est une région d'acceptation de l'hypothèse $\mathcal{H}_1$ pour le signal i $(1 \leq i \leq n)$.
La vraisemblance est mise en œuvre dans le but de quantifier l'adéquation entre une distribution de probabilité et un échantillon : plus grande est la vraisemblance de l'échantillon, meilleure est l'adéquation. Elle est utilisée par les tests d'hypothèses dont le but est de discriminer entre deux groupes de distributions qui affirment chacun contenir la distribution qui a engendré l'échantillon.

Le log du rapport de vraisemblance est égal à,

$$\log \frac{L(x_1^i, ..., x_s^i | \mathcal{H}_1)}{L(x_1^i, ..., x_s^i | \mathcal{H}_0)} = \log \frac{\prod_{t=1}^{r-1} P_{\theta_0}(x_t^i) \prod_{t=r}^{s} P_{\theta_1}(x_t^i)}{\prod_{t=1}^{s} P_{\theta_0}(x_t^i)}.$$

Après simplification, nous obtenons

$$\log \frac{L(x_1^i, ..., x_s^i | \mathcal{H}_1)}{L(x_1^i, ..., x_s^i | \mathcal{H}_0)} = \sum_{t=r}^{s} \log \frac{P_{\theta_1}(x_t^i)}{P_{\theta_0}(x_t^i)}. \quad (3)$$

Notons

$$S_s^i(r) = \sum_{t=r}^{s} \log \frac{P_{\theta_1}(x_t^i)}{P_{\theta_0}(x_t^i)}. \quad (4)$$

Pour résoudre un tel problème, nous utilisons le principe du maximum du log du rapport de vraisemblance généralisé car l'instant de rupture r est inconnu [1] .
Posons alors,

$$S_s^i := \max_{1 \leq r \leq s} S_s^i(r), \quad (5)$$

et la règle de décision d^i, telle que $d^i = 0$ ou 1 est accordée respectivement à $\mathcal{H}_0$ ou $\mathcal{H}_1$, est donnée par,

$$d^i = \begin{cases} 0 & \text{si } S_s^i < h_\alpha \\ 1 & \text{si } S_s^i \geq h_\alpha. \end{cases}$$

Nous désirons détecter l'instant de rupture le plus tôt possible. Ainsi avant d'estimer cet instant r, nous allons d'abord déterminer le temps d'arrêt t_a, qui est l'instant minimum où le choix $d^i = 1$ est pris. Cet instant t_a nous permet de minimiser le retard à la détection. Il vient que,

$$t_a = \min\{s : S_s^i \geq h_\alpha\}. \quad (6)$$

Une fois que l'instant d'arrêt t_a est déterminé, nous sommes alors sous l'hypothèse $(\mathcal{H}_1)$, il reste à déterminer l'instant de rupture.

Estimation de r : Il existe un instant r où il y a un saut, le problème revient alors à estimer cet instant. En utilisant le principe du maximum de vraisemblance, nous obtenons [1],

$$r^* = arg \max_{1\leq r\leq t_a} \log \left[\prod_{t=1}^{r-1} P_{\theta_0}(x_t^i) \prod_{t=r}^{t_a} P_{\theta_1}(x_t^i) \right]. \quad (7)$$

où $\log \left[\prod_{t=1}^{r-1} P_{\theta_0}(x_t^i) \prod_{t=r}^{t_a} P_{\theta_1}(x_t^i) \right]$ est le log vraisemblance et r^* est l'estimateur de l'instant de rupture r.

En multipliant et en divisant par $\prod_{t=r}^{t_a} P_{\theta_0}(x_t^i)$, nous obtenons que,

$$r^* = arg \max_{1\leq r\leq t_a} \left[\log \frac{\prod_{t=r}^{t_a} P_{\theta_1}(x_t^i)}{\prod_{t=r}^{t_a} P_{\theta_0}(x_t^i)} + \log \prod_{t=1}^{t_a} P_{\theta_0}(x_t^i) \right]. \quad (8)$$

Le dernier terme étant constant par rapport à r, nous avons,

$$r^* = arg \max_{1\leq r\leq t_a} \left[\sum_{t=r}^{t_a} \log \frac{P_{\theta_1}(x_t^i)}{P_{\theta_0}(x_t^i)} \right]. \quad (9)$$

Finalement,

$$r^* = arg \max_{1\leq r\leq t_a} S_{t_a}^i(r), \quad (10)$$

$S_{t_a}^i(r)$ ainsi défini représente le CUSUM du signal i. Elle est la somme cumulée du log du rapport de vraisemblance pour les observations de r à t_a.

La figure 5 illustre en détails la procédure de l'estimation de r pour chaque signal i , $1 \leq i \leq n$ et $\forall\ 1 \leq c \leq C_{max}$.

Remarque : Les paramètres avant la rupture θ_0 et après la rupture θ_1 ne sont pas nécessairement connus. Nous utilisons alors le log du rapport de vraisemblance généralisé [1], qui est :

$$\log \frac{\hat{L}(x_1^i, ..., x_s^i|\mathcal{H}_1)}{\hat{L}(x_1^i, ..., x_s^i|\mathcal{H}_0)} = \sup_{\theta_1,\theta_0} \sum_{t=r}^{s} \log \frac{P_{\theta_1}(x_t^i)}{P_{\theta_0}(x_t^i)} =: \hat{S}_s^i(r). \quad (11)$$

Dans d'autres cas, il n'y a que le paramètre après la rupture qui est inconnu. De ce fait le log du rapport de vraisemblance généralisé [1] est :

$$\log \frac{\hat{L}(x_1^i, ..., x_s^i|\mathcal{H}_1)}{\hat{L}(x_1^i, ..., x_s^i|\mathcal{H}_0)} = \sup_{\theta_1} \sum_{t=r}^{s} \log \frac{P_{\theta_1}(x_t^i)}{P_{\theta_0}(x_t^i)} =: \hat{S}_s^i(r). \quad (12)$$

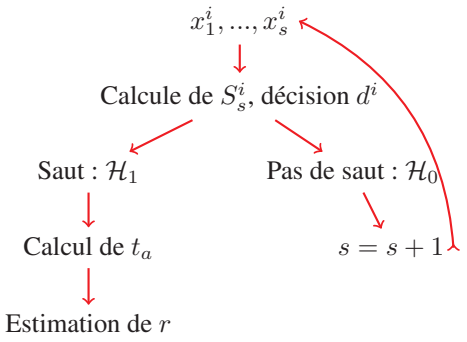


FIGURE 5 – Schéma illustrant l'algorithme de détection de rupture.

4 Résultats obtenus et étude comparative

Dans cette partie, nous appliquons l'algorithme du CUSUM sur les données fournies par le LHVP provenant de l'appareil d'enregistrement. Cet appareil mesure les points caractéristiques, C_t, par la méthode de la dérivée seconde. Ces derniers seront comparés à ceux que nous obtenons par la méthode du CUSUM.

Les signaux de fluorescence sont représentés sur la figure 1(a) et les points caractéristiques correspondants sur le tableau 1.

En appliquant la méthode du CUSUM, nous obtenons des instants de ruptures notés r. Les points caractéristiques réels sont déduis en faisant une interpolation (spline par exemple) ou une pondération sur les r entiers. Les résultats obtenus sont représentés sur le tableau 1 : la colonne $\hat{C}_{\partial^2}$ représente les cycles livrés par l'appareil standard ; la colonne t_a représente le temps d'arrêt ; la colonne $\hat{C}_{CUP}$ représente les points caractéristiques obtenus via une pondération autour des r et $\hat{C}_{CUS}$ ceux obtenus via une interpolation spline.

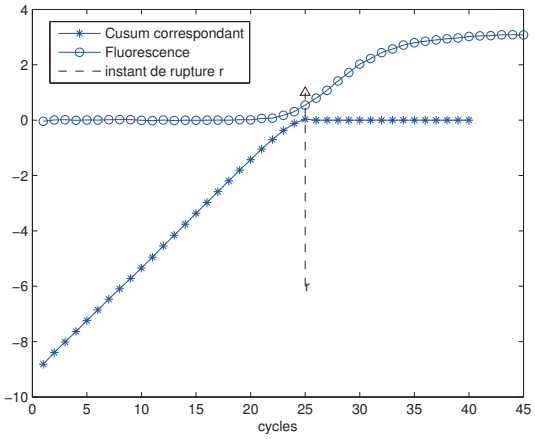
Nous constatons que les points caractéristiques obtenus via la méthode du CUSUM sont cohérents avec ceux livrés par l'appareil de mesure standard (tab. 1) car nous restons linéaire par rapport aux résultats fournis par ce dernier (fig. 9).

Deux exemples de comportement sur le graphe du CUSUM associé à sa fluorescence sont représentés sur les figures 6(a) et 6(b).

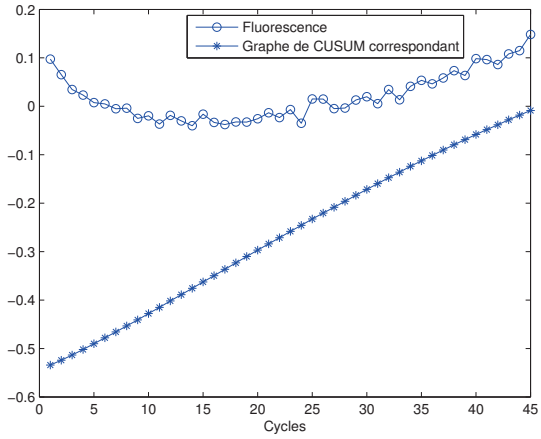
En représentant sur la même figure les points caractéristiques obtenus par la méthode de la dérivée seconde par rapport à ceux obtenus par la méthode du CUSUM, nous obtenons la figure 7.

Le graphe des cycles calculés via la pondération par rapport aux cycles obtenus par la méthode de la dérivée seconde, (fig. 9(b)), présente quelques fluctuations par rapport à ceux obtenus par l'interpolation spline (fig. 9(a)).

Nous pouvons alors déduire que l'interpolation est meilleure que la pondération. Dans les deux cas, nous res-



(a) Signal de fluorescence avec un effet de saut et son graphe de CUSUM correspondant



(b) Signal de fluorescence sans un effet de saut et son graphe de CUSUM correspondant

FIGURE 6 – Deux types de comportement des données et leur CUSUM correspondant. Un cycle seuil est déterminé à partir de l'instant de rupture pour des signaux avec saut et cet instant est l'instant auquel la courbe du CUSUM atteint son maximum.

tons linéaire par rapport aux points caractéristiques obtenus par l'appareil standard (fig. 9) donc le C_0 peut en être déduit après étalonnage. Nous observons un biais constant sur les valeurs obtenues par rapport à celles de l'appareil de mesure standard. Nous allons faire de l'apprentissage par validation croisée afin d'estimer ce biais et d'en appliquer une correction.

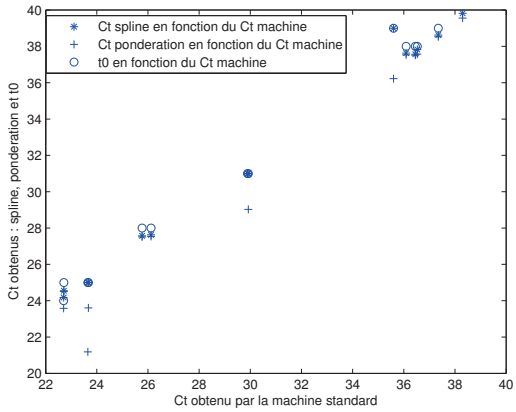


FIGURE 7 – Points caractéristiques obtenus par la méthode de la dérivée seconde par rapport aux notre et aux instants de rupture t_0.

Validation croisée :

Le biais rencontré est constant. Nous allons alors calculer cette correction par apprentissage en utilisant la validation croisée par *leave one out*. Elle consiste à faire la moyenne de la différence entre le point caractéristique obtenu par la méthode de la dérivée seconde et celui que nous obtenons

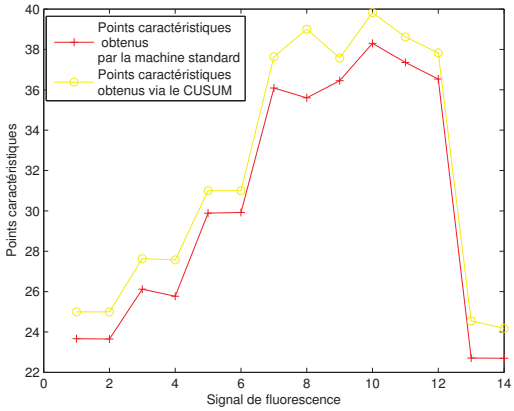


FIGURE 8 – Comparaison des points caractéristiques : ceux obtenus par la méthode de la dérivée seconde contre ceux que nous obtenons par la méthode du CUSUM.

sans la valeur courante (éq. 13).
Les résultats obtenus sont donnés sur le tableau 2.

$$correction(i) = \frac{1}{N-1} \sum_{j=1;j\neq i}^{N-1} difference(j), \forall\, 1 \leq i \leq N, \tag{13}$$

où N est le nombre de signaux avec rupture et $difference$ est la différence entre nos cycles obtenus ($\hat{C}_{CUS}$) et ceux trouvés par la méthode de la dérivée seconde ($\hat{C}_{\partial^2}$).

$$difference(i) = \hat{C}_{CUS}(i) - \hat{C}_{\partial^2}(i), \forall\, 1 \leq i \leq N.$$

Ainsi, nous déterminons les points caractéristiques corrigés ($\hat{C}_t$ corrigé) en retirant les valeurs de la correction à nos

Signaux	$\hat{C}_{\partial^2}$	t_a	r	$\hat{C}_{CUP}$	$\hat{C}_{CUS}$
1	23.22	25	25	24.79	25
2	23.57	25	25	24.63	24.79
3	26.96	28	28	26.63	28
4	26.86	28	28	26.75	28
5	28.72	30	30	30.99	30
6	28.67	30	30	29.63	29.85
7	34.64	36	36	35.67	36
8	34.80	36	36	35.83	36
9	36.06	38	38	37.58	37.75
10	35.89	37	37	37.09	36.91
11	35.53	37	37	36.62	36.80
12	35.68	37	36	35.58	36.17
13	36.10	37	37	33.48	37
14	36.23	38	38	37.62	37.79
15	36.01	37	37	38.36	37
16	35.85	37	37	36.53	36.74
17	22.05	24	24	23.72	24
18	22.10	24	24	23.54	23.78
19	0	45 (=n)	0	0	0
20	0	45 (=n)	0	0	0

TABLE 1 – Tableau des différents points caractéristiques obtenus en appliquant le CUSUM (2 dernières colonnes) et ceux de l'appareil d'enregistrement (2^{ieme} colonne). r est l'instant de rupture déterminé par la méthode du CUSUM ; t_a représente le temps d'arrêt. Les deux derniers signaux n'ont pas de points caractéristiques. Ce sont des signaux sans rupture.

cycles obtenus ($\hat{C}_{CUS}$) :

$$\hat{C}_t \text{ corrigé}(i) = \hat{C}_{CUS}(i) - correction(i), \forall\, 1 \leq i \leq N.$$

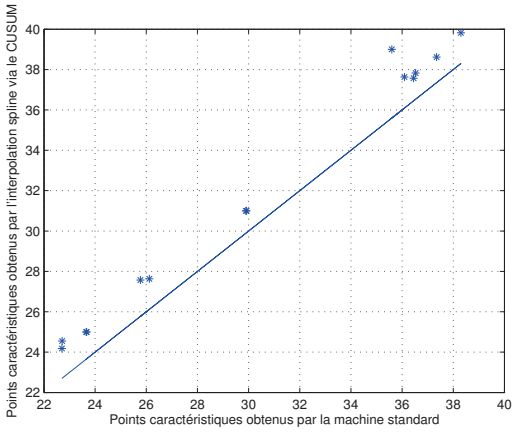
Un point caractéristique est obtenu avec un retard de 1.39 cycles en moyen. En se basant sur nos objectifs correspondant aux tolérances des appareils de mesure industriel (tab. 4). Nous pouvons donc conclure que nos points caractéristiques respectent les performances à atteindre en terme de retard à la détection et de précision du $\hat{C}_t$.

Avant la correction, les points caractéristiques que nous obtenons sont au dessus de ceux de l'appareil standard de 1.13 cycle en moyen et le retard à la détection est de 1.39 cycle en moyen (dernière ligne).

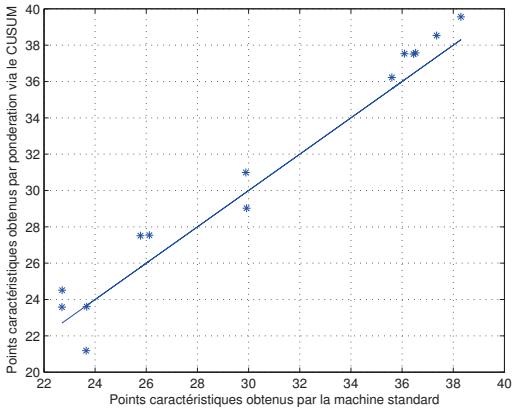
Après correction (tab. 2 et tab. 3) nous sommes de 0.26 cycle en moyen au dessus des points caractéristiques obtenus par la méthode de la dérivée seconde et de 1.65 cycles en moyen de retard à la détection.

	Minimum	Maximum	Moyenne
$\lvert\hat{C}_{\partial^2}$ - $\hat{C}_t$ corrigé $\rvert$, précision du cycle seuil :	0.01 cycle	0.73 cycle	0.26 cycle

TABLE 3 – Résumé, tableau de précision des points caractéristiques obtenus par la méthode du CUSUM.



(a) Points caractéristiques obtenus par interpolation spline par rapport à ceux obtenus par la méthode de la dérivée seconde.



(b) Points caractéristiques obtenus par pondération par rapport à ceux obtenus par la méthode de la dérivée seconde.

FIGURE 9 – Comparaison des points caractéristiques obtenus par pondération et par interpolation.

	Minimum	Objectif	Obtenu
Entre 15 et 35 cycles, précision du cycle seuil	$+-$ 2 ct	$+-$ 0.5 ct	$+-$ 0.26 ct
Information de quantifications après le cycle seuil	10 cycles	4 cycles	1.65 ct

TABLE 4 – Performances à atteindre.

Retard détection	Différence	Correction	Cycle corrigé	Erreur
1.13	1.78	1.22	23.77	0.55
1.34	1.22	1.25	23.53	0.04
1.13	1.03	1.27	26.72	0.24
1.13	1.14	1.26	26.73	0.13
1.13	1.28	1.25	28.74	0.02
1.28	1.18	1.26	28.58	0.09
1.13	1.36	1.25	34.74	0.10
1.13	1.20	1.26	34.73	0.07
1.38	1.69	1.23	36.51	0.45
1.22	1.02	1.27	35.63	0.26
1.33	1.27	1.25	35.54	0.01
1.96	0.49	1.30	34.8	0.18
1.13	0.90	1.27	35.72	0.38
1.34	1.56	1.23	36.55	0.32
1.13	0.99	1.27	35.72	0.29
1.39	0.89	1.27	35.46	0.39
1.13	1.95	1.21	22.78	0.73
1.35	1.68	1.23	22.54	0.44
1.39	**1.13**	...	...	**0.26**

TABLE 2 – Tableau des différences entre les points caractéristiques obtenus par la méthode de la dérivée seconde ($\hat{C}_{\partial^2}$) et les notres, ainsi que les points caractéristiques obtenus après correction ($\hat{C}_t$ corrigé). Les valeurs en gras représentent les moyennes de la colonne ; retard détection = $t_a - \hat{C}_{CUS}$; différence = $|\hat{C}_{\partial^2} - \hat{C}_{CUS}|$; cycle corrigé = $\hat{C}_t$ corrigé ; erreur = $|\hat{C}_{\partial^2} - \hat{C}_t$ corrigé$|$.

5 Conclusion et perspectives

La détermination d'un point de montée du signal de fluorescence peut être faite en temps continu et sans connaissance au préalable du comportement total du signal. Pour cela, nous avons utilisé la méthode du CUSUM qui est une technique de détection de rupture. Nous obtenons des points caractéristiques avec un biais constant par rapport à ceux fournis par l'appareil de mesure standard tout en minimisant le retard à la détection. Pour corriger ce de biais, nous avons alors utilisé une technique d'apprentissage à savoir la validation croisée. Avec cette méthode nous avons obtenu des résultats satisfaisants avec 1.65 cycles de retard à la détection.

Il reste à étendre l'algorithme en travaillant directement sur les images de fluorescences au lieu de la moyenne spatiale de ces dernières. Ce qui facilitera notamment l'estimation de la variance pour notre modèle statistique en supposant le signal ergodique.

REMERCIEMENTS

Le projet Genetic EquipemeNt for biothrEat environmental Analysis and SurveillancE (GENEASE) dont est issu ce travail est financé par l'Agence Nationale de la Rercherche [3]. Il est composé de Bertin technologies [4], du CEA LETI [5], du CEA SBTN [6] et du LITIS ; le LHVP [7] en est un prescripteur.

3. http://www.agence-nationale-recherche.fr/

Références

[1] M. BASSEVILLE et I.V. NIKIFOROV : *Detection of Abrupt Changes : Theory and Application (Prentice Hall information and system sciences series)*. Prentice Hall, 1993.

[2] J. BERTHIER et P. SILBERZAN : *Microfluidics for Biotechnology, Second Edition*. Artech House, 2009.

[3] S.A. BUSTIN, V. BENES, J.A. GARSON, J. HELLEMANS, J. HUGGETT, M. KUBISTA, R. MUELLER, T. NOLAN, M.W. PFAFFL et G.L. SHIPLEY : The MIQE guidelines : minimum information for publication of quantitative real-time PCR experiments. *Clinical chemistry*, 55(4):611–622, 2009.

[4] R.O. DUDA, P.E. HART et D.G. STORK : *Pattern Classification (2nd Edition)*. Wiley-Interscience, 2000.

[5] M. KUBISTA, J.M. ANDRADE, M. BENGTSSON, A. FOROOTAN, J. JONÁK, K. LIND, B. Sjögreen L. Strömbom A. Stahlberg SINDELKA, R. Sjöback et N. ZORIC : The real-time Polymerase Chain Reaction. *Molecular aspects of medicine*, 27(2-3):95–125, 2006.

[6] K.J. LIVAK et T.D. SCHMITTGEN : Analysis of relative gene expression data using real-time quantitative PCR and the $2^{-\Delta\Delta C_T}$ method. *Methods*, 25(4):402–408, 2001.

[7] K.B. MULLIS et F.A. FALOONA : Specific synthesis of dna in vitro via a polymerase-catalyzed chain reaction. *Methods in enzymology*, 155:335–350, 1987.

[8] E.S. PAGE : Continuous inspection schemes. *Biometrika*, 41(1-2):100–115, 1954.

[9] R.G. RUTLEDGE : Sigmoidal curve-fitting redefines quantitative real-time PCR with the prospective of developing automated high-throughput applications. *Nucleic acids research*, 32(22):e178, 2004.

[10] A. SCHAEFER, M. JUNG, H.J. MOLLENKOPF, I. WAGNER, C. STEPHAN, F. JENTZMIK, K. MILLER, M. LEIN, G. KRISTIANSEN et K. JUNG : Diagnostic and prognostic implications of microRNA profiling in prostate carcinoma. *International journal of cancer*, 126(5):1166–1176, 2010.

[11] P. TASSI : *Méthodes statistiques*. Economica Paris, 1985.

[12] C. TSE et J. CAPEAU : Quantification des acides nucléiques par PCR quantitative en temps réel. *Ann Biol Clin*, 61(3):279–293, 2003.

4. Bertin technologies http://www.bertin.fr/

5. Commissariat à l'Energie Atomique - Laboratoire d'Electronique et de Technologies de l'Information, http://www.leti.fr

6. Commissariat à l'Energie Atomique - Service de Biochimie et Toxicologie Nucléaire, http://www-dsv.cea.fr/sbtn/

7. Laboratoire d'Hygiène de la Ville de Paris