

Cours sur les
Séries chronologiques

Par
Yakoub BOULAROUK

Table des matières

| | | |
|----------|---|-----------|
| 1 | Introduction | 3 |
| 1.1 | Exemples de séries chronologiques | 4 |
| 1.2 | Objectifs de l'analyse des séries chronologiques | 4 |
| 1.3 | Modèles de séries chronologiques | 8 |
| 1.3.1 | Modèles additifs | 9 |
| 1.3.2 | Procédure de modélisation de séries chronologiques | 11 |
| 1.4 | Modèles stationnaires et fonction d'autocorrélation | 12 |
| 1.4.1 | Modèles stationnaires | 12 |
| 1.4.2 | Fonction d'autocorrélation | 14 |
| 1.5 | Tendance et saisonnalité | 15 |
| 1.5.1 | Élimination de la tendance en l'absence de saisonnalité | 17 |
| 1.5.2 | Élimination de la tendance et de la saisonnalité | 20 |
| 1.6 | Exercices | 22 |
| 2 | Processus linéaires | 24 |
| 2.1 | Processus stationnaires | 24 |

| | | |
|----------|--|-----------|
| 2.2 | Introduction aux processus ARMA | 26 |
| 2.3 | Processus linéaire | 28 |
| 2.4 | Estimations des fonctions caractéristiques | 31 |
| 2.4.1 | Estimation de la moyenne | 31 |
| 2.4.2 | Estimation de la fonction d'autocovariance | 33 |
| 2.5 | Exercices | 35 |
| 3 | Modèles ARMA. | 37 |
| 3.1 | Processus ARMA | 37 |
| 3.2 | La fonction d'autocovariance et la fonction d'autocorrélation . | 42 |
| 3.2.1 | La fonction d'autocovariance | 43 |
| 3.2.2 | La fonction d'autocorrélation | 44 |
| 3.2.3 | La fonction d'autocorrélation partielle | 45 |
| 3.3 | Estimation des paramètres | 48 |
| 3.3.1 | Estimateur de Yule Walker | 48 |
| 3.3.2 | La fonction de vraisemblance pour l'ARMA | 49 |
| 3.3.3 | L'estimateur du Maximum de vraisemblance | 51 |
| 3.4 | Le critère d'information d'Akaike corrigé | 52 |
| 3.5 | Exercices | 53 |
| 4 | Prévision des processus ARMA | 57 |
| 4.1 | Choix du prédicteur | 58 |
| 4.2 | Prévision des processus autorégressif | 60 |
| 4.3 | Prévision des processus autorégressif à moyenne mobile | 64 |

4.4 Exercices 65

Table des figures

| | | |
|-----|--|----|
| 1.1 | Population des Etats-Unis 1790-1980. | 4 |
| 1.2 | Grèves aux États-Unis | 5 |
| 1.3 | Résultats des matchs de baseball, star, 1933-1995. | 5 |
| 1.4 | Nombre mensuel de décès accidentels aux USA 1973-1978. . . | 6 |
| 1.5 | Problème de détection de signal. | 6 |
| 1.6 | Fenêtre glissante avec un pas de $2q + 1 = 5$ | 19 |
| 1.7 | Estimation par moyenne mobile de la tendance de la figure 1.4. | 19 |
| 3.1 | Dow Jones | 46 |
| 3.2 | Autocorrélations partielle | 47 |

Chapitre 1

Introduction

Dans ce chapitre, nous présentons quelques idées de base sur l'analyse des séries chronologiques (temporelles) et les processus stochastiques. Les concepts de stationnarité de l'autocovariance et des fonctions d'autocovariance empiriques sont particulièrement importants. Certaines techniques standard pour l'estimation, l'élimination de la tendance et de la saisonnalité (d'une période connue) sont décrites pour une série chronologique observée. Celles-ci sont illustrées par référence aux ensembles de données de la section 1.1. Les calculs dans tous les exemples peuvent être effectués à l'aide du logiciel R dont la version téléchargeable est disponible sur le site internet <https://www.r-project.org/>. La plupart des sujets abordés dans ce chapitre seront développés plus en détail dans les sections suivantes du cours.

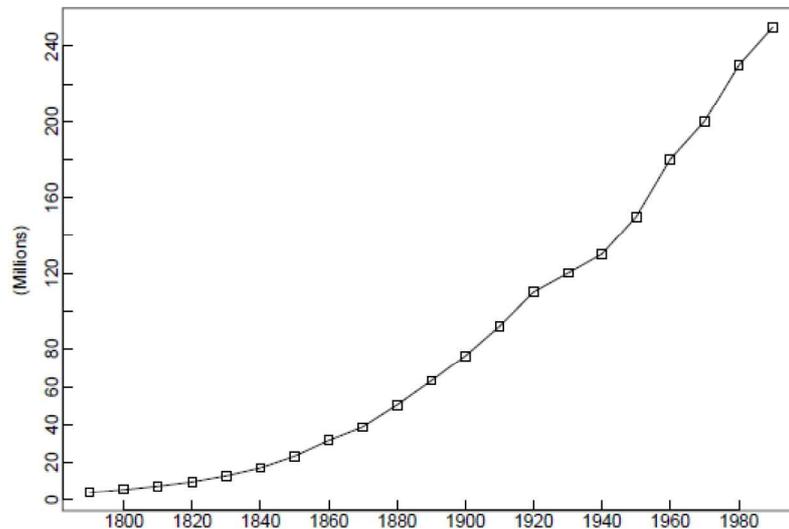


FIGURE 1.1 – Population des Etats-Unis 1790-1980.

1.1 Exemples de séries chronologiques

Dans cette partie, nous présentons quelques séries chronologiques célèbres qui vont nous servir d'exemples de séries sur lesquelles nous appliquons les différentes techniques expliquées dans ce cours.

1.2 Objectifs de l'analyse des séries chronologiques

Les exemples exposés dans les figures (1,2,3,4,5) représentent un échantillon de séries chronologiques rencontrées dans les domaines de l'ingénierie, de la science, de la sociologie et de l'économie. Notre but dans ce cours est

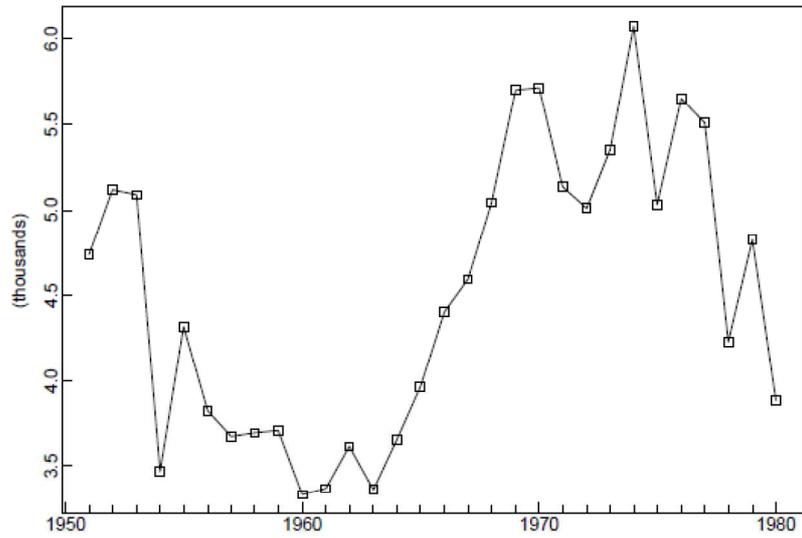


FIGURE 1.2 – Grèves aux États-Unis

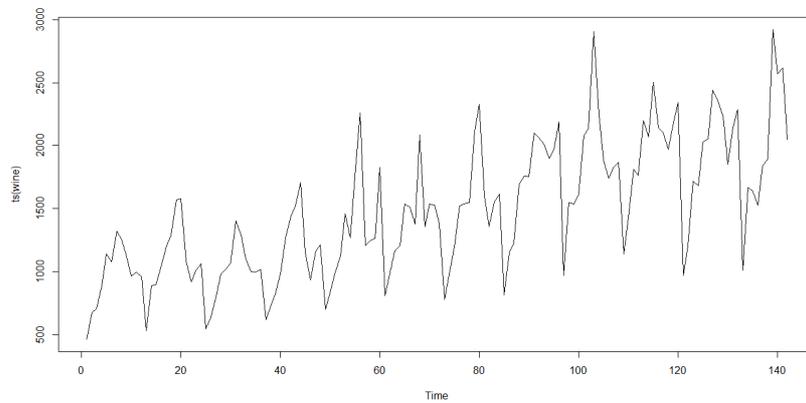


FIGURE 1.3 – Résultats des matchs de baseball, star, 1933-1995.

d'étudier des techniques pour tirer des inférences de telles séries. Avant cela, cependant, il est nécessaire de mettre en place un modèle de probabilité hypothétique pour représenter les données. Après avoir choisi une famille ap-

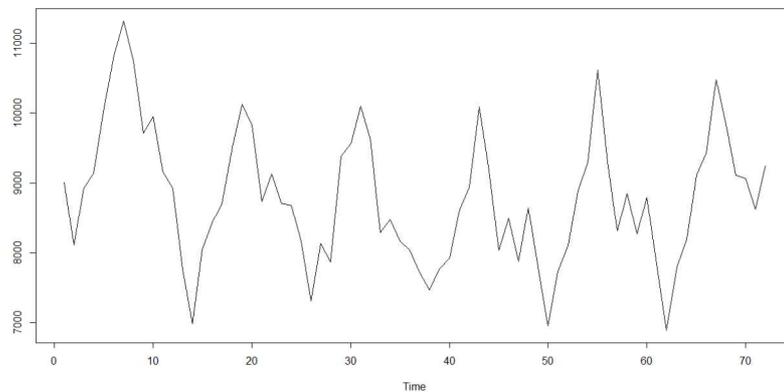


FIGURE 1.4 – Nombre mensuel de décès accidentels aux USA 1973-1978.

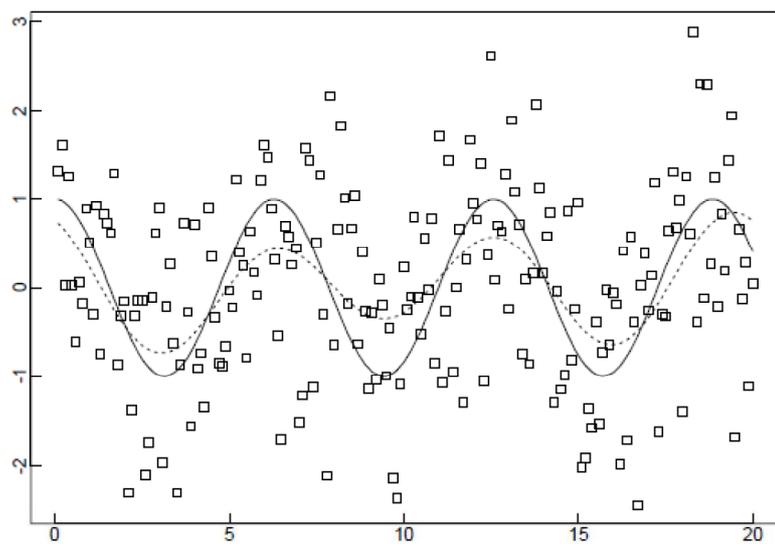


FIGURE 1.5 – Problème de détection de signal.

propriété de modèles, il est alors possible d'estimer des paramètres, de vérifier la qualité de l'ajustement aux données et, éventuellement, d'utiliser le modèle ajusté pour améliorer notre compréhension du mécanisme générant la série.

Une fois qu'un modèle satisfaisant ait été développé, il peut être utilisé de diverses façons selon le domaine d'application particulier.

Le modèle peut être utilisé simplement pour fournir une description compacte des données. Par exemple, nous pouvons représenter les données sur les décès accidentels de la figure (1.4) comme étant la somme d'une tendance donnée et des termes saisonniers et aléatoires. Pour l'interprétation des statistiques économiques telles que les chiffres du chômage, il est important de reconnaître la présence de composantes saisonnières et de les supprimer afin de ne pas les confondre avec les tendances à long terme. Ce processus est connu comme ajustement saisonnier. D'autres applications des modèles de séries temporelles incluent la séparation (ou le filtrage) du bruit des signaux comme dans la figure (1.6). En testant des hypothèses telles que le réchauffement planétaire à l'aide de données de température enregistrées, en prévision d'une série à partir d'observations d'une autre, par exemple en prédisant les futures ventes en utilisant les données de dépenses publicitaires et en contrôlant les valeurs futures d'une série en ajustant les paramètres. Les modèles de séries chronologiques sont également utiles dans les études de simulation. Par exemple, la performance d'un réservoir dépend fortement des apports quotidiens aléatoires d'eau au système. Si ceux-ci sont modélisés comme une série temporelle, alors nous pouvons utiliser le modèle ajusté pour simuler un grand nombre de séquences indépendantes d'entrées quotidiennes. Connaissant la taille et le mode de fonctionnement du réservoir, nous pouvons déterminer la fraction des séquences d'entrée simulées qui font

que le réservoir s'épuise dans une période de temps donnée. Cette fraction sera alors une estimation de la probabilité de vide du réservoir à un certain moment dans la période donnée.

1.3 Modèles de séries chronologiques

Une partie importante de l'analyse d'une série chronologique est la sélection d'un modèle stochastique (ou d'une classe de modèles) approprié pour les données. Pour tenir compte de la nature éventuellement imprévisible des observations futures, il est naturel de supposer que chaque observation x_t est une valeur réalisée d'une certaine variable aléatoire X_t .

Définition 1 *Une série chronologique est un ensemble d'observations x_t d'un processus X_t , chacune étant enregistrée à un instant t . On peut distinguer deux types de séries chronologiques :*

- *Les séries temporelles à temps discrets (le type auquel ce cours est principalement consacré) est celle dans laquelle l'ensemble T des instants auxquels les observations sont faites est un ensemble discret, par exemple, lorsque des observations sont faites à des intervalles de temps fixes.*
- *Les séries temporelles continues obtenues lorsque les observations sont enregistrées en continu sur un certain intervalle de temps.*

Définition 2 *Un modèle de série chronologique est une équation précisant la façon dont les composantes s'articulent les unes par rapport aux autres pour constituer la série chronologique.*

Ils existent de très nombreux modèles et parmi eux le modèle additif auquel nous nous limiterons.

1.3.1 Modèles additifs

Nous considérons dans cette section une série y_t admettant une décomposition additive

$$Y_t = f_t + S_t + X_t, t = 1 \dots n, \quad (1.3.1)$$

où f_t est la composante tendancielle, S_t la composante saisonnière et Q_t représente l'erreur ou l'écart au modèle.

Comme nous l'avons dit en introduction,

- la tendance f_t exprime un mouvement à moyen terme de la série. Elle est le plus souvent modélisée par une fonction polynomiale du temps.
- la composante saisonnière S_t exprime un phénomène qui se reproduit de manière analogue sur chaque intervalle de temps successif. L'étendue de cet intervalle qui est constante est appelée période et sera notée P dans la suite. La plupart du temps, nous supposons que la composante saisonnière est constante sur chaque période P , c'est-à-dire

$$S_{t+P} = S_t, \forall t.$$

Cela revient à dire que l'effet net du saisonnier sur une période est nul ; ce qui est naturel puisqu'il est repris dans la tendance générale de la série chronologique. Il s'agit là du modèle le plus simple dans lequel le saisonnier est caractérisé par P coefficients c_1, \dots, c_P . Lorsque $P = 4$, la série est trimestrielle ; lorsque $P = 12$, la série est mensuelle... On suppose par ailleurs que l'effet du saisonnier est en moyenne nul sur une période, ce qui signifie que

$$\sum_{i=1}^P c_i = 0,$$

— les erreurs sont des variables aléatoires centrées.

Remarque 1 *Dans ce modèle, l'amplitude de la série reste constante au cours du temps. Ceci se traduit graphiquement par des fluctuations autour de la tendance f_t constantes au bruit près.*

Imaginons que nous étudions la série des températures moyennes relevées chaque mois à la wilaya de Mila depuis janvier 2007. Que peut-on dire des composantes présentes ?

- la série f_t représente la tendance générale (réchauffement ? cycle ?).
- les données étant mensuelles, la période est donc un an et $p = 12$.
- des valeurs $S_1 = -1$ et $S_6 = 28$ signifient que le mois de janvier est plus froid de 29° par rapport à l'ensemble de l'année alors que le mois de juin est plus chaud de 29° .
- une fluctuation irrégulière $x_{14} = -2$ signifie qu'il a fait 2° de moins

que prévu pour un mois de février en 2008 (c'est-à-dire ce que nous laissaient prévoir la tendance et l'effet saisonnier pour février 2008).

1.3.2 Procédure de modélisation de séries chronologiques

- Tracer la série et examiner les principales caractéristiques du graphe, en vérifiant notamment s'il existe
 1. une tendance,
 2. une composante saisonnière,
 3. tout changement brusque évident dans le comportement,
 4. toute observation éloignée.
- Enlever les composantes de tendance et saisonnières pour obtenir des résidus stationnaires. Pour atteindre cet objectif, il peut parfois être nécessaire d'appliquer une transformation préliminaire aux données. Par exemple, si l'ampleur des fluctuations semble croître de façon approximativement linéaire avec le niveau de la série, par exemple la série représenté à la figure 1.1, alors la série transformée $\log x_1, \dots, \log x_n$ aura des fluctuations d'une magnitude plus constante. Si certaines données sont négatives, ajoutez une constante positive à chacune des valeurs de données pour vous assurer que toutes les valeurs sont positives avant de prendre des logarithmes. Il existe plusieurs façons de supprimer la tendance et la saisonnalité, certaines impliquant l'esti-

mation des composantes et en les soustrayant des données, et d'autres en fonction de la différence des données, c'est-à-dire en remplaçant la série originale Y_t par $Z_t = Y_t - Y_{t-d}$ pour un entier positif d . Quelle que soit la méthode utilisée, l'objectif est de produire une série stationnaire centrée.

- Choisir un modèle pour adapter les résidus, en utilisant diverses statistiques empiriques comprenant la fonction d'autocorrélation empirique.
- La prévision sera réalisée en prévision de processus centré, puis en inversant les transformations décrites ci-dessus pour arriver aux prévisions de la série originale Y_t .

1.4 Modèles stationnaires et fonction d'autocorrélation

1.4.1 Modèles stationnaires

Une série chronologique $(X_t)_{t \in \mathbb{Z}}$ est dite stationnaire si elle a des propriétés statistiques semblables à celles de la série "décalée dans le temps" $(X_{t+h})_{t \in \mathbb{Z}}$, pour chaque entier h . En restreignant l'attention sur les propriétés qui ne dépendent que des moments de premier et second ordre de X_t , nous pouvons rendre cette idée précise avec les définitions suivantes.

Définition 3 Soit X_t un processus avec $E(X_t^2) < \infty$.

La moyenne est

$$\mu_X(t) = E(X_t)$$

La fonction d'autocovariance est

$$\gamma_X(r, s) = \text{Cov}(X_r, X_s) = E[(X_r - E(X_r))(X_s - E(X_s))] \quad \forall r, s \in \mathbb{Z}.$$

Définition 4 Soit X_t un processus stochastique,

— X_t est stationnaire au second ordre (faiblement stationnaire) si

1. $\mu_X(t)$ est indépendant de temps.
2. $\gamma_X(t+h, t)$ est indépendante de temps.

X_t est strictement stationnaire (fortement stationnaire) si

(X_1, \dots, X_n) et $(X_{1+h}, \dots, X_{n+h})$ ont la même loi de probabilité.

Remarque 2 Il est facile de vérifier que si X_t est strictement stationnaire et $E(X_t^2) < \infty$ pour tout t , alors X_t est également faiblement stationnaire.

Exemple 1.4.1 L'exemple le plus connu de processus stationnaire est le processus bruit blanc. Un Bruit Blanc est une suite de v.a.r. ε_t telle que :

$$E(\varepsilon_t) = 0 \quad \forall t \in T$$
$$\gamma_\varepsilon(h) = E(\varepsilon_t \varepsilon_{t-h}) = \begin{cases} \sigma_\varepsilon^2 & \text{si } h = 0 \\ 0 & \text{si } h \neq 0 \end{cases}$$

Il s'agit d'une suite de v.a.r. homoscédastiques et non autocorrélées (et même indépendantes, c'est pourquoi on parle aussi de processus i.i.d. pour indépendamment et indépendamment distribué).

Remarque 3 Vu la condition (2) de la définition (4), quand on utilise le terme "Fonction de covariance" par rapport à un processus stationnaire X_t , on entend la fonction γ_X d'une variable, définie par :

$$\gamma_X(h) = \gamma_X(h, 0) = \gamma_X(t + h, t)$$

La fonction $\gamma_X(\cdot)$ sera appelée fonction d'autocovariance et $\gamma_X(h)$ sa valeur au retard h .

1.4.2 Fonction d'autocorrélation

Les principales caractéristiques temporelles d'un processus sont données par l'autocorrélation (simple) et l'autocorrélation partielle.

Définition 5 Soit X_t est un processus stationnaire,

— La fonction d'autocovariance de X_t au retard h est

$$\gamma_X(h) = \text{Cov}(X_{t+h}, X_t)$$

— La fonction d'autocorrélation de X_t au retard h

$$\text{cor}(X_{t+h}, X_t) = \rho_X(h) = \frac{\gamma_X(h)}{\gamma_X(0)}$$

1.5 Tendance et saisonnalité

Lorsqu'on souhaite analyser une série chronologique représentée sous la forme d'un tableau de chiffres, il faut tout d'abord représenter la série sous la forme d'un graphe. Dans les représentations des figures (1.1) et (1.2) des pages 6 et 7, on peut observer dans certains cas une composante déterministe qui se présente sous la forme d'une tendance ou d'un cycle saisonnier. Une série chronologique peut également présenter ces deux comportements en même temps.

Bien entendu, cette tendance ou cette saisonnalité ne sont pas toujours faciles à déceler. Il peut même arriver que des séries ne contiennent aucune de ces composantes déterministes. Il se peut aussi que la tendance n'apparaisse qu'après avoir transformé les données par une fonction, par exemple, logarithmique.

Nous supposons que toute série chronologique Y_t puisse être mise sous la forme de la décomposition vue déjà dans 1.3.1

$$Y_t = f_t + S_t + X_t$$

où :

- f_t est appelé trend (ou tendance) et est soit une fonction décrite par un nombre fini de paramètres, par exemple, une fonction linéaire du temps $a+bt$ ou un polynôme en t , soit, plus généralement, une fonction lisse en temps t .

- S_t est appelé saisonnalité et est une fonction périodique. Soit d la période de S_t , on a alors : $S_t = S_{t+d}$ pour tout t . On supposera également que S_t ne contient pas de tendance, ce qu'on écrira :

$$\sum_{j=1}^d S_{t+j} = 0$$

En d'autres termes, toute la tendance est modélisée dans f_t .

- $f_t + S_t$ est la composante déterministe du modèle.
- X_t est, par opposition, la composante aléatoire, supposée de moyenne nulle (parce qu'elle représente les composantes (erreurs) non systématiques), mais possédant en général une structure de corrélation non nulle. C'est bien évidemment la partie la plus intéressante à modéliser. Nous verrons comment, sous certaines hypothèses, nous pourrons ajuster un modèle pour X_t qui permettra de prévoir les futures valeurs de la série chronologique. Notons que, si X_t est de moyenne non nulle, on pourra remplacer f_t , par $f_t + EX_t$, et X_t sera remplacé par $W_t = X_t - EX_t$. On a alors $EW_t = 0$.

Dans un premier temps, nous allons estimer la composante déterministe du signal. De cette façon, nous n'aurons plus qu'à nous intéresser à la partie aléatoire qui constitue le véritable enjeu de ce cours.

1.5.1 Elimination de la tendance en l'absence de saisonnalité

Supposons, dans un premier temps, que la partie déterministe du modèle soit uniquement composée d'une tendance f_t :

$$Y_t = f_t + X_t \quad (1.5.1)$$

Dans cette section, nous donnons trois méthodes pour estimer f_t .

Estimation paramétrique

Cette méthode consiste à estimer la tendance par la méthode des moindres carrés ordinaires. Supposons que l'on observe la série de la figure (1.3) de la page 8. Il semble naturel d'estimer la composante f_t de cette série par une fonction linéaire

$$\hat{f}_t = \hat{a} + \hat{b}t$$

où \hat{a} et \hat{b} sont des estimateurs des coefficients de la fonction linéaire estimant f_t . Pour trouver ces estimateurs par la méthode des moindres carrés, il faut minimiser l'erreur quadratique commise en remplaçant f_t par \hat{f}_t dans 1.5.1. En conséquence, si on observe la série y_1, \dots, y_n , il faut trouver les coefficients \hat{a} et \hat{b} minimisant

$$\sum_{t=1}^n (y_t - at - b)^2 \quad (1.5.2)$$

La solution de ce problème de minimisation est,

$$\begin{aligned}\hat{a} &= \frac{\text{Cov}(Y, T)}{\sigma_T} \\ \hat{b} &= \bar{Y} - \hat{a}\bar{T}\end{aligned}$$

Estimation non paramétrique

Dans certaines situations, il n'est pas facile de trouver le degré du polynôme d'ajustement pour f_t . Par exemple, dans la figure (1.2), il n'est pas possible d'utiliser la méthode des moindres carrés car le polynôme utilisé au départ pour \hat{f}_t n'est ni linéaire, ni quadratique. On pourrait utiliser un polynôme avec un degré élevé, mais le nombre de paramètres à estimer serait important et rendrait les calculs fastidieux. Dans cette situation, on a recourt à la théorie non paramétrique de l'estimation de la tendance, qui ne la suppose pas polynomiale a priori. Pour comprendre cette technique, supposons que f_t soit linéaire dans un intervalle $[t - q, t + q]$. Dans ce cas, un bon estimateur du trend est donné par

$$\hat{f}_t = \frac{1}{2q + 1} \sum_{j=-q}^q y_{t+j}$$

On peut à présent calculer \hat{f}_t pour chaque valeur de t en calculant cette moyenne sur les $2q$ observations autour de t . On peut aussi voir l'intervalle $[t - q; t + q]$ comme une "fenêtre" sur les observations, que l'on déplace lorsque t varie. A chaque valeur de t , l'estimateur \hat{f}_t calcule la moyenne des obser-

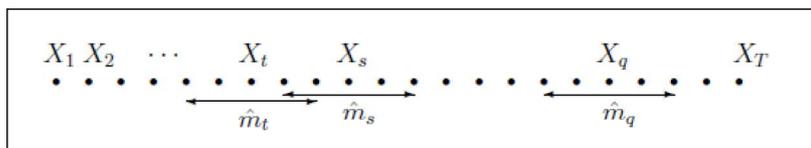


FIGURE 1.6 – Fenêtre glissante avec un pas de $2q + 1 = 5$.

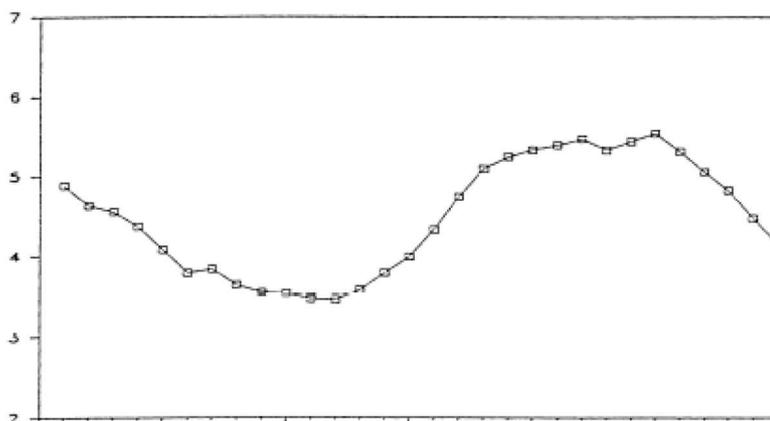


FIGURE 1.7 – Estimation par moyenne mobile de la tendance de la figure 1.4.

vations tombant dans cette fenêtre glissante (on dit alors qu'on effectue une estimation par moyenne mobile). Pour éviter que des problèmes se posent aux bords de la série (lorsque $t \leq q$ ou $t > n - q$), on pose :

$$\begin{cases} X_t = X_1, \text{ lorsque } t < 1 \\ X_t = X_n, \text{ lorsque } t > n \end{cases}$$

1.5.2 Elimination de la tendance et de la saisonnalité

Supposons à présent que, en plus d'une tendance f_t , la série comporte une composante saisonnière S_t . On considère donc le modèle

$$Y_t = f_t + S_t + X_t$$

où, comme nous l'avons spécifié plus haut,

- S_t est une fonction périodique en t $S_t = S_{t+d}$ pour un certain $d > 0$
- S_t ne comporte pas de composante "tendance", et donc :

$$\sum_{t=1}^d S_{t+j} = 0$$

Voyons comment éliminer la composante saisonnière sur l'exemple de la figure (1.4) page 8, dans lequel nous observons des données mensuelles dont la période est visiblement $d = 12$. Dans cet exemple, il est utile de réorganiser les données $(y_t)_{t=1,n}$ dans un tableau dont les entrées sont les années (1973 à 1978) et le mois (1 à 12) :

$$(y_t)_{t=1,n} = (y_{jk})_{j=1..6(\text{ans}) \text{ et } k=1,\dots,12(\text{mois})}$$

avec la relation : $y_{jk} = y_{k+12(j-1)}$.

Dans la figure 1.4, on observe que le tendance ne varie pas beaucoup et il semble naturel de le supposer constant sur une année. Un estimateur naturel

de la tendance pour l'année j est donc :

$$\hat{f}_j = \frac{1}{12} \sum_{k=1}^{12} y_{jk} \quad j = 1 \dots 6$$

Remarquons que cette composante saisonnière ne dépend que du mois et non de l'année. En conséquence, un estimateur de cette composante S_j est donné par la moyenne empirique sur le même mois de chaque année des résidus :

$$\hat{S}_k = \frac{1}{6} \sum_{j=1}^6 (y_{jk} - \hat{f}_{jk}) \quad j = 1 \dots 12$$

La méthode que nous venons d'appliquer sur un exemple peut être généralisée au cas d'une période non nécessairement égale à 12. Elle nécessite toutefois une hypothèse sur le comportement du trend, qui doit varier faiblement.