

# Étonnante précision de la méthode des moindres carrés pour des séries chronologiques issues de modèles linéaires fortement perturbés

Stéphane Junca(\*)

## 1. Introduction

L'étude de séries statistiques en classe demande souvent du temps pour rentrer les données. Ce qui nous limite dans la taille des échantillons. Pour gagner du temps et traiter très rapidement des exemples de plus grande taille, j'utilise dans mes classes de préparation au C.A.P.E.S. des séries de la forme  $(k, k + e_k)_{k=1}^n$ , facile à rentrer et à illustrer graphiquement avec ma calculatrice à écran rétroprojectable. Pour des « petites » perturbations  $e_k$ , on n'est pas surpris de retrouver précisément la pente 1 avec la méthode des moindres carrés. On est alors tenté de prendre des perturbations de plus en plus grandes. À la grande surprise de ma classe, la méthode des moindres carrés résiste très bien à ce genre de traitement. Pour éclaircir ce mystère, cet article propose de nombreux exemples et des explications mathématiques et historiques de cette étonnante stabilité de la méthode des moindres carrés.

Il faut savoir que cette méthode comprend bien plus que le problème de l'ajustement affine. En général, il s'agit de trouver  $p$  paramètres solutions d'un système linéaire rectangulaire. En pratique, on a beaucoup plus d'équations que d'inconnues et il s'agit de trouver la solution au sens des moindres carrés [10].

Le cas  $p = 1$  est déjà fait dès le Collège sans bien sûr le présenter de cette manière. En effet lorsque l'on choisit de n'associer à une série statistique  $(x_k)_{k=1}^n$  qu'un seul nombre représentatif : la moyenne  $m$ , on résout le système linéaire *surdéterminé* :  $m = x_k, k = 1, \dots, n$  à une inconnue et avec  $n$  équations au sens des moindres carrés. C'est-à-dire que l'on prend l'unique nombre  $m$  qui minimise l'écart quadratique

$$\text{moyen : } \frac{1}{n} \sum_{k=1}^n (x_k - m)^2.$$

Le cas  $p = 2$  n'est fait qu'au Lycée pour certaines Terminales dans le cadre d'ajustement affine d'un nuage de points  $(x_k, y_k)_{k=1}^n$ . Il en est ainsi pour la section ES, des séries technologiques (par exemple la série sciences et technologies de la gestion) et des séries professionnelles. L'utilisation croissante de matériel

(\*) IUFM et Université de Nice, Laboratoire J. A. Dieudonné, UMR CNRS 6621.

informatique ne fera que renforcer cette tendance. D'autant plus que son utilisation est très courante après le bac en coordonnées semilog ou log-log, en physique, en chimie, en biologie, en économie, en sciences humaines, ... Pour le problème de l'ajustement affine, les inconnues sont les coefficients de la droite  $(\alpha, \beta)$ , et les équations sont  $y_k = \alpha x_k + \beta$ ,  $k = 1, \dots, n$  avec  $n > 2$ . Une fois encore la solution s'obtient en minimisant la moyenne des écarts quadratiques  $(y_k - (\alpha x_k + \beta))^2$ . Ainsi, la méthode nous fournit toujours une droite. Mais cette droite est-elle bien pertinente ? Par exemple, si le nuage de points est un échantillon de points d'une parabole, la droite fournie est sans intérêt. Il faut donc avoir plus d'informations sur le nuage de points. Dans cet article nous ne traiterons que des cas issus de perturbations du modèle linéaire. De plus nous supposons que la série des abscisses est *arithmétique*, ce qui est fréquent pour des séries chronologiques. Dans ce cadre nous verrons que cette méthode est très efficace et très stable.

Commençons d'abord par bien poser le problème. On suppose que deux caractères statistiques  $y$  et  $x$  sont reliés de manière linéaire ou affine :

$$y = \alpha x + \beta \quad (1)$$

En pratique les mesures de  $y$  en fonction de celles de  $x$  sont perturbées. La perturbation sera représentée par la série  $(e_k)_{k=1}^n$ . On obtient ainsi le modèle linéaire perturbé :

$$y_k := \alpha x_k + \beta + e_k, k = 1, \dots, n \quad (2)$$

À partir de la série statistique  $(x_k, y_k)_{k=1}^n$ , on se propose de retrouver une approximation de  $\alpha$  et  $\beta$  grâce à la célèbre *méthode des moindres carrés (MMC)* découverte par Carl Friedrich Gauss en 1795<sup>(1)</sup>, alors qu'il n'avait que 18 ans ! Gauss avait déjà obtenu l'optimalité (en un certain sens statistique) de la MMC pour estimer les coefficients inconnus  $\alpha$  et  $\beta$  lors de calculs astronomiques. Ainsi, il retrouva quelques années plus tard, à la surprise générale, par le calcul l'astre Cérès<sup>(2)</sup> que les astronomes avaient perdu de vue (au sens propre). Et, à trente ans, il devint le directeur de l'observatoire de l'Université de Göttingen.

Aujourd'hui, l'utilisation des calculatrices et des tableurs nous donne facilement accès à la méthode des moindres carrés. Avec une simple calculatrice on va fabriquer des séries statistiques vérifiant le modèle (2).  $(e_k)_{k=1}^n$  représentera une suite de perturbations déterministes ou aléatoires. Ensuite, on demandera à notre calculatrice de nous fournir une approximation des coefficients de la droite du modèle théorique (1). Vous pourrez apprécier la qualité de l'approximation de la pente  $\alpha$ . Au cours de cet article, on fera des perturbations de plus en plus fortes, pour pousser la méthode jusqu'à ses derniers retranchements. On traitera des cas de perturbations déterministes avec  $e_k = (-1)^k$  ou  $\sin(k\alpha)$ . Ensuite on simulera des perturbations aléatoires indépendantes : le cas envisagé par Gauss. On terminera par des

(1) Cette découverte n'est publiée qu'en 1809 dans [4], soit quatre ans après Adrien-Marie Legendre dans [11]. Il en résulte une querelle de priorité entre Legendre et Gauss.

(2) Cérès est un astéroïde de très grande taille.

perturbations aléatoires non indépendantes avec la loi binomiale. On donnera le point de vue de Legendre et de Gauss, démontrant l'efficacité de leur méthode.

Les démonstrations mathématiques des résultats énoncés sont disponibles sur le site de l'A.P.M.E.P.

## 2. Notations, formules et caractère arithmétique du temps

Pour une série statistique à deux variables  $(x_k, y_k)_{k=1}^n$ , on peut chercher une droite d'équation  $y = ax + b$ , qui minimise au mieux l'écart quadratique moyen

$$R_n(a, b) := \frac{1}{n} \sum_{k=1}^n (y_k - [ax_k + b])^2.$$

Par la suite, on va étudier l'effet de la taille de l'échantillon sur les coefficients de la droite cherchée. On indicera ainsi les coefficients de la droite des moindres carrés par  $n$ . Des calculs classiques montrent que la droite optimale pour ce critère a pour coefficients :

$$a_n := \frac{s_{xy}}{s_x^2}, \quad b_n := \bar{y} - a_n \bar{x}$$

avec les notations usuelles pour les moyennes, les variances, la covariance, le coefficient de corrélation, et la relation entre la formule des résidus et le coefficient de corrélation :

$$\bar{x} := \frac{1}{n} \sum_{k=1}^n x_k, \quad \bar{y} := \frac{1}{n} \sum_{k=1}^n y_k, \quad \bar{e} := \frac{1}{n} \sum_{k=1}^n e_k, \quad s_x^2 := \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})^2, \quad s_y^2 := \frac{1}{n} \sum_{k=1}^n (y_k - \bar{y})^2,$$

$$s_{xy} := \frac{1}{n} \sum_{k=1}^n (x_k - \bar{x})(y_k - \bar{y}), \quad \rho_n := \frac{s_{xy}}{s_x s_y}, \quad R_n = R_n(a_n, b_n) = s_y^2 (1 - \rho_n^2).$$

Pour une série chronologique, la série statistique  $(x_k)_{k=1}^n$ , est bien déterminée, elle représente les différents temps de la mesure du caractère  $y$ . Il arrive souvent que les mesures soient prises à intervalle de temps régulier. Dans ce cas, la série statistique  $(x_k)_{k=1}^n$  est souvent une **suite arithmétique**. Quitte à changer d'unité de temps et d'origine des temps, on supposera pour toute la suite que :

$$x_k = k \quad \text{pour tout } k \tag{3}$$

Le comportement arithmétique de la série  $(x_k)_{k=1}^n$  va jouer un rôle essentiel dans la stabilité de la MMC. Dans ce cas, on peut déjà calculer plus explicitement certains indicateurs statistiques :

$$\bar{x} = \frac{n+1}{2}, \quad s_x^2 = \frac{n^2-1}{12}, \quad s_{xy} = \alpha \times s_x^2 + s_{xe} \quad \text{où } s_{xe} := \left( \frac{1}{n} \sum_{k=1}^n k e_k \right) - \frac{n+1}{2} \bar{e}.$$

### 3. Étude déterministe du modèle affine

On s'intéresse au modèle :  $y = \text{constante}$ . Quitte à faire une translation sur l'axe des  $y$ , on peut supposer, sans perte de généralité, que la constante est nulle. Ainsi, la série  $(y_k)_{k=1}^n$  devient simplement les valeurs des perturbations :

$$y_k = 0 \times k + 0 + e_k = e_k, \quad k = 1, \dots, n.$$

On considère l'exemple suivant  $e_k = (-1)^k$  avec  $n = 10$ .

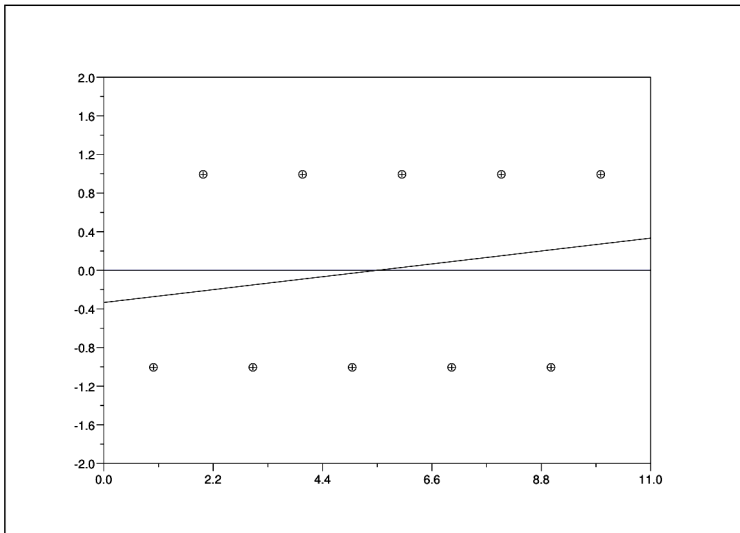


FIG. 1. Perturbation déterministe de taille 1 de  $y = 0$  :  $e_k = (-1)^k$

Dans la figure 1, on a représenté le nuage de points  $(k, (-1)^k)_{k=1}^n$ , la droite horizontale cherchée et la droite trouvée par la MMC. On remarquera l'excellente approximation du modèle réel  $y = 0$  par la droite fournie par la MMC alors que les points sont loin d'être alignés sur la même droite. En effet, la MMC nous fournit la droite :  $y = 0,060\ 606... \times x + 0,333...$ . Ou remarque aussi que la MMC nous fournit le coefficient de la pente avec une erreur de l'ordre de  $10^{-2}$  alors que pour l'ordonnée à l'origine, l'erreur est moins bonne, elle est de l'ordre de  $10^{-1}$ . Pour comprendre ce résultat, une explication de Legendre s'impose ([2, 11]) :

« De tous les principes qu'on peut proposer pour cet objet, je pense qu'il n'en est pas de plus général, de plus exact, ni d'une application plus facile, que celui dont nous avons fait usage dans les recherches précédentes, et qui consiste à rendre minimum la somme des carrés des erreurs. Par ce moyen il s'établit entre les erreurs une sorte d'équilibre qui, empêchant les extrêmes de prévaloir, est très propre à faire connaître l'état du système le plus proche de la vérité. »

Donnons maintenant une justification mathématique :

**Théorème 1 [Perturbation déterministe de  $y = \alpha x + \beta$ ]**

La méthode des moindres carrés, appliquée à la série chronologique  $(k, y_k)_{k=1}^n$  satisfaisant le modèle (2) et (3), fournit la droite d'équation :  $y = a_n x + b_n$ , avec les majorations suivantes :

$$|a_n - \alpha| \leq \frac{18C_n}{n^2 - 1}, \quad |b_n - \beta| \leq \frac{10C_n}{n - 1}, \quad \text{où } C_n = \max_{1 \leq j \leq n} \left| \sum_{k=1}^j e_k \right|.$$

Ainsi, les quantités importantes à contrôler pour estimer la précision de la MMC sont les **sommes des perturbations** plutôt que la taille des perturbations. Lorsque  $C_n$  est majorée indépendamment de  $n$ , on a donc une vitesse de convergence en  $n^{-2}$  de  $(a_n)$  vers  $\alpha$ . Cette bonne vitesse de convergence provient de la variance de la série arithmétique  $(x_k)_{k=1}^n = (k)_{k=1}^n$ . En effet, on peut montrer que  $s_x^2$  est de l'ordre de  $n^2$  et que  $a_n - \alpha$  est inversement proportionnel à  $s_x^2$ .

Pour la figure 1, nous avons  $C_n = 1$  pour tout  $n$ . Donnons d'autres exemples pour lesquels les sommes des perturbations sont bornées :

**Exemples 1 [Perturbations  $(e_k)$  telles que  $(C_n)$  soit bornée] :**

$e_k = (-1)^k, \sin(k\theta), \cos(k\eta)$  pour  $\eta \in ]0, 2\pi[$ , une suite périodique de moyenne nulle, les combinaisons linéaires de telles perturbations.

On va tester numériquement un exemple de perturbation périodique de moyenne nulle. On considère une perturbation mensuelle sur 3 ans, (donc sur 36 mois), d'une quantité devant vérifier  $y = 1 \times x$ , avec  $e_k = 3 \sin\left(\frac{\pi}{6}k\right)$  qui représente une perturbation mensuelle de période 12. Ainsi, on obtient la figure 2. La MMC nous fournit la droite  $y = 0,95x + 0,96$ .

Poursuivons l'étude de cet exemple pour un nombre d'années  $N$  compris entre 1 et 20 ans. Le nombre de mois est donc :  $n = 12 \times N$ . Nous obtenons les résultats suivants :

| N     | 1    | 2    | 3    | 4    | 5    | 6    | 7     | 8     | 9     | 10    | 20    |
|-------|------|------|------|------|------|------|-------|-------|-------|-------|-------|
| $a_n$ | 0,53 | 0,88 | 0,95 | 0,97 | 0,98 | 0,99 | 0,990 | 0,993 | 0,994 | 0,995 | 0,999 |
| $b_n$ | 3,05 | 1,46 | 0,96 | 0,71 | 0,57 | 0,47 | 0,40  | 0,35  | 0,31  | 0,282 | 0,141 |

Ces résultats numériques montrent bien que  $(a_n)$  converge bien plus vite vers 1 que  $(b_n)$  vers 0.  $b_n$  est ainsi bien plus sensible aux perturbations que  $a_n$ . Pour renforcer cette différence de stabilité entre la pente estimée par la MMC et l'ordonnée à l'origine estimée par la MMC, ajoutons la remarque suivante :

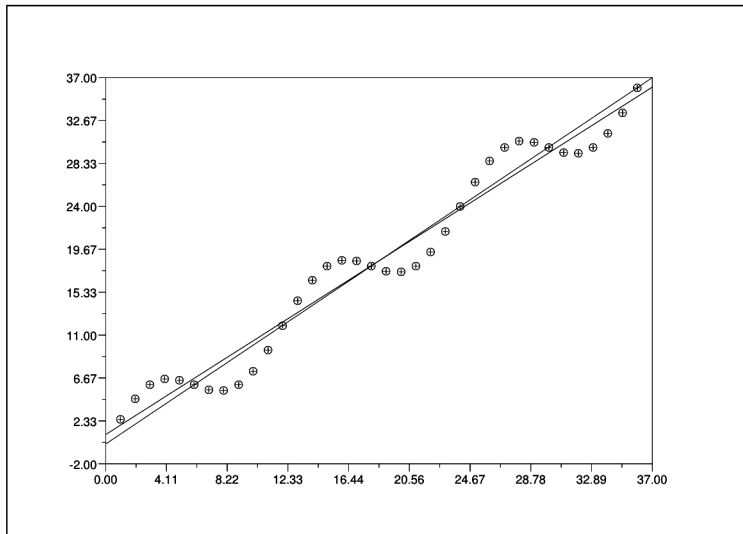


FIG. 2. Perturbation déterministe de taille 3 de  $y = x : y_k = 1 \times k + 3 \sin\left(\frac{\pi}{6}k\right)$

**Remarque 1 [Effet d'une perturbation de moyenne non nulle].**

$(b_n)$  répercutera directement une perturbation de moyenne non nulle. En revanche,  $(a_n)$  sera insensible à une telle perturbation. En effet, si  $(C_n)$  est bornée, alors cela signifie que la moyenne des perturbations est asymptotiquement nulle (et même de l'ordre de  $n^{-1}$ ). Si l'on remplace l'hypothèse  $(C_n)$  est bornée par il existe un nombre  $\delta$  et une constante  $C > 0$  tels que pour tout  $n$ ,

$$\left| \sum_{k=1}^n (e_k - \delta) \right| \leq C,$$

c'est à dire que  $(e_k)$  est une suite de perturbations du nombre  $\delta$ , alors  $(a_n)$  continue à converger vers  $\alpha$  en  $n^{-2}$ . En revanche,  $(b_n)$  ne converge plus vers  $\beta$  mais vers  $\beta + \delta$  en  $n^{-1}$ .

Pour avoir des exemples de perturbations avec  $C_n$  de l'ordre de  $n$ , il suffit de considérer les suites  $(ke_k)$  avec  $(e_k)$  provenant des exemples 1. Ainsi, si  $C_n$  est de l'ordre de  $n$ , on perd une bonne estimation de  $\beta$  mais on garde une estimation correcte de la pente. Mais, pour  $C_n$  plus grand que  $n^2$ , la perturbation est trop forte et la MMC ne nous permet plus de retrouver la pente  $\alpha$ . On laisse au lecteur le plaisir de traiter numériquement et graphiquement de telles perturbations. En revanche, grâce à des exemples de perturbations aléatoires, on va naturellement obtenir dans les sections 4 puis 5 des perturbations dont les sommes sont de l'ordre de  $\sqrt{n}$ , puis  $n\sqrt{n}$ .

#### 4. Étude probabiliste d'une variable affine perturbée

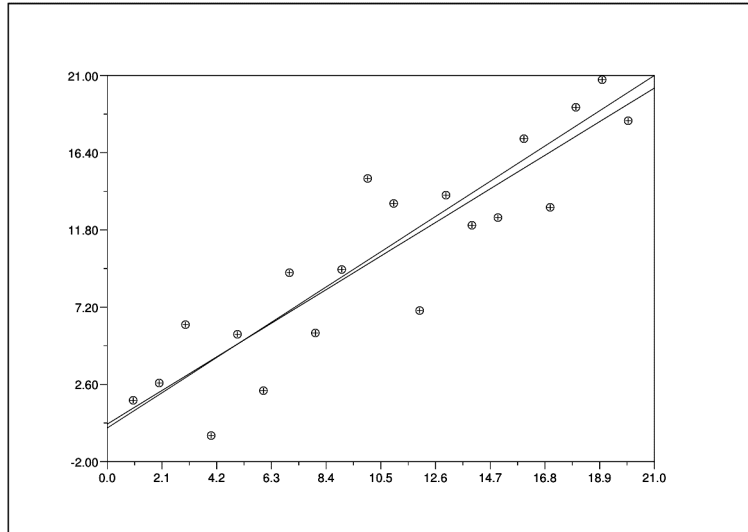


FIG. 3. Perturbation aléatoire d'ordre 5 de  $y = x : y_k = k + \mathcal{U}([-5,5])$

Pour la figure 3,  $\mathcal{U}([-5,5])$  désigne une perturbation suivant la loi uniforme sur l'intervalle  $[-5,5]$ . Pour l'obtenir, il suffit de calculer  $10 \times rand - 5$  où  $rand$  est la touche de la calculatrice qui simule la loi uniforme sur  $[0,1]$ . Le modèle exact est  $y = x$  et la MMC nous fournit, par exemple, pour  $n = 20$ , la droite d'équation  $y = 0,95x + 0,24$  qui est une bonne approximation de l'équation cherchée.

Attention, si l'on avait oublié de centrer la perturbation, en prenant  $y_k = k + 10 \times rand$ , on aurait obtenu par la MMC la droite  $y = 0,95x + 5,24$ . En effet,  $10 \times rand$  simule la loi uniforme de  $[0,10]$  de moyenne 5. Et, d'après la remarque 1, on aurait eu un décalage de +5 pour l'estimation de  $\beta$ .

Pour expliquer le résultat de la figure 3, il suffit de voir que la variable aléatoire  $S_n = e_1 + \dots + e_n$  simule une marche aléatoire. On s'attend donc à avoir  $S_n$  de l'ordre de  $C_n = constante \times \sqrt{n}$ . Ainsi, à l'aide du théorème 1, on pourra démontrer le résultat suivant :

**Théorème 2 [Perturbation aléatoire de  $y = \alpha x + \beta$ ]**

Soit  $(e_k)$  une suite de variables indépendantes, d'espérances nulles et de variances  $\sigma^2$ . Alors, pour un **niveau de confiance** de 99%, il existe une constante  $D > 0$  telle que la méthode des moindres carrés appliquée à la série chronologique  $(k, y_k)_{k=1}^n$  satisfaisant le modèle (2) fournit la droite d'équation  $y = a_n x + b_n$  avec les majorations suivantes pour tout  $n \geq 2$  :

$$P\left(|a_n - \alpha| \leq \frac{D}{n\sqrt{n}} \text{ et } |b_n - \beta| \leq \frac{D}{\sqrt{n}}\right) \geq 0,99.$$

En fait Gauss a montré dans ce cadre que la MMC est le **meilleur estimateur** linéaire par rapport à  $(y_1, \dots, y_n)$ , des coefficients de la droite (voir [1, 9, 13]). C'est-à-dire que parmi tous les estimateurs linéaires de  $\alpha$  et  $\beta$ ,  $a_n$  et  $b_n$  sont de variances minimales. On peut d'ailleurs montrer ici que  $\text{Var}(a_n) \sim 48\sigma^2 n^{-3}$  et  $\text{Var}(b_n) \sim 7\sigma^2 n^{-1}$ . Ceci permet de trouver précisément les intervalles de confiance précédents.

## 5. Pile ou face : Binomiale et méthode des moindres carrés

Soit  $(X_k)_{k \geq 1}$  une suite de variables aléatoires indépendantes de loi de Bernoulli  $\mathcal{B}(1/2)$ .  $X_k$  vaut 1, avec la probabilité 1/2, si l'on a obtenu un succès, et elle vaut 0 sinon.

Ainsi  $y_k = X_1 + \dots + X_k$  qui est le nombre de succès obtenus suit la loi binomiale

$\mathcal{B}(k, 1/2)$ . En centrant les variables  $X_k$ ,  $\tilde{X}_k := X_k - \frac{1}{2}$ , et en notant  $e_k = \tilde{X}_1 + \dots + \tilde{X}_k$ ,

qui est d'espérance nulle et d'écart type  $\frac{\sqrt{k}}{2}$ , on peut écrire la série des  $y_k$  sous la forme du modèle linéaire perturbé (2) :

$$y_k = X_1 + \dots + X_k = \frac{k}{2} + (\tilde{X}_1 + \dots + \tilde{X}_k) = \frac{k}{2} + e_k.$$

On remarque que, contrairement à l'exemple de la figure 3, la suite  $(y_k)$  est toujours croissante ; de plus, les perturbations sont de plus en plus fortes et non indépendantes.

On a obtenu la figure 4 pour une simulation de  $(k, y_k)_{k=1}^n$ . La MMC fournit comme approximation de la droite d'équation  $y = \frac{x}{2}$ , l'équation :

$y = (0,43\dots) \times x - (1,28\dots)$  avec  $n = 50$ . Ainsi on retrouve une approximation de la pente, mais l'on perd complètement l'ordonnée à l'origine.

Si l'on prend  $n \leq 30$  la MMC donne des résultats rarement précis, et même pour  $n = 50$ , il arrive que la pente de la droite obtenue par la MMC soit de l'ordre de 0,2 ou 0,7. Eu revanche, on peut démontrer le résultat suivant pour de très fortes perturbations de la relation  $y = px$ .

### Théorème 3 [MMC et binomiale]

Soient  $p \in ]0,1[$  et  $(X_k)$  une suite de variables de Bernoulli  $\mathcal{B}(p)$  indépendantes, et  $y_k = X_1 + \dots + X_k$ . Alors, pour un **niveau de confiance** de 99%, il existe une constante  $D > 0$  telle que la méthode des moindres carrés appliquée à la série chronologique  $(k, y_k)_{k=1}^n$  fournit la droite d'équation :  $y = a_n x + b_n$  avec les majorations suivantes pour tout  $n \geq 2$  :

$$P\left(|a_n - p| \leq \frac{D}{\sqrt{n}}\right) \geq 0,99.$$



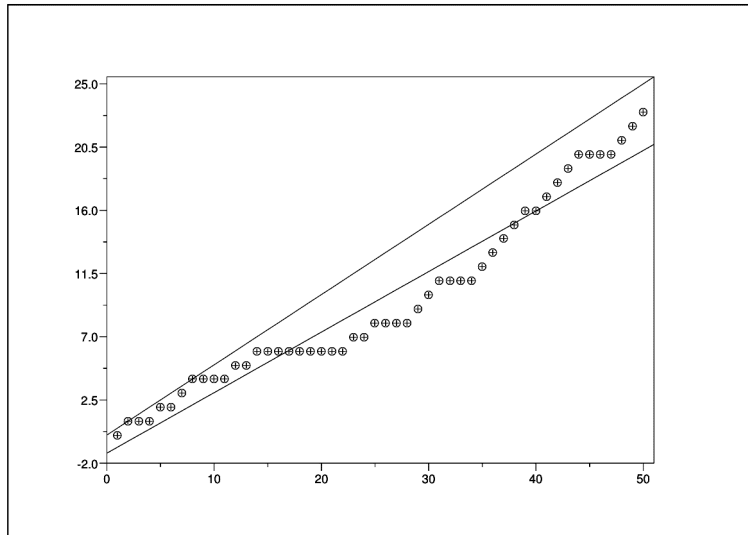


FIG. 4. Pile ou Face :  $y_k = \mathcal{B}(k, 1/2)$

Ainsi, la MMC nous permet d'estimer une probabilité  $p$  avec une erreur en  $\frac{1}{\sqrt{n}}$ . C'est le même ordre de convergence que l'on obtient par les fourchettes de sondages usuelles (voir [7]). Cependant  $\frac{X_1 + \dots + X_n}{n}$  reste le meilleur estimateur statistique de  $p$  (voir par exemple [8]).

## 6. Quelques remarques sur le modèle linéaire

Les théorèmes précédents prouvent que la méthode des moindres carrés est stable vis-à-vis de perturbations non négligeables. Cependant, on va voir que pour des perturbations dont les sommes sont bornées, le résidu moyen ne converge pas vers zéro. On verra aussi que le coefficient de corrélation converge vers  $\pm 1$  ou 0. On se demandera aussi pourquoi la pente  $\alpha$  est toujours obtenue avec une erreur environ  $n$  fois plus petite que l'ordonnée à l'origine  $\beta$ . Enfin on verra que tous les résultats précédents s'écrouleront lorsque  $(x_k)_{k \geq 1}$  sera une série bornée.

### 6.1. Résidu quadratique moyen et coefficient de corrélation

On n'avait pas parlé du coefficient de corrélation. Il n'est pas au programme du lycée. On sait que son interprétation est délicate. Il y a plusieurs tests statistiques associés à cet indicateur (cf [9]). Une idée circule souvent : « Si  $\rho \approx 1$ , alors les points sont presque alignés ». Si  $\rho^2 = 1$ , ils sont en effet parfaitement alignés. Mais, on peut montrer le résultat suivant :

**Théorème 4 [Comportement du résidu moyen et du coefficient de corrélation]**

Si, pour tout  $n$ ,  $\left| \sum_{k=1}^n e_k \right| \leq C$ , et  $\left( \frac{1}{n} \sum_{k=1}^n e_k^2 \right)$  converge vers un nombre  $s^2 > 0$ , alors il existe une constante  $D > 0$  telle que la méthode des moindres carrés appliquée à la série chronologique  $(k, y_k)_{k=1}^n$  du modèle (2) fournit un résidu moyen  $R_n$  et un coefficient de corrélation  $\rho_n$  tels que :

$$\left| R_n - \frac{1}{n} \sum_{k=1}^n e_k^2 \right| \leq \frac{D}{n^2}, \quad |1 - \rho_n^2| \leq \frac{D}{|\alpha| n^2} \text{ si } \alpha \neq 0 \text{ et } |1 - \rho_n^2| \leq \frac{D}{n} \text{ si } \alpha = 0.$$

Dans tous nos exemples de la section 3, les hypothèses de ce théorème sont vérifiées. De plus, comme  $\bar{e}$  convergeait vers 0, on a aussi  $s_c^2$  (qui dépend de  $n$ ) qui converge vers  $s^2$ . Ainsi, l'écart quadratique moyen converge :

$$\lim_{n \rightarrow +\infty} R_n = s^2.$$

En fait, ceci traduit très bien le genre de perturbation que l'on faisait subir à la droite d'équation (1).  $R_n$  permet d'estimer la variance de la perturbation.

Par contre, le coefficient de corrélation a deux types de comportements. Peut-on se fier à cet indicateur qui tend vers  $\pm 1$  ou 0 suivant les cas ?

Ainsi, si  $\alpha \neq 0$ ,  $(\rho_n)$  converge toujours vers 1, et même rapidement, alors que les points sont loin d'être alignés,  $\rho_n$  est égal à 0,98 pour l'exemple de la figure 2. On a ainsi une famille d'exemples qui contredisent bien la croyance « plus  $\rho$  est proche de  $\pm 1$ , plus les points sont presque alignés ». En revanche, si l'on pense en terme

d'erreur relative :  $\frac{y_k - (a_n k + b_n)}{y_k}$ , il n'y a pas de contradiction. En effet, si  $\alpha \neq 0$ ,

alors  $y_k$  est de l'ordre de  $\alpha k$  et  $\frac{y_k - (a_n k + b_n)}{y_k} = \frac{(\alpha - a_n)k + (\beta - b_n) + e_k}{y_k}$  converge

vers 0 quand  $n \rightarrow +\infty$ .

Pour  $\alpha = 0$ , le comportement du coefficient de corrélation est complètement différent :  $\rho_n \rightarrow 0$ . En fait, si les variables sont décorrélées ( $s_{xy} = 0$ ), alors le coefficient de corrélation qui est proportionnel à la covariance est nul. Comme on perturbe la relation  $y = 0$ , on a en fait  $y_k = e_k$ . C'est-à-dire que l'on ne mesure que la perturbation. Et la série  $(e_k)$  est de plus en plus décorrélée de  $(x_k = k)$ . En effet, dans tous les cas présentés, on a toujours  $\lim_{n \rightarrow +\infty} s_{x_e} = 0$ .

**6.2. L'approximation de la pente  $\alpha$  est meilleure que celle de  $\beta$** 

Pour comprendre ce phénomène simplement, voici deux analogies.

On se donne deux points A et B de coordonnées respectives  $(x_A, y_A)$ ,  $(x_B, y_B)$  avec  $x_A < x_B$  et on cherche l'équation de la droite passant par ces deux points :

$$y = \alpha x + \beta, \quad \alpha = \frac{y_B - y_A}{x_B - x_A}, \quad \beta = y_A - \alpha x_A.$$

Cependant on fait une erreur sur les ordonnées d'ordre  $\varepsilon$ . Imaginez par exemple la situation où les points de coordonnées  $(x_A, 0)$  et  $(x_B, 0)$  se trouvent à l'intersection des lignes d'un quadrillage alors que les valeurs de  $y_A$  et  $y_B$  sont approchées et les points A et B sont placés à l'aide d'une règle graduée. La pente  $\alpha$  dépend linéairement de  $y_A, y_B$  et inversement proportionnellement à la longueur de l'intervalle. En notant  $\Delta\alpha$  l'erreur que l'on commet sur  $\alpha$  et  $L := |x_B - x_A|$ , on a donc :

$$|\Delta\alpha| \leq \frac{|\Delta y_A| + |\Delta y_B|}{L} \leq \frac{2|\varepsilon|}{L}.$$

Ainsi, on retrouve un résultat bien connu pour le tracé de droite : plus l'intervalle est grand, plus la précision s'améliore.

En revanche  $\Delta\beta = \Delta y_A - (\Delta\alpha)x_A$ . Si, comme dans les exemples présentés, L est grand et  $x_A$  est de l'ordre de 1, alors on a une plus grande erreur sur  $\beta$  :

$\Delta\beta \approx \Delta y_A = \varepsilon$ . Plus précisément, supposons que les erreurs sur  $y_A$  et  $y_B$  sont indépendantes de moyennes nulles et d'écart type  $\varepsilon > 0$  alors  $\Delta\beta$  est de moyenne nulle et d'écart type  $\varepsilon\sqrt{1 + 2x_A L^{-1} + 2x_A^2 L^{-2}} \sim \varepsilon$  quand  $L \rightarrow +\infty$ .

On pourrait penser avoir une meilleure estimation de  $\beta$  si la série des abscisses était centrée entre  $x_A = -n$  et  $x_B = +n$ . On a alors  $\beta := \frac{y_A + y_B}{2}$  et

$$\Delta\beta = \frac{\Delta y_A + \Delta y_B}{2}.$$

Cette dernière égalité montre que  $|\Delta\beta| \leq \varepsilon$ . Donc en général l'erreur reste d'ordre  $\varepsilon$ . Si on suppose à nouveau que les erreurs sur  $y_A$  et  $y_B$  sont indépendantes de moyennes

nulles et d'écart type  $\varepsilon > 0$ , alors  $\Delta\beta$  est de moyenne nulle et d'écart type  $\frac{\varepsilon}{\sqrt{2}}$ . Alors

que  $\frac{\sqrt{2}\varepsilon}{L}$  est l'écart type de  $\Delta\alpha$ .

Voici une deuxième analogie pour mieux comprendre ce phénomène. Imaginez un grand champ rectangulaire dans le sens Nord-Sud. Au Nord se trouve l'entrée N d'une fourmilière et au sud une autre entrée S. Ces deux entrées ne sont pas visibles par un aigle volant au dessus du champ. En revanche, ce dernier, à l'œil perçant, voit une file de fourmis rouges qui serpente au milieu du champ. Cette file vient de l'entrée Nord et se dirige vers l'entrée Sud. Plus l'oiseau voit de fourmis, plus il sera capable de bien évaluer la direction précise empruntée par les fourmis. Cela lui donne une famille de parallèles. Pour localiser l'entrée Nord (i.e.  $\beta$ ) il doit choisir la bonne parallèle. La file zigzaguant, on se rend compte que l'on aura plus de mal à trouver l'entrée Sud ou Nord que la direction des fourmis.

Que se passe-t-il si l'on cherche  $M$  le milieu des deux entrées (i.e. la série est centrée) ? Supposons de plus que la file soit assez longue pour que l'on ait quasiment la valeur exacte de la direction (i.e.  $\alpha$  est connu). Comme toutes les lignes envisagées sont parallèles on comprend que l'on fait la même erreur sur  $M$ ,  $N$ ,  $S$ , puisque le calcul d'un de ces points revient à choisir une parallèle. D'ailleurs la recherche de  $M$  revient à la recherche du centre de gravité de la file de fourmis. En fait, comme l'on fait une erreur sur la pente et que l'on calcule les coordonnées de  $N$  et de  $S$  à partir de  $M$  et de la pente, on pourra avoir une précision légèrement moins bonne sur  $N$  et  $S$ . Ce qui correspond bien au résultat de notre première analogie : trouver l'équation de la droite passant par deux points.

Terminons par une étude numérique de l'effet du centrage de la série des abscisses. Pour la série  $(k, \sin(k))_{k=1}^n$ , on laisse le soin au lecteur de vérifier que le centrage améliore ou donne au moins la même estimation de  $\beta$ . L'exemple suivant montre que la situation peut être plus compliquée. On prend un intervalle des abscisses de longueur  $n - 1$  avec la série  $(k, \sin(\sqrt{k}))_{k=1}^n$  et la même série en centrant

les abscisses :  $\left(k - \frac{n+1}{2}, \sin(\sqrt{k})\right)_{k=1}^n$ . On a bien sûr exactement la même estimation

pour la pente  $\alpha$ . En revanche, on obtient respectivement des valeurs approchées de  $\beta$  dans le tableau suivant :

| $n =$                            | 40    | 45    | 46    | 47    | 50     | 59    | 60    |
|----------------------------------|-------|-------|-------|-------|--------|-------|-------|
| $x_k = k, b_n =$                 | 0,29  | 0,05  | 0,002 | -0,04 | -0,162 | -0,43 | -0,44 |
| $x_k = k - \frac{n+1}{2}, b_n =$ | -0,32 | -0,25 | -0,24 | -0,22 | -0,168 | -0,07 | -0,01 |

Ainsi le centrage de la série des abscisses donne parfois une meilleure approximation de  $\beta$  et parfois une plus mauvaise approximation. Une étude numérique plus poussée montre que ce phénomène continue pour de plus grandes valeurs de  $n$ .

### 6.3. Cas d'une série des abscisses bornée

Que se passe-t-il sur un intervalle de temps borné ? Un point très important dans cette note est que la suite  $(x_k)$  était non bornée, et que  $s_x^2$  est de l'ordre de  $n^2$ . Cela représente le cas d'une série chronologique où l'on mesure  $y_k$  à intervalles de temps réguliers  $\tau = 1$ . En général, avec  $\tau > 0$ , on aurait  $s_x^2 = \frac{n^2 - 1}{12} \tau^2$ . Ainsi, tous les résultats des sections précédentes auraient des majorations avec des termes en  $\tau^{-2}$ . Ce qui est d'autant meilleur lorsque  $\tau$  est grand, mais d'autant plus mauvais lorsque  $\tau$  est petit. Par exemple, supposons que l'on mesure  $y_k$  avec des intervalles de temps de plus en plus petits pour que la suite  $(x_k)$  soit bornée. Pour fixer les idées, considérons l'exemple suivant :  $\tau = n^{-1}$ ,  $x_k = \frac{k}{n}$ ,  $k = 1, \dots, n$ . Cette fois-ci,  $s_x^2$  est de

l'ordre de  $1/12$ , et la MMC n'est même plus capable de retrouver la pente quand  $n \rightarrow +\infty$ .

Exemple numérique : Considérons l'exemple de la figure 2 :  $y = x$  et l'équation du modèle perturbé s'écrit :  $\tilde{y} = x + 3\sin\left(\frac{x\pi}{6}\right)$ . Pour  $x_k = k/n$  et  $n = 36$ , on obtient par la MMC l'équation :  $y \approx 2,5x + 0,02$ . On pourrait objecter qu'il faut étudier le phénomène sur une année et non un mois. Mais cela n'arrange rien : pour  $x_k = 12 \times k/n$  et  $n = 36$ , on obtient par la MMC l'équation :  $y \approx 0,5x + 2,9$ .

## Références

- [1] Ake Björck, *Numerical Methods for Least Squares Problems*. SIAM, 1996.
- [2] Jean-Luc Chabert, Évelyne Barbin, Michel Guillemot, Anne Michel-Pajus, Jacques Borowczyk, Ahmed Djebbar, Jean-Claude Martzloff, *Histoires d'algorithmes. Du caillou à la puce*, Chapitre 9, Résolution de systèmes d'équations linéaires, p. 320, p. 324-331, p. 333-337. Belin, 1994.
- [3] William Feller, *An Introduction to Probability Theory and Its Applications*, Vol I & II, 1970.
- [4] Carl Friedrich Gauss, *Theory of the motion of the heavenly bodies moving about the Sun in conic sections*, 1809.
- [5] Michel Henry, *L'enseignement des probabilités*. IREM de Besançon, 1994.
- [6] Michel Henry, *Autour de la modélisation en probabilités*. Commission inter-IREM Statistique et Probabilités, Coordination : Michel Henry, Presses Universitaires Franc-Comtoises, 2001.
- [7] Michel Henry, *Simulation d'un sondage*. Bulletin de l'APMEP n° 444, p. 88-96, 2003.
- [8] Jean Jacod, *Cours de Probabilités* :  
<http://www.proba.jussieu.fr/pageperso/jacod.html>, 2001.
- [9] Kendall & Stuart, *The advanced Theory of Statistics*. Ch. 19 du Vol 2 & Vol 3, 1961 & 1966.
- [10] Patrick Lascaux, *Analyse numérique matricielle appliquée à l'art de l'ingénieur, tome 1, méthodes directes*. Dunod, 2004.
- [11] Adrien-Marie Legendre, *Nouvelles méthodes pour la détermination des orbites des comètes*. Courcier, Paris, 1806.
- [12] Emmanuel Lesigne, *Pile ou face. Une introduction aux théorèmes limites du calcul des probabilités*. Ellipses, 2001.
- [13] Guy Morlat, article *Statistique* de l'encyclopédie Universalis, 2003.
- [14] Petrov, *Sums of Independent Random Variables*. Springer-Verlag, 1972.
- [15] Encyclopédie en ligne Wikipédia, article en français : « Méthode des moindres carrés », articles en anglais : « Least squares », « Linear regression ».