

REGRESSIONS et EQUATIONS INTEGRALES

Jean Jacquelin

Exemples d'applications à diverses fonctions :

Gaussienne (pp.6-8)

Puissance, Exponentielle, Logarithme, Weibull (pp.15-20)

Sinusoïdale (pp.21-36)

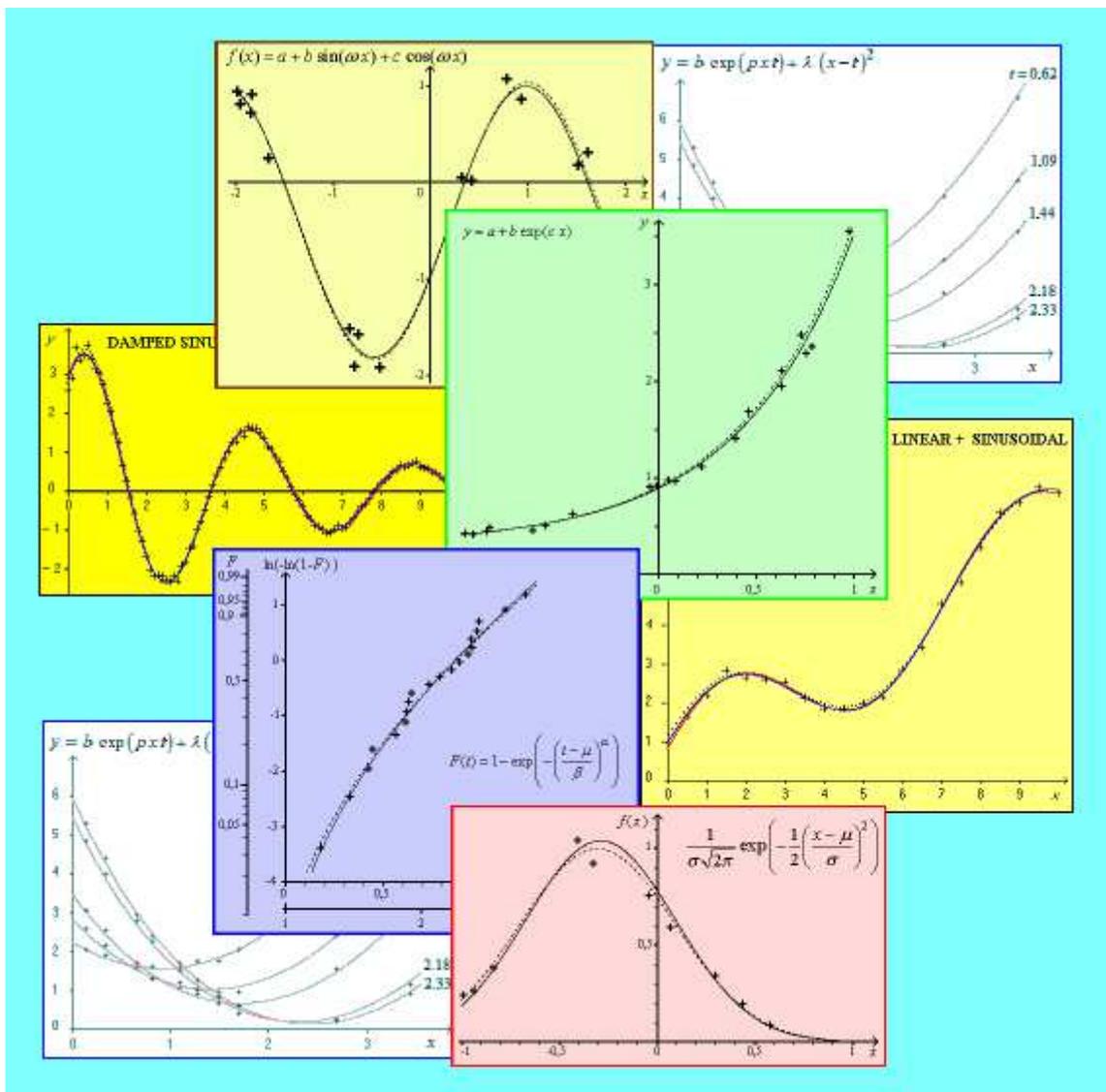
Logistic (pp.37-46)

Generalization of the sinusoidal regression (pp. 47-63)

Damped sinusoidal regression (pp. 64-70)

Double exponential or double power (pp. 71-74)

Régressions multi-variables (pp.75-84)



REGRESSIONS et EQUATIONS INTEGRALES

Jean Jacquelin

Résumé :

Le but principal de la présente publication est d'attirer l'attention sur une approche peu usitée pour la résolution de certains problèmes de régression non-linéaire.

La méthode est basée sur un principe de linéarisation par équation différentielle et/ou intégrale, dont l'exposé constitue la partie essentielle de ce papier.

Le processus de calcul se distingue fondamentalement des méthodes habituellement utilisées par le fait que la procédure n'est pas récursive et ne comporte donc pas de boucle d'itération.

Afin de donner un aperçu plus concret, des exemples de régression non-linéaires sont traités avec des applications numériques détaillées : régressions sur fonctions puissance, exponentielle, logarithme, ainsi que des fonctions usuelles en statistiques : densité de probabilité de Gauss, distribution de Weibull.

REGRESSIONS and INTEGRAL EQUATIONS

Abstract :

The main aim of this paper is to draw attention to a method rarely used to solve some regression problems.

In many cases, a differential and/or integral equation allows to turn a difficult problem of non-linear regression into a simple linear regression, which is the key part of the presentation.

The computation process is fundamentally different from the usual ones, since it isn't recursive. So, it doesn't require an iterative loop.

In order to give a more concrete view, some examples of non linear regressions are treated with detailed numerical examples : functions power, exponential, logarithm and some functions currently used in statistics : Gaussian Function, Weibull distribution.

REGRESSION et EQUATION INTEGRALE

Jean Jacquelin

La première version du papier *Régression et équations intégrale* date du 14/01/2009.
La présente version a été mise à jour le 27/04/09.

1. Introduction

La présente étude se situe dans le cadre général des problèmes de régressions. Par exemple, on connaît les coordonnées d'une série de n points : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$ et l'on cherche à ce que la courbe représentative d'une fonction $y = F(a, b, c, \dots; x)$ passe au plus près de ces points, ceci en optimisant les valeurs des paramètres a, b, c, \dots .

Le cas bien connu de la régression linéaire ne mérite qu'un bref rappel, ce qui sera fait en Appendix 1. Pour certains cas apparemment non linéaires et bien que cela puisse échapper à première vue, il est possible de revenir à une régression linéaire. Le cas de la fonction de répartition gaussienne en est un exemple : il sera traité en Appendix 2.

Hors les cas simples précédents, on est confronté à un véritable problème de régression non linéaire. La littérature sur le sujet est très étendue. Une revue, même sommaire, nous éloignerait de l'objectif du présent papier. Nous n'en aurons pas besoin ici car notre parti est de ramener certains problèmes non linéaires à une régression linéaire sans processus itératif ou récursif (si non, où serait l'originalité par rapport à des méthodes couramment utilisées ?).

Dès le paragraphe suivant, on entre dans le vif du sujet : c'est-à-dire les possibilités de ramener un problème non linéaire à une forme linéaire grâce à une équation différentielle et/ou intégrale convenable. La discussion préliminaire montre que, sauf cas particuliers, une équation intégrale est mieux adaptée à la résolution par calcul numérique qu'une équation différentielle, dans le contexte de ce genre de problèmes.

Le principe de l'utilisation d'une équation intégrale sera exposé et mis en pratique en prenant pour exemple la fonction de distribution gaussienne. D'autres exemples d'applications de la méthode de régression associée à une équation intégrale sont exposés de façon détaillée dans deux papiers joints :

- *Régressions non linéaires des genres : puissance, exponentielle, logarithme, Weibull.*
- *Régression sinusoidale .*

2. Principe de linéarisation par équation différentielle et/ou intégrale :

Commençons par un sommaire concernant les approximations des dérivées et/ou des primitives par calcul numérique. Etant donnés n points (x_k, y_k) situés à proximité de la courbe représentative d'une fonction $y(x)$ et étant donnée une autre fonction $g(x)$, on peut calculer les approximations pour les dérivées et/ou intégrales suivantes, avec $g_k = g(x_k)$:

$$D_k = \frac{g_{k+1}y_{k+1} - g_{k-1}y_{k-1}}{x_{k+1} - x_{k-1}} \simeq \left(\frac{d}{dx} g(x)y(x) \right)_{(x=x_k)}$$

$$DD_k = \frac{D_{k+1} - D_{k-1}}{x_{k+1} - x_{k-1}} \simeq \left(\frac{d^2}{dx^2} g(x)y(x) \right)_{(x=x_k)}$$

Et ainsi de suite, pour les dérivées suivantes, si nécessaire.

$$S_k \simeq \int_{x_1}^x g(u)y(u)du \quad \begin{cases} S_1 = 0 & \text{et pour } k = 2 \rightarrow n : \\ S_k = S_{k-1} + \frac{1}{2}(g_k y_k + g_{k-1} y_{k-1})(x_k - x_{k-1}) \end{cases}$$

$$SS_k \simeq \int_{x_1}^x \left(\int_{x_1}^v g(u)y(u)du \right) dv \quad \begin{cases} SS_1 = 0 & \text{et pour } k = 2 \rightarrow n : \\ SS_k = SS_{k-1} + \frac{1}{2}(S_k + S_{k-1})(x_k - x_{k-1}) \end{cases}$$

Et ainsi de suite, pour les intégrales suivantes, si nécessaire.

Il va sans dire que les points doivent être préalablement ordonnés selon les x_k croissants.

Il serait possible d'utiliser des méthodes de dérivation et/ou d'intégration numérique plus sophistiquées. Rien n'empêche non plus de prendre la (ou les) borne(s) inférieure(s) d'intégration autres que x_1 et même différentes entre elles pour les intégrations successives. Mais cela compliquerait les formules et alourdirait les explications. Pour faire simple, restons en aux formules les plus élémentaires possibles, du moins à ce stade de l'exposé.

Revenons maintenant à la formulation initiale du problème: optimiser les paramètres a, b, c, \dots d'une fonction $y(a, b, c, \dots; x)$ de telle sorte que sa courbe représentative passe au plus près de n points donnés (x_k, y_k) . Bien évidemment, les expressions littérales des dérivées et des primitives de cette fonction dépendent de a, b, c, \dots . Mais, leurs valeurs approchées calculées selon les formules précédentes, c'est-à-dire les valeurs numériques $D_k, DD_k, \dots, S_k, SS_k, \dots$ sont obtenues uniquement à partir des données (x_k, y_k) et **sans avoir besoin de connaître** a, b, c, \dots : cette observation est fondamentale dans la compréhension de la méthode qui va être exposée.

Supposons que la fonction $y(a, b, c, \dots; x)$ soit solution d'une équation différentielle et/ou intégrale linéaire telle que :

$$y(x) = A \Phi(x) + B \int G(x)y(x)dx + C \int \int H(x)y(x)dx^2 + \dots + \alpha \frac{d}{dx} g(x)y(x) + \beta \frac{d^2}{dx^2} h(x)y(x) + \dots$$

avec $\Phi(x), G(x), H(x), \dots, g(x), h(x), \dots$ des fonctions données ne dépendant pas de a, b, c, \dots et les coefficients $A, B, C, \dots, \alpha, \beta, \dots$ dépendant de a, b, c, \dots

Les valeurs approximatives sont donc respectivement :

$$\Phi_k = \Phi(x_k); G_k = G(x_k); H_k = H(x_k); \dots; \alpha_k = \alpha(x_k); \beta_k = \beta(x_k); \dots$$

$$D_k = \frac{g_{k+1}y_{k+1} - g_{k-1}y_{k-1}}{x_{k+1} - x_{k-1}}$$

$$DD_k = \frac{\Delta_{k+1} - \Delta_{k-1}}{x_{k+1} - x_{k-1}} \quad \text{avec} \quad \Delta_k = \frac{h_{k+1}y_{k+1} - h_{k-1}y_{k-1}}{x_{k+1} - x_{k-1}}$$

$$S_1 = 0 \quad ; \quad S_k = S_{k-1} + \frac{1}{2}(G_k y_k + G_{k-1} y_{k-1})(x_k - x_{k-1})$$

$$\left\{ \begin{array}{l} SS_1 = 0 \quad ; \quad SS_k = SS_{k-1} + \frac{1}{2}(\Xi_k + \Xi_{k-1})(x_k - x_{k-1}) \\ \text{avec : } \Xi_k = 0 \quad ; \quad \Xi_k = \Xi_{k-1} + \frac{1}{2}(H_k y_k + H_{k-1} y_{k-1})(x_k - x_{k-1}) \end{array} \right.$$

Si l'on remplace les dérivées et/ou primitives littérales par leurs approximations, l'équation cesse d'être exactement vérifiée. On considère alors la somme des écarts quadratiques :

$$\sum_{k=1}^n \varepsilon_k^2 = \sum_{k=1}^n (-y_k + A \Phi_k + B S_k + C SS_k + \dots + \alpha D_k + \beta DD_k + \dots)^2$$

La relation est linéaire relativement à $A, B, C, \dots, \alpha, \beta, \dots$. Ainsi, on se trouve ramené à une régression linéaire classique qui permet de calculer les valeurs optimums de $A_o, B_o, C_o, \dots, \alpha_o, \beta_o, \dots$. Finalement, puisque $A, B, C, \dots, \alpha, \beta, \dots$ sont des fonctions connues de a, b, c, \dots on aura à résoudre le système d'équations : $A(a,b,c,\dots)=A_o; B(a,b,c,\dots)=B_o; \dots; \alpha(a,b,c,\dots)=\alpha_o; \beta(a,b,c,\dots)=\beta_o; \dots$ pour obtenir les valeurs optimum des paramètres a, b, c, \dots

Des conditions complémentaires sont à prendre en considération, concernant le choix de l'équation différentielle et/ou intégrale. Outre qu'elle doit être linéaire relativement aux coefficients (mais non au sens des fonctions elles-mêmes, puisqu'on dispose du choix des $G(x), H(x), \dots, g(x), h(x), \dots$), l'équation doit, de préférence comporter autant de coefficients $A_o, B_o, \dots, \alpha_o, \beta_o, \dots$ qu'il y a de paramètres initiaux a, b, c, \dots à optimiser. S'il y en a moins, une (ou des) régression supplémentaire sera nécessaire pour calculer les coefficients ne figurant pas explicitement dans l'équation.

De plus, pour ne pas surcharger l'exposé, on a considéré une forme réduite d'équation différentielle et/ou intégrale. En fait, elle pouvait aussi comporter plusieurs fonctions $\Phi(x)$ différentes, plusieurs dérivées différentes (correspondant à des $g(x)$ différents), plusieurs intégrales différentes (correspondant à des $G(x)$ différents) et ainsi de suite pour les dérivées multiples et intégrales multiples.

On voit donc que l'on dispose de possibilités très nombreuses pour adapter une équation différentielle et/ou intégrale au problème à traiter. Toutefois, des contingences pratiques limitent ce choix. L'une des principales pierres d'achoppement résulte des difficultés inhérentes aux dérivations numériques. En effet, dans les cas où les points donnés ne sont pas régulièrement répartis, s'ils sont peu nombreux et insuffisamment proches les uns des autres et si, pour aggraver encore la situation, les valeurs des y_k ne sont pas assez précises, les dérivées calculées deviennent très fluctuantes, très dispersées, rendant inefficace la régression linéaire qui s'en suit. Au contraire, même dans ces cas difficiles, les intégrations numériques conservent une bonne stabilité (ce qui ne veut pas dire que les inévitables déviations sont faibles, mais au moins elles restent amorties, ce qui est essentiel pour la robustesse du procédé). Sauf cas particulier, il est donc largement préférable de s'orienter vers une équation intégrale plutôt qu'une équation comportant une fonction dérivée.

La généralité de la présentation qui vient d'être faite peut donner l'impression que la méthode est ardue et difficile à mettre en œuvre. Hors c'est tout le contraire lorsque l'on cesse de parler d'une façon abstraite, couvrant trop de cas différents et lorsque l'on s'applique à résoudre un cas concret.

L'un des exemples les plus spectaculaires est celui de la régression sinusoïdale (que nous évoquons seulement, sans approfondir ici, mais qui sera traitée en détail dans l'article joint : *Régression sinusoïdale*) :

Il s'agit d'optimiser les paramètres a , b , c et ω de l'équation :

$$y(x) = a + b \sin(\omega x) + c \cos(\omega x)$$

Cette fonction est solution de l'équation différentielle :

$$y(x) = A + B \frac{d^2 y}{dx^2} \quad \text{avec : } A = a \quad \text{et} \quad B = -\frac{1}{\omega^2}$$

C'est une équation linéaire relativement à A et B , qui sont eux-mêmes des fonctions (très simples) de a et ω . Qui plus est, les paramètres b et c n'interviennent plus directement. On est donc dans un cas typique et des plus aisés d'applicabilité de la méthode, sauf que s'agissant d'une dérivée seconde, il vaut mieux s'abstenir ! Heureusement, il n'y a pas de contre-indication à priori pour utiliser une équation intégrale dont la fonction sinusoïdale est solution. Ce n'est guère plus compliqué et donne en général des résultats largement satisfaisants (cette étude est exposée de façon détaillée dans le papier joint : *Régression sinusoïdale*).

Un premier exemple, plus simple, montre très clairement le processus de calcul : Dans le paragraphe suivant, la méthode de régression avec équation intégrale est appliquée à la fonction densité de probabilité de Gauss.

3. Exemple : Cas de la fonction densité de probabilité de Gauss :

Nous considérons la fonction de densité de probabilité, à deux paramètres σ et μ , définie par :

$$f(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right) \quad [1]$$

La notation générale $y(x)$ des paragraphes précédents se trouve donc être remplacée par $f(x)$ en raison de la spécificité de ce cas.

L'intégration [2] conduit à l'équation intégrale [3] dont $f(x)$ est solution :

$$\int_{x_1}^x (t-\mu) f(t) dt = -\sqrt{\frac{\pi}{2}} \sigma (f(x) - f(x_1)) \quad [2]$$

$$\begin{cases} f(x) - f(x_1) = A \int_{x_1}^x f(t) dt + B \int_{x_1}^x t f(t) dt \\ \text{avec : } A = \frac{\mu}{\sigma} \sqrt{\frac{2}{\pi}} \quad \text{et} \quad B = -\frac{1}{\sigma} \sqrt{\frac{2}{\pi}} \end{cases} \quad [3]$$

C'est une équation intégrale linéaire avec la particularité de comporter deux intégrales simples, ce qui entre dans les extensions mentionnées à la fin du paragraphe précédent. On calcule les approximations respectives, la première étant notée S avec $G(x) = 1$ et la seconde notée T avec $G(x) = x$:

$$\begin{cases} S_1 = 0 \\ S_k = S_{k-1} + \frac{1}{2}(f_k + f_{k-1})(x_k - x_{k-1}) \quad k = 2 \rightarrow n \end{cases} \quad [4]$$

$$\begin{cases} T_1 = 0 \\ T_k = T_{k-1} + \frac{1}{2}(x_k f_k + x_{k-1} f_{k-1})(x_k - x_{k-1}) \quad k = 2 \rightarrow n \end{cases} \quad [5]$$

En remplaçant $f(x_k)$ par f_k , ainsi que $f(x_1)$ par f_1 et les intégrales par S_k et T_k respectivement, l'équation [3] n'est plus exactement vérifiée. On cherche à minimiser la somme des carrés des écarts :

$$\sum_{k=1}^n \varepsilon_k^2 = \sum_{k=1}^n (-(f_k - f_1) + A S_k + B T_k)^2 \quad [6]$$

Remarquons que, si l'on avait choisi une autre borne inférieure d'intégration que x_1 , cela aurait entraîné le changement de f_1 , mais aussi des valeurs numériques différentes pour S_k et T_k , le tout se compensant et ne modifiant pas le résultat final.

La relation [6] n'est autre que l'équation de base d'une régression linéaire dont on sait calculer la solution optimum A_1, B_1 :

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} \sum (S_k)^2 & \sum S_k T_k \\ \sum S_k T_k & \sum (T_k)^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum (y_k - y_1) S_k \\ \sum (y_k - y_1) T_k \end{pmatrix} \quad [7]$$

Avec conventionnellement : $\sum \equiv \sum_{k=1}^n$. On déduit ensuite σ_1 et μ_1 d'après [3] :

$$\sigma_1 = -\frac{1}{B_1} \sqrt{\frac{2}{\pi}} \quad ; \quad \mu_1 = -\frac{A_1}{B_1} \quad [8]$$

En résumé, le processus de calcul numérique est le suivant :

Données : $(x_1, f_1), (x_2, f_2), \dots, (x_k, f_k), \dots, (x_n, f_n)$

- Calcul des S_k :

$$\begin{cases} S_1 = 0 \\ S_k = S_{k-1} + \frac{1}{2}(f_k + f_{k-1})(x_k - x_{k-1}) \quad k = 2 \rightarrow n \end{cases}$$

- Calcul des T_k :

$$\begin{cases} T_1 = 0 \\ T_k = T_{k-1} + \frac{1}{2}(x_k f_k + x_{k-1} f_{k-1})(x_k - x_{k-1}) \quad k = 2 \rightarrow n \end{cases}$$

- Calcul de : $\sum (S_k)^2, \sum S_k T_k, \sum (T_k)^2,$
 $\sum (y_k - y_1) S_k, \sum (y_k - y_1) T_k$

- Calcul de A_1 et B_1 :

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} \sum (S_k)^2 & \sum S_k T_k \\ \sum S_k T_k & \sum (T_k)^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum (y_k - y_1) S_k \\ \sum (y_k - y_1) T_k \end{pmatrix}$$

- Calcul de σ_1 et μ_1 : $\sigma_1 = -\frac{1}{B_1} \sqrt{\frac{2}{\pi}}$; $\mu_1 = -\frac{A_1}{B_1}$

Résultat : σ_1 et μ_1 sont les valeurs approchées de σ et μ

Pour illustrer ce calcul (figure 1), les données numériques (Table 1) ont été générées de la manière suivante : Les x_k ont été tirés au hasard sur la plage des abscisses considérées. A partir de valeurs "exactes" données σ_e et μ_e , (définissant la fonction $f(x)$ dite "exacte" dont la courbe représentative est tracée en pointillés sur la figure 1), on a calculé les $f(x_k)$ exacts correspondants avec l'équation [9]. Ils ont été ensuite affectés de déviations dont l'amplitude a été tirée au hasard entre - et + 10% de $f(x_k)$, ce qui a donné, après arrondis, les valeurs numériques f_k indiquées sur la Table 1.

Cette modélisation outrancière de l'imprécision sur les ordonnées répond au souci de lisibilité de la figure, de telle sorte que les points dits "expérimentaux", figurés par des croix, soient assez éloignés de la courbe en pointillés. Dans le même esprit, un nombre exagérément faible de points a été choisi de façon à ce que les défauts soient mis en évidence sur la figure 1 par une différence nette entre les courbes "exactes" en pointillés et celles en trait plein représentatives des résultats de calculs intermédiaires et final. Le fait que les points ne soient pas répartis à intervalles constants selon les abscisses est aussi un facteur fortement aggravant la difficulté.

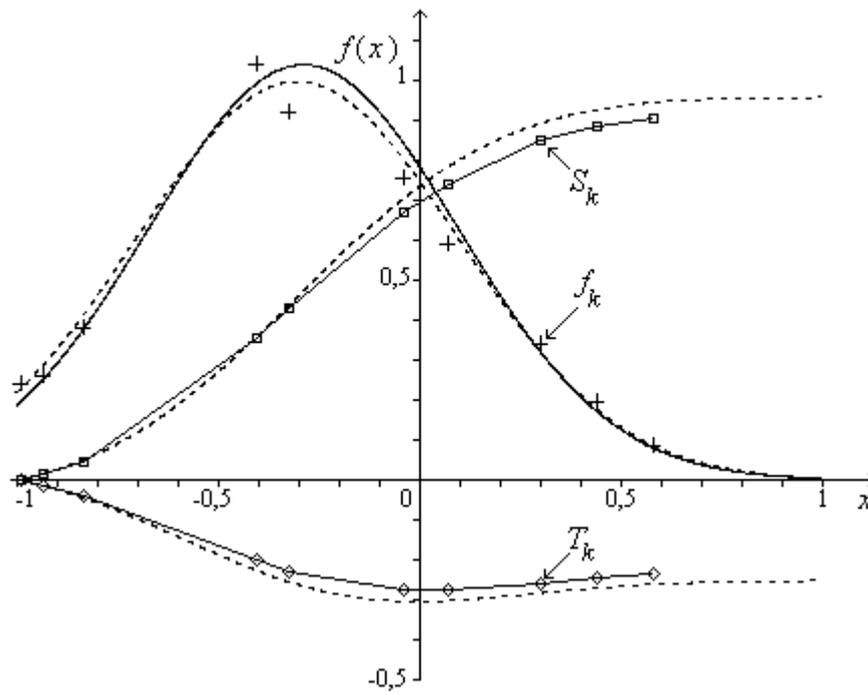


Figure 1 : Fonction densité de probabilité de Gauss, exemple de régression.

| k | x_k | f_k | S_k | T_k | |
|-----|--------|-------|----------|------------|-----------------------|
| 1 | -0,992 | 0,238 | 0 | 0 | |
| 2 | -0,935 | 0,262 | 0,01425 | -0,0137104 | $\sigma_e = 0,4$ |
| 3 | -0,836 | 0,38 | 0,046029 | -0,0415616 | $\mu_e = -0,3$ |
| 4 | -0,404 | 1,041 | 0,352965 | -0,201022 | |
| 5 | -0,326 | 0,922 | 0,429522 | -0,229147 | $\sigma_1 = 0,383915$ |
| 6 | -0,042 | 0,755 | 0,667656 | -0,276331 | $\mu_1 = -0,289356$ |
| 7 | 0,068 | 0,589 | 0,741576 | -0,275872 | |
| 8 | 0,302 | 0,34 | 0,850269 | -0,259172 | |
| 9 | 0,439 | 0,193 | 0,88678 | -0,246335 | |
| 10 | 0,58 | 0,083 | 0,906238 | -0,236968 | |

Table 1 : Valeurs numériques correspondantes à l'exemple de la figure 1.

Sur la figure 1, le tracé des courbes représentatives des intégrales "exactes" et des points (x_k, S_k) et (x_k, T_k) fait clairement apparaître la cause principale de déviations dans cette méthode de calcul : L'intégration numérique, bien que plus favorable que ne serait la dérivation, n'est pas parfaite, ce qui entraîne des déviations sur le résultat (σ_1, μ_1) .

Pour se faire une opinion objective des qualités et défauts de la méthode qui vient d'être exposée, il faudrait mener une étude expérimentale systématique sur un très grand nombre de cas et d'exemples. Ceci reste à faire, dans l'état d'avancement actuel de l'étude.

Il est à priori certain que les déviations, causées par le défaut inhérent aux intégrations numériques, seront considérablement réduites si les points sont assez nombreux et leurs abscisses réparties à intervalles assez réguliers.

4. Commentaires :

Il serait déraisonnable d'imaginer que la méthode présentée ici peut remplacer celles qui sont couramment utilisées, implantées dans les logiciels commerciaux et qui bénéficient d'une longue histoire d'études, d'expérimentations et de fiabilisation. On peut même prévoir avec quasi certitude que les méthodes de régression non linéaires qui ont fait leurs preuves, en travaillant par approximations successives, convergent vers un résultat plus précis qu'une méthode directe, sans calcul itératif. Alors on se demande bien quel peut être la motivation du présent travail.

Certes, en général les méthodes récursives nécessitent de connaître au départ une première approximation, au moins un ordre de grandeur, du résultat que l'on cherche. Ce n'est pas un handicap en général car le praticien ne part pas dans l'inconnu total. On pourrait penser à la méthode de régression avec équation intégrale pour, éventuellement, satisfaire ce besoin de première approximation. Mais c'est un besoin bien marginal, donc il ne faut pas voir là une motivation sérieuse.

Certes, une méthode de principe simple, aisée à programmer, telle que celle présentée ici, pourrait séduire quelques utilisateurs potentiels dans des situations particulières où l'on cherche à avoir la maîtrise totale des calculs que l'on exécute : L'utilisateur de logiciels commerciaux est bien satisfait des résultats qu'ils fournissent, mais peut parfois regretter de ne pas savoir ce que fait précisément le logiciel sophistiqué qu'il manipule. Néanmoins ce serait une piètre motivation pour la présente étude que de vouloir fournir un outil moins performant que ce qui existe, dans le seul but de répondre à un sentiment de frustration à l'usage d'outils dont on ne connaît pas exactement le mécanisme.

En fait, il faut voir dans ce papier, non pas une motivation utilitaire dans le cas spécifique de la distribution de Gauss, mais au contraire l'intention d'attirer l'attention sur une idée plus générale : les nombreuses possibilités offertes par les équations intégrales pour transformer un problème de régression non linéaire en une régression linéaire et en déduire un processus de calcul de principe non itératif.

Il est hors de question de vouloir concurrencer ce qui a déjà l'avantage d'exister et qui mieux est, de bien fonctionner. Par contre, pour aider à résoudre de futurs problèmes, parmi les voies possibles il serait dommage d'en oublier une : celle qui fait l'objet du présent papier et dont le paragraphe 2 constitue l'essentiel de la présentation.

Appendix 1 : Régression linéaire (rappel)

Lorsque la fonction $y = F(a, b, c, \dots; x)$ que l'on cherche à optimiser peut se mettre sous la forme : $y = a f(x) + b g(x) + c h(x) + \dots$, selon le nombre de paramètres a, b, c, \dots et avec des fonctions $f(x), g(x), h(x), \dots$ connues, le processus est linéaire relativement aux paramètres à optimiser.

Encore plus généralement, si la fonction $y = F(a, b, c, \dots; x)$ peut être transformée et mise sous la forme : $F(x,y) = A f(x,y) + B g(x,y) + C h(x,y) + \dots$ avec des fonctions connues : $F(x,y), f(x,y), g(x,y), h(x,y), \dots$, $A(a,b,c,\dots), B(a,b,c,\dots), C(a,b,c,\dots), \dots$ le processus est encore linéaire relativement aux coefficients A, B et C , bien qu'il ne le soit plus relativement à a, b, c, \dots . Mais il relève toujours d'une régression linéaire. En effet, la méthode "des moindres carrés" consiste à chercher le minimum de :

$$\left\{ \begin{array}{l} \mathcal{E}^2_{(A,B,C,\dots)} = \sum_{k=1}^n (F_k - (A f_k + B g_k + C h_k + \dots))^2 \\ F_k \equiv F(x_k, y_k); f_k \equiv f(x_k, y_k); g_k \equiv g(x_k, y_k); h_k \equiv h(x_k, y_k) \end{array} \right.$$

Les dérivées partielles relatives à A, B, C, \dots conduisent au système d'équations dont les solutions A_0, B_0, C_0, \dots sont optimum :

$$\left\{ \begin{array}{l} \left(\frac{\partial(\mathcal{E}^2)}{\partial A} \right)_{A_0, B_0, C_0, \dots} = - \sum_{k=1}^n (F_k - (A_0 f_k + B_0 g_k + C_0 h_k + \dots)) f_k = 0 \\ \left(\frac{\partial(\mathcal{E}^2)}{\partial B} \right)_{A_0, B_0, C_0, \dots} = - \sum_{k=1}^n (F_k - (A_0 f_k + B_0 g_k + C_0 h_k + \dots)) g_k = 0 \\ \left(\frac{\partial(\mathcal{E}^2)}{\partial C} \right)_{A_0, B_0, C_0, \dots} = - \sum_{k=1}^n (F_k - (A_0 f_k + B_0 g_k + C_0 h_k + \dots)) h_k = 0 \\ \dots \end{array} \right.$$

La résolution de ce système linéaire, écrit conventionnellement avec $\sum \equiv \sum_{k=1}^n$ conduit à :

$$\begin{pmatrix} A_0 \\ B_0 \\ C_0 \\ \dots \end{pmatrix} = \begin{pmatrix} \sum f_k^2 & \sum f_k g_k & \sum f_k h_k & \dots \\ \sum f_k g_k & \sum g_k^2 & \sum g_k h_k & \dots \\ \sum f_k h_k & \sum g_k h_k & \sum h_k^2 & \dots \\ \dots & \dots & \dots & \dots \end{pmatrix}^{-1} \begin{pmatrix} \sum F_k f_k \\ \sum F_k g_k \\ \sum F_k h_k \\ \dots \end{pmatrix}$$

Ensuite, on obtient les valeurs optimum de a, b, c, \dots correspondantes par résolution du système suivant, dont les inconnues sont a_0, b_0, c_0, \dots :

$$\begin{cases} A(a_0, b_0, c_0, \dots) = A_0 \\ B(a_0, b_0, c_0, \dots) = B_0 \\ C(a_0, b_0, c_0, \dots) = C_0 \\ \dots \end{cases}$$

qui est un système d'équations non linéaires dans la mesure où les fonctions $A(a,b,c,\dots)$, $B(a,b,c,\dots)$, $C(a,b,c,\dots)$, ... ne sont pas linéaires. Mais cela n'empêche pas que la régression qui a été faite est linéaire, donc que ce cas a bien sa place dans le présent paragraphe.

Bien sûr, ceci peut être encore étendu en considérant plus de variables, par exemple x, y, z, t, \dots , au lieu de seulement x, y et donc de travailler en 3D., ou 4D., ... au lieu de 2D.. Tout ce qui précède figure dans la littérature de façon plus détaillée et surtout mieux structurée, avec des présentations adaptées à une théorie générale. Ici, le propos était seulement un bref rappel, avec les notations spécifiques cohérentes avec celles utilisées par la suite.

Appendix 2 : Régression linéaire, cas de la fonction de répartition gaussienne

Nous considérons la fonction de répartition gaussienne non centrée, à deux paramètres σ et μ , définie par :

$$F(x) = \frac{1}{\sqrt{2\pi} \sigma} \int_{-\infty}^x \exp\left(-\frac{1}{2}\left(\frac{t-\mu}{\sigma}\right)^2\right) dt \quad [11]$$

Un exemple est représenté sur la figure 2 (courbe en pointillés).

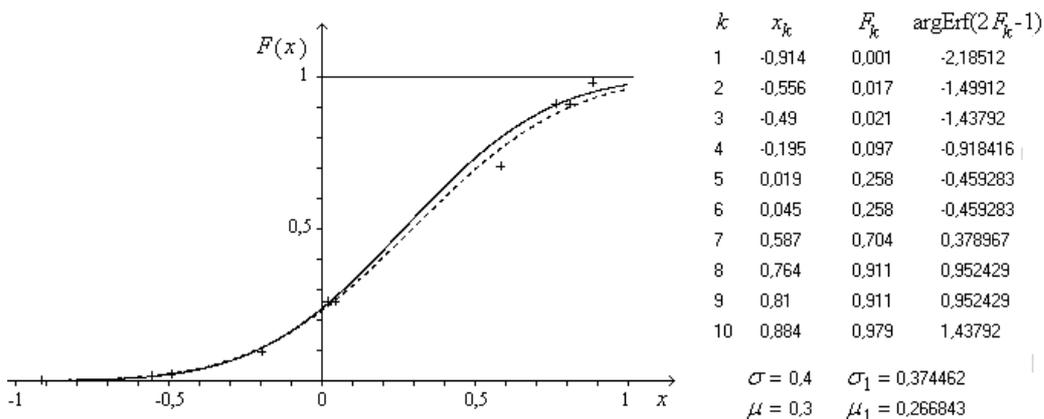


Figure 2 : Exemple de régression, cas d'une fonction de répartition gaussienne.

Les données sont les points dit "expérimentaux" : $(x_1, F_1), (x_2, F_2), \dots, (x_k, F_k), \dots, (x_n, F_n)$ qui, sur l'exemple de la figure 11, présentent une certaine dispersion par rapport à leur positions théoriques respectives $(x_k, F(x_k))$ sur la courbe en pointillés représentative de $F(x)$.

Une forme équivalente d'écriture de $F(x)$ se réfère à la fonction Erf , dite "fonction d'erreur" et définie par :

$$\operatorname{erf}(z) = \frac{2}{\sqrt{\pi}} \int_0^z \exp(-\tau^2) d\tau \quad [12]$$

Le changement de variable $t = \mu + \sqrt{2} \sigma \tau$ dans [11] donne la relation :

$$F(x) = \frac{1}{\sqrt{\pi}} \int_{-\infty}^{\frac{x-\mu}{\sqrt{2}\sigma}} \exp(-\tau^2) d\tau = \frac{1}{2} + \frac{1}{2} \operatorname{erf}\left(\frac{x-\mu}{\sqrt{2}\sigma}\right) \quad [13]$$

La fonction réciproque ou "inverse" de Erf est désignée par $\operatorname{Erf}^{(-1)}$, ou Erfinv , ou argErf . Nous utiliserons cette dernière notation.

Ainsi, la relation réciproque de [13] s'écrit :

$$\frac{x-\mu}{\sqrt{2}\sigma} = \operatorname{argErf}(2F(x)-1) \quad [14]$$

Ce qui conduit à la relation linéaire relativement à A et B définis par :

$$y(x) = \operatorname{argErf}(2F(x)-1) = A x + B \quad \begin{cases} A = \frac{1}{\sqrt{2}\sigma} \\ B = -\frac{\mu}{\sqrt{2}\sigma} \end{cases} \quad [15]$$

Il s'agit donc d'une régression linéaire sous sa forme la plus élémentaire, relativement aux points (x_k, y_k) avec les y_k calculés préalablement par :

$$y_k = \operatorname{argErf}(2F_k - 1) \quad [16]$$

Les valeurs optimum A_1, B_1 sont les solutions du système suivant :

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} \sum (x_k)^2 & \sum x_k \\ \sum x_k & n \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k x_k \\ \sum y_k \end{pmatrix} \quad [17]$$

avec conventionnellement : $\sum \equiv \sum_{k=1}^n$. On déduit ensuite σ_1 et μ_1 d'après [15] :

$$\sigma_1 = \frac{1}{\sqrt{2} A_1} \quad ; \quad \mu_1 = -\frac{B_1}{A_1} \quad [18]$$

Pour l'exemple traité, les valeurs numériques σ_1 et μ_1 obtenues sont indiquées sur la figure 1, où la courbe représentative de la fonction correspondante est tracée en trait plein. Elle est voisine de la courbe "théorique" en pointillé.

En fait, l'exemple a été choisi intentionnellement avec un très faible nombre de points et une forte dispersion pour que les deux courbes soient bien distinctes l'une de l'autre, ce qui est plutôt dépréciatif et peu représentatif de la qualité de ce qui est obtenu le plus souvent.

En résumé, le processus de calcul numérique, très simple, est le suivant :

Données : $(x_1, F_1), (x_2, F_2), \dots, (x_k, F_k), \dots, (x_n, F_n)$

- Calcul des y_k : $y_k = \text{argErf}(2F_k - 1)$

- Calcul de $\sum x_k, \sum (x_k)^2, \sum y_k, \sum y_k x_k$

- Calcul de A_1 et B_1 :

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} \sum (x_k)^2 & \sum x_k \\ \sum x_k & n \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k x_k \\ \sum y_k \end{pmatrix}$$

- Calcul de σ_1 et μ_1 : $\sigma_1 = \frac{1}{\sqrt{2} A_1}$; $\mu_1 = -\frac{B_1}{A_1}$

Résultat : σ_1 et μ_1 sont les valeurs approchées de σ et μ

Si l'on ne dispose pas de la fonction argErf implémentée dans le logiciel utilisé, un exemple de listing pour les fonctions Erf et argErf est donné page suivante.

Note : Ce qui a est présenté en Appendix 1 et 2 est bien connu. Toutefois, il était utile d'attirer l'attention sur la différence fondamentale entre les problèmes de régressions rappelés en Appendix 1 et ceux considérés au §.2 du texte principal. Il est également utile de donner un exemple de différence notable entre les régressions appliquées à la distribution de Gauss, d'une part dans le cas aisé de la **fonction de répartition** (Appendix 2) et d'autre part dans le cas difficile de la **fonction densité** (§.3 du texte principal).

Listing pour les fonctions Erf et argErf :

Les valeurs approchées de $\text{Erf}(x)$ sont obtenues avec au moins huit chiffres significatifs après la virgule. On utilise le développement limité suivant :

$$\text{Erf}(x) \approx \frac{2x}{\sqrt{\pi}} \sum_{k=0}^{30} \frac{(-1)^k x^{2k}}{k!(2k+1)} \quad \begin{cases} |x| < 2,7 \\ |\text{Erf}(x)| < 0,999866 \end{cases}$$

complété par le développement limité asymptotique :

$$\text{Erf}(x) \approx \pm 1 - \frac{e^{-x^2}}{x\sqrt{\pi}} \sum_{k=0}^5 \frac{(-1)^k (2k+1)!!}{x^{2k}} \quad \begin{cases} + \text{ si } x > 2,7 ; - \text{ si } x < -2,7 \\ (2k+1)!! = 1 * 3 * \dots * (2k+1) \\ 0,999865 < |\text{Erf}(x)| < 1 \end{cases}$$

La fonction $\text{argErf}(y)$ est calculée par la méthode de Newton-Raphson. Le résultat $\text{argErf}(y)$ est obtenu avec au moins huit chiffres significatifs après la virgule si : $|y| < 0,999\,999\,999\,998 \rightarrow |\text{argErf}(y)| < 5$. Au delà de ce domaine, le résultat n'est pas significatif.

Le listing (page suivante), écrit en langage Pascal, ne comporte que du vocabulaire et syntaxe élémentaires. Il ne devrait pas y avoir de difficulté pour le traduire dans tout autre langage souhaité.

```

Function Erf(x:extended):extended;
var
  y,p:extended;
  k:integer;
begin
  y:=0;
  p:=1;
  if ((x>-2.7) and (x<2.7)) then
  begin
    for k:=0 to 30 do
    begin
      y:=y+p/(2*k+1);
      p:=-p*x*x/(k+1);
    end;
    y:=y*2*x/sqrt(pi);
  end else
  begin
    for k:=0 to 5 do
    begin
      y:=y+p;
      p:=-p*(2*k+1)/(2*x*x);
    end;
    y:=y*exp(-x*x)/(x*sqrt(pi));
    if x>0 then y:=1-y else y:=-1-y;
  end;
  Erf:=y;
end;

```

```

Function argErf(y:extended):extended;
var
  x:extended;
  k:integer;
begin
  x:=0;
  for k:=1 to 30 do
  begin
    x:=x+exp(x*x)*sqrt(pi)*(y-Erf(x))/2;
  end;
  argErf:=x;
end;

```

On pourra effectuer des tests par comparaison des résultats du calcul avec les valeurs ci-contre (faire également les tests avec les mêmes valeurs mais négatives) :

| $x = \text{argErf}(y)$ | $\text{Erf}(x) = y$ |
|------------------------|---------------------|
| 0.001 | 0.001128378791 |
| 0.1 | 0.112462916 |
| 1. | 0.8427007929 |
| 2. | 0.995322265 |
| 2.699 | 0.9998648953 |
| 2.701 | 0.9998664351 |
| 4. | 0.9999999846 |
| 5. | 0.9999999999984 |

**REGRESSIONS NON LINEAIRES des genres :
PUISSANCE, EXPONENTIELLE, LOGARITHME, WEIBULL**

Jean Jacquelin

La première version de ce papier date du 18/01/2009.

La présente version a été mise à jour le 23/04/09.

Résumé :

Les paramètres de fonctions puissance, exponentielle, logarithme, Weibull sont optimisés par une méthode de régression non itérative faisant appel à une équation intégrale appropriée.

**NON LINEAR REGRESSIONS of the kinds :
POWER, EXPONENTIAL, LOGARITHM, WEIBULL**

Abstract :

The parameters of functions Power, Exponential, Logarithm, Weibull are optimized thanks to a non iterative process of regression using a convenient integral equation.

**REGRESSIONS NON LINEAIRES des genres :
 PUISSANCE, EXPONENTIELLE, LOGARITHME, WEIBULL**

Jean Jacquelin

1. Introduction

Les deux cas de régressions suivants seront traités simultanément :

$$y = a + b X^c \quad (X > 0) \quad (1)$$

$$y = a + b \exp(c x) \quad (2)$$

En effet, dans le cas (1), si les données sont les points $(X_1, y_1), \dots, (X_k, y_k), \dots, (X_n, y_n)$, on calculera préalablement les $x_k = \ln(X_k)$ (3)

ce qui ramène au cas (2) pour lequel les données sont : $(x_1, y_1), \dots, (x_k, y_k), \dots, (x_n, y_n)$

Diverses autres formes d'équations se ramènent au même cas :

- L'équation : $y = a + b' \exp(c(x - \mu))$ est identique à (2) en posant : $b = b' \exp(-c \mu)$

- L'équation : $y = \alpha + \beta \ln(x - \gamma)$ revient à (2) en intervertissant les notations x et y , ce qui répond donc aux cas de régressions de fonctions logarithme à trois paramètres.

- Etc. En particulier, le cas de l'équation de Weibull à trois paramètres sera traité au §.3.

La méthode utilisée a été décrite au §.2 de l'article : "Régression et équation intégrale".

2. Régression dans le cas de la fonction $y(x) = a + b \exp(c x)$

L'intégration de la fonction $y(x)$ donne :

$$\int_{x_1}^x y(u) du = a(x - x_1) + \frac{b}{c} \exp(c x) - \frac{b}{c} \exp(c x_1) \quad (4)$$

et en reportant $\exp(c x)$ tiré de (2) :

$$\int_{x_1}^x y(u) du = a(x - x_1) + \frac{1}{c} (y - a) - \frac{b}{c} \exp(c x_1) \quad (5)$$

D'où l'équation intégrale qui sera utilisée :

$$y - (a + b \exp(c x_1)) = -a c (x - x_1) + c \int_{x_1}^x y(u) du \quad (6)$$

Les approximations des valeurs numériques l'intégrale pour les $x = x_k$ sont calculées avec :

$$\begin{cases} S_1 = 0 & \text{et pour } k = 2 \rightarrow n : \\ S_k = S_{k-1} + \frac{1}{2} (y_k + y_{k-1}) (x_k - x_{k-1}) \end{cases} \quad (7)$$

En remplaçant dans (6) les valeurs exactes inconnues par leurs approximations respectives, l'équation n'est plus exactement vérifiée :

$$y_k - y_1 \cong -a c (x_k - x_1) + c S_k \quad (8)$$

On cherche à minimiser la somme des écarts quadratiques :

$$\sum_{k=1}^n \varepsilon_k^2 = \sum_{k=1}^n (A(x_k - x_1) + B S_k - (y_k - y_1))^2 \quad (9)$$

avec : $A = -a c$; $B = c$ (10)

Il s'agit donc d'une régression linéaire relativement aux coefficients A et B , dont les valeurs optimum A_1 et B_1 sont classiquement obtenues (avec conventionnellement $\sum \equiv \sum_{k=1}^n$) :

$$\begin{pmatrix} A_1 \\ B_1 \end{pmatrix} = \begin{pmatrix} \sum (x_k - x_1)^2 & \sum (x_k - x_1) S_k \\ \sum (x_k - x_1) S_k & \sum (S_k^2) \end{pmatrix}^{-1} \begin{pmatrix} \sum (y_k - y_1)(x_k - x_1) \\ \sum (y_k - y_1) S_k \end{pmatrix} \quad (11)$$

Ensuite, (10) donne les valeurs optimum a_1 et c_1 : $a_1 = -\frac{A_1}{B_1}$; $c_1 = B_1$ (12)

La forme de l'équation intégrale choisie n'ayant conduit qu'à optimiser deux paramètres (le troisième intervenant dans les valeurs numériques mais n'apparaissant pas explicitement), une seconde régression est nécessaire pour l'obtenir. En fait, compte tenu de la forme de l'équation (2), on aura un résultat encore meilleur en effectuant la régression linéaire relativement aux deux paramètres a et b :

$$\sum_{k=1}^n \varepsilon_k^2 = \sum_{k=1}^n (a + b \exp(c_1 x_k) - y_k)^2 \quad (13)$$

et en posant : $c_2 = c_1$; $\theta_k = \exp(c_2 x_k)$ (14)

$$\begin{pmatrix} a_2 \\ b_2 \end{pmatrix} = \begin{pmatrix} n & \sum \theta_k \\ \sum \theta_k & \sum \theta_k^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k \\ \sum y_k \theta_k \end{pmatrix} \quad (15)$$

En résumé, la procédure de calcul est la suivante :

Données dans le cas : $y = a + b X^c$
 $(X_1, y_1), (X_2, y_2), \dots, (X_k, y_k), \dots, (X_n, y_n)$
 - Calcul des $x_k = \ln(X_k)$

Données dans le cas $y = a + b \exp(c x)$:
 $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

- Classement des points par ordre croissant des x_k
- Calcul des S_k , équation (7)
- Calcul de $\sum (x_k - x_1)^2, \sum (x_k - x_1) S_k, \sum S_k^2,$
 $\sum (y_k - y_1)(x_k - x_1), \sum (y_k - y_1) S_k$
- Calcul de B_1 , système [11]
- avec $c_2 = c_1 = B_1$, calcul des θ_k (relations 12 et 14)
- Calcul de $\sum \theta_k, \sum (\theta_k)^2, \sum y_k, \sum y_k \theta_k$
- Calcul de a_2 et b_2 , système [15]

Résultat :
 a_2, b_2 et c_2 sont les valeurs approchées de a, b et c

Pour illustrer ce calcul (Figure 1), les données numériques ont été générées de la manière suivante : Les x_k ont été tirés au hasard sur la plage des abscisses considérées. A partir de valeurs "exactes" données a_e , b_e et c_e , définissant la fonction $y(x)$ dite "exacte" par l'équation (2), dont la courbe représentative est tracée en pointillés sur la figure, on a calculé les $y(x_k)$ exacts correspondants. Ils ont été ensuite affectés de déviations dont l'amplitude a été tirée au hasard entre $-$ et $+$ 10% de $y(x_k)$, ce qui a donné, après arrondis, les valeurs numériques y_k indiquées et représentées par des croix sur la figure.

Finalement, le résultat (a_2 , b_2 , c_2) est reporté dans l'équation (2) de la fonction dont la courbe représentative est tracée en trait plein.

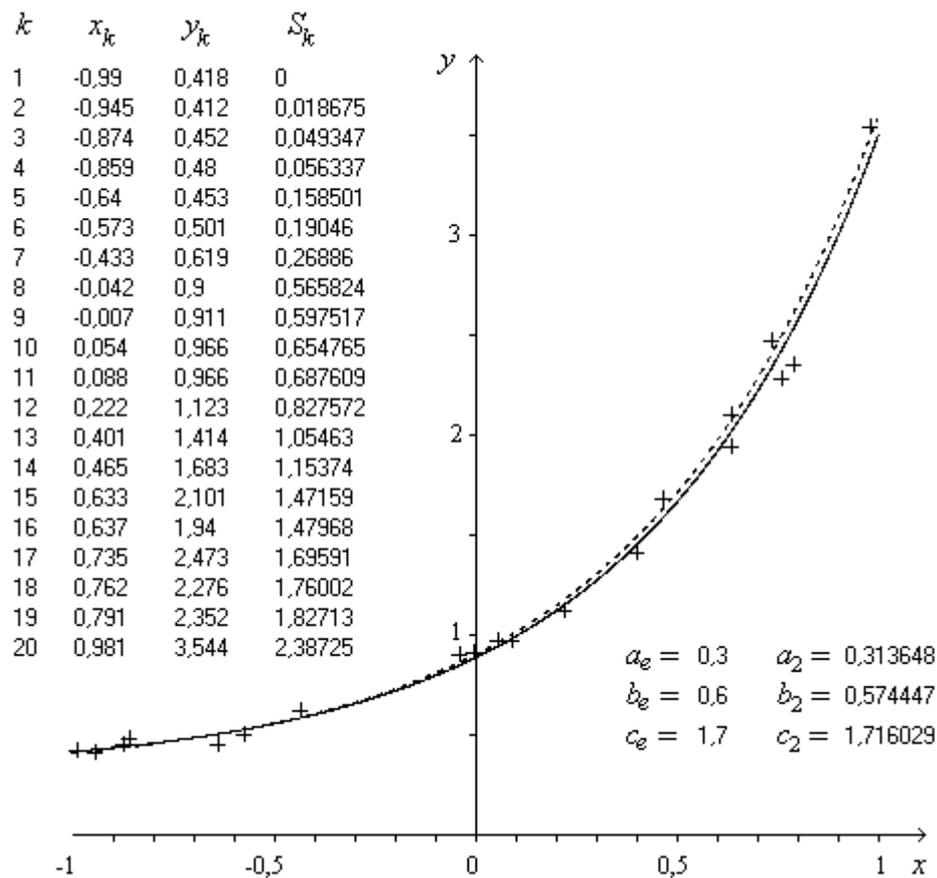


Figure 1 : Exemple de régression pour la fonction $y = a + b \exp(c x)$

Pour se faire une opinion objective des qualités et défauts de la méthode qui vient d'être exposée, il faudrait mener une étude expérimentale systématique sur un très grand nombre de cas et d'exemples. Il est à priori certain que les déviations, causées par le défaut inhérent aux intégrations numériques, seront considérablement réduites si les points sont assez nombreux et leurs abscisses réparties à intervalles assez réguliers.

3. Régression dans le cas de fonction de répartition de Weibull à trois paramètres

La fonction de répartition de Weibull à trois paramètres (α , β et μ) est définie par :

$$F(t) = 1 - \exp\left(-\left(\frac{t-\mu}{\beta}\right)^\alpha\right) \quad (16)$$

Les données étant $(t_1, F_1), \dots, (t_k, F_k), \dots, (t_n, F_n)$ on cherche à optimiser μ , β et α de telle sorte que la relation (16) soit approximativement et au mieux satisfaite pour les n points. La fonction réciproque de (16) est :

$$t = \mu + \beta(-\ln(1-F))^{1/\alpha} \quad (17)$$

Et en posant : $x = \ln(-\ln(1-F)) \quad (18)$

et : $y = t ; a = \mu ; b = \beta ; c = \frac{1}{\alpha} \quad (19)$

On voit que l'on est ramené au cas précédent : $y = a + b \exp(c x) \quad (2)$

La procédure de calcul s'en déduit immédiatement en transposant les notations :

Données :

$$(t_1, F_1), \dots, (t_k, F_k), \dots, (t_n, F_n)$$

Procédure :

- Classement des points par ordre croissant des F_k

- Calcul des $x_k = \ln(-\ln(1-F_k))$

- Calcul des S_k :

$$\begin{cases} S_1 = 0 & \text{et pour } k = 2 \rightarrow n : \\ S_k = S_{k-1} + \frac{1}{2}(t_k + t_{k-1})(x_k - x_{k-1}) \end{cases}$$

- Calcul de B :

$$\begin{pmatrix} A \\ B \end{pmatrix} = \begin{pmatrix} \sum (x_k - x_1)^2 & \sum (x_k - x_1) S_k \\ \sum (x_k - x_1) S_k & \sum (S_k^2) \end{pmatrix}^{-1} \begin{pmatrix} \sum (t_k - t_1)(x_k - x_1) \\ \sum (t_k - t_1) S_k \end{pmatrix}$$

- On obtient $\alpha_c = \frac{1}{B}$

- Calcul des $\theta_k = \exp(B x_k)$

- Calcul de β_c et μ_c :
$$\begin{pmatrix} \mu_c \\ \beta_c \end{pmatrix} = \begin{pmatrix} n & \sum \theta_k \\ \sum \theta_k & \sum \theta_k^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum t_k \\ \sum t_k \theta_k \end{pmatrix}$$

Résultat : α_c , β_c et μ_c sont les approximations de α , β et μ

Dans les représentations graphiques, il est d'usage de porter $\ln(t_k)$ en abscisses et $\ln(-\ln(1-F_k))$ en ordonnées. Ceci est un héritage de la méthode graphique de linéarisation en usage dans le cas où $\mu = 0$. Pour respecter cette tradition, il nous faut permuter les axes et porter en abscisses $\ln(t_k)$ et x_k en ordonnées.

La loi de Weibull s'applique généralement à la défaillance de matériaux ou d'objets. La variable t étant le temps, les t_k et les F_k se trouvent, de fait, ordonnés par valeurs croissantes.

Un exemple de régression est présenté Figure 2. Pour simuler une expérience, les données numériques (t_k, F_k) ont été générées à partir de valeurs "exactes" α_e, β_e et μ_e , définissant la fonction $F(t)$ dite "exacte" par l'équation (16), dont la courbe représentative est tracée en pointillés sur la figure. A chaque F_k donné, correspond une valeur de t calculée par (17). Ces valeurs de t sont ensuite affectées de déviations aléatoires simulant les aléas expérimentaux, ce qui donne les valeurs t_k indiqués. Finalement, le résultat (α_c, β_c et μ_c), reporté dans l'équation (16), donne la fonction "calculée" $F_c(t)$ (dont la courbe représentative est tracée en trait plein) :

$$F_c(t) = 1 - \exp\left(-\left(\frac{t - \mu_c}{\beta_c}\right)^{\alpha_c}\right) \quad (20)$$

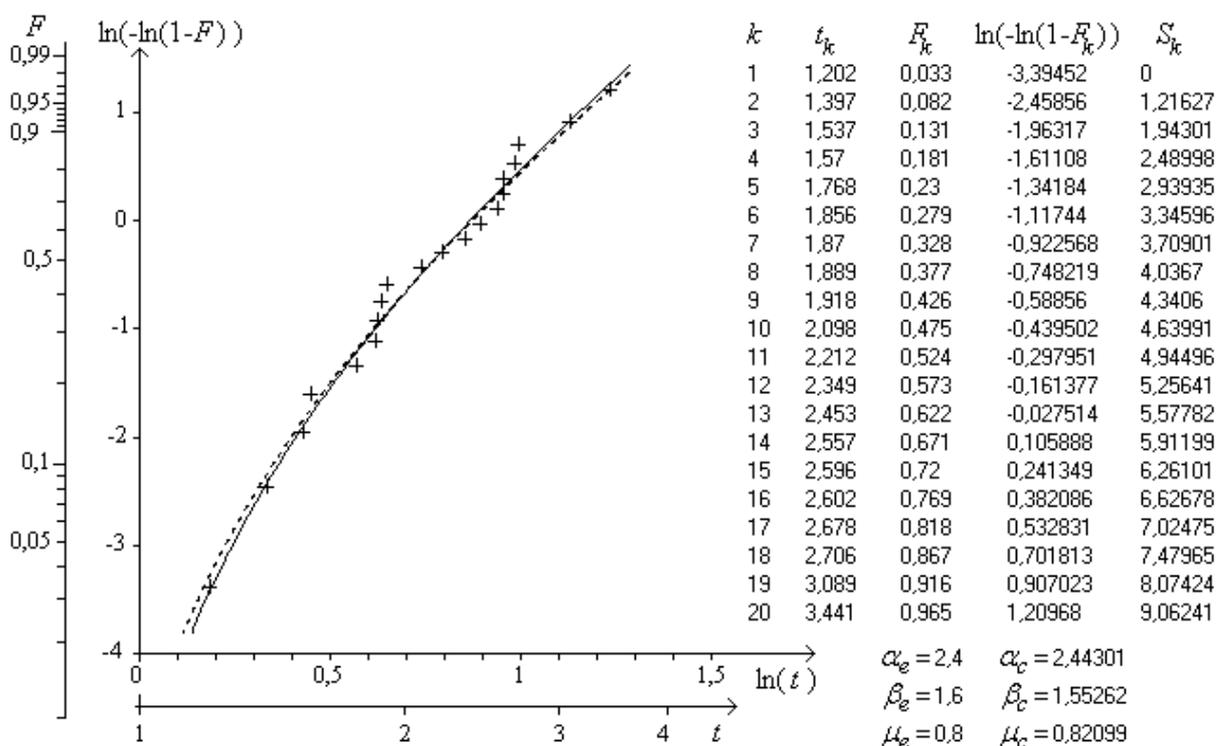


Figure 2 : Exemple dans le cas d'une fonction de répartition de Weibull à trois paramètres.

La représentation dans le système d'axes employé en pratique montre que la présence du paramètre μ_c non nul ne permet pas la linéarisation graphique habituelle, inconvénient auquel on s'attendait, bien sûr. La régression effectuée par le biais de l'équation intégrale a permis la linéarisation et l'obtention de la courbe qui se substitue ainsi à la droite traditionnelle, en améliorant ainsi sensiblement l'adéquation.

4. Conclusion

Les exemples précédents (§.2 et 3) montrent comment un problème de régression non linéaire est ramené à une régression linéaire grâce à une équation intégrale appropriée. De cette façon, les procédures itératives habituelles sont remplacées par un simple calcul linéaire.

REGRESSION SINUSOÏDALE

Jean Jacquelin

La première version du papier *Régression sinusoidale* date du 09/01/2009.

La présente version a été mise à jour le 15/02/2009

1. Introduction

Au titre apparemment anodin de "Régression sinusoidale", le sous-titre :

"Un cauchemar d'optimisation"

conviendrait bien pour ajouter une touche de réalisme. En effet, il faut avoir été concerné par ce problème pour comprendre vraiment la pertinence de ce mot. Mais de quoi s'agit-il donc ?

Ainsi que pour nombre de problèmes du même genre, les données consistent en n points expérimentaux $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$. On cherche à ajuster les paramètres d'une fonction $y = f(x)$ de telle sorte que sa courbe représentative passe "au plus près" des points donnés. En l'occurrence, il s'agit de la fonction sinusoidale suivante qui comporte quatre paramètres a, b, c et ω :

$$f(x) = a + b \sin(\omega x) + c \cos(\omega x) \quad [1]$$

Cette fonction peut être présentée sous la forme équivalente :

$$\begin{cases} f(x) = a + \rho \sin(\omega x + \varphi) \\ \rho = \sqrt{b^2 + c^2} \quad ; \quad b = \rho \cos(\varphi) \quad ; \quad c = \rho \sin(\varphi) \end{cases} \quad [2]$$

L'expression "au plus près" sous-entend un critère d'optimisation. En l'occurrence, on considère la somme des carrés des écarts :

$$\mathcal{E}_{(a,b,c,\omega)}^2 = \sum_{k=1}^n (y_k - f(x_k))^2 = \sum_{k=1}^n (y_k - (a + b \sin(\omega x_k) + c \cos(\omega x_k)))^2 \quad [3]$$

C'est cette somme que l'on tend à rendre minimum, d'où le nom générique de "méthode des moindres carrés".

Une situation rêvée se présente lorsque l'on connaît à priori la valeur de ω . En effet, la relation [1] est alors linéaire relativement aux paramètres à optimiser (a, b et c). Ce cas bien connu ne mérite qu'un bref rappel, qui sera fait au paragraphe suivant.

Dans tous les autres cas, on est en présence d'un problème de régression (ou d'optimisation) non linéaire, du fait que la somme des carrés dépend non linéairement de ω .

Une situation presque aussi favorable se présente lorsque l'on connaît une "assez bonne" valeur approchée de ω , ce qui permet l'initialisation d'une quelconque méthode de régression non linéaire, dont il existe diverses descriptions dans la littérature et dont certaines sont implémentées dans des logiciels appropriés. En parler plus longuement ici sortirait du cadre limité auquel le présent papier est volontairement restreint.

Mais la situation cauchemardesque redoutée n'est pas loin. En effet, encore faut-il que la valeur initialement estimée de ω soit "assez bonne"... Et c'est là que la fonction sinusoïdale se distingue d'autres fonctions non linéaires plus accommodantes : Que les x_k soient distribués sur de nombreuses périodes, qu'ils soient aléatoirement répartis ou encore que les valeurs des y_k soient entachées d'imprécisions et voilà la condition "assez bonne" qui doit être remplacée par "très bonne", voire "avec grande précision". Autrement dit, il faudrait quasiment connaître d'avance le ω que l'on cherche !

La méthode originale proposée au §.3 tend à apporter un début de réponse à cette gageure. Certes, il serait abusif de prétendre que sa robustesse est totale : on observera quelques unes de ses insuffisances au §.4. Néanmoins, grâce au premier résultat obtenu, nous verrons, au §.5, qu'un mode de régression original (en dents de scie) permet de mieux approcher ω par une linéarisation améliorée. Finalement, le §.6 présente un résumé des performances observées au cours d'expérimentations systématiques. Un synoptique de l'ensemble du processus, qui ne fait intervenir aucun mode de calcul itératif, est donné en Annexe.

Avant d'entrer dans le vif du sujet, un avertissement doit être donné en ce qui concerne certaines des figures présentées (1, 2, 3, 7, 10) . Elles interviennent en tant qu'illustrations des processus de calcul. Pour ce faire, on est bien obligé de fixer des données numériques qui ne sont qu'un exemple parmi les innombrables cas possibles. Au vu de ces seules figures, il serait aberrant de se faire une opinion, favorable ou non, sur l'efficacité de la méthode en question. Ceci d'autant plus que l'exemple est dépréciatif car ayant été sélectionné de telle sorte que des défauts apparaissent exagérément pour que les explications données dans le texte puissent s'y référer sans ambiguïté.

On voit sur la figure 1 que les points dits "expérimentaux" sont peu nombreux, très mal répartis et entachés d'une forte dispersion. On se doute qu'il ne s'agit pas vraiment de mesures expérimentales, mais d'une simulation, réalisée de la façon suivante : Une fonction dite "exacte" est donnée, avec les coefficients a_e , b_e , c_e et ω_e indiqués sur la figure. La courbe représentative est tracée en pointillés. Les x_k ont été tirés au hasard sur la plage considérée des abscisses, puis arrondies aux valeurs indiquées sur la figure. Les valeurs exactes des y correspondants sont d'abord calculées par l'équation [1]. Pour donner les y_k , elles ont été ensuite affectées de déviations aléatoires de telle sorte que l'écart quadratique moyen σ_e atteigne environ 10% de l'amplitude ρ_e de la sinusoïde, ce qui représente une dispersion très forte et généralement bien plus grande que celles rencontrées en pratique courante.

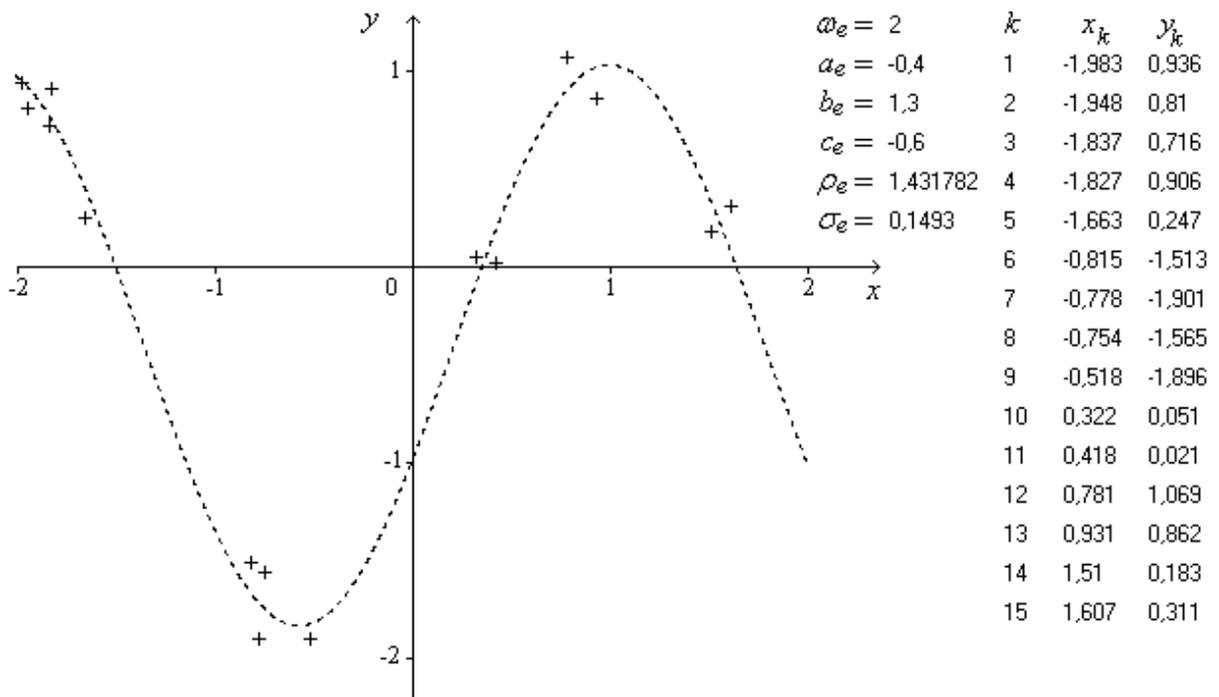


Figure 1 : Sinusoïde "exacte" et données numériques de l'exemple.

L'écart quadratique moyen est défini par :
$$\sigma = \sqrt{\frac{1}{n} \sum_{k=1}^n (f(x_k) - y_k)^2}$$
 [4]

La fonction $f(x)$ étant calculée avec les paramètres a, b, c et ω correspondants à l'exemple considéré.

Les valeurs arrondies des y_k , indiquées sur la figure 1, constituent les données de l'exemple numérique illustrant les paragraphes suivants. Il convient de noter que les valeurs exactes de a_e, b_e, c_e et ω_e qui sont reportées sur la figure, ne sont pas utilisées par la suite (sauf ω_e dans le cas particulier "simpliste" du §.2). Elles doivent être oubliées dans tout le processus que nous verrons et qui n'a, pour base des calculs, que les (x_k, y_k) donnés. Bien que, pour mémoire, la sinusoïde "exacte" apparaisse en pointillés sur les autres figures, cela ne veut pas dire que la fonction [1] avec les paramètres $(a_e, b_e, c_e, \omega_e)$ soit utilisée, ce qui n'est pas.

Les figures ont un autre rôle qui sera apprécié des personnes désirant mettre au point un programme d'ordinateur : Les données et résultats numériques portées dessus permettent, si besoin est, de reproduire exactement les calculs de l'exemple donné ici. Ainsi, un logiciel de régression, construit selon le principe décrit, pourra être vérifié et éventuellement corrigé.

2. Cas où ω est connu à priori

La valeur $\omega = \omega_e$ étant fixée, l'optimisation ne porte alors que sur les paramètres a, b et c de [1]. Les dérivées partielles de [3], relativement aux paramètres à optimiser, conduisent au système de trois équations :

$$\begin{cases} \left(\frac{\partial \mathcal{E}^2}{\partial a} \right)_{(a_0, b_0, c_0)} = -2 \sum_{k=1}^n (y_k - (a_0 + b_0 \sin(\omega_e x_k) + c_0 \cos(\omega_e x_k))) = 0 \\ \left(\frac{\partial \mathcal{E}^2}{\partial b} \right)_{(a_0, b_0, c_0)} = -2 \sum_{k=1}^n (y_k - (a_0 + b_0 \sin(\omega_e x_k) + c_0 \cos(\omega_e x_k))) \sin(\omega_e x_k) = 0 \\ \left(\frac{\partial \mathcal{E}^2}{\partial c} \right)_{(a_0, b_0, c_0)} = -2 \sum_{k=1}^n (y_k - (a_0 + b_0 \sin(\omega_e x_k) + c_0 \cos(\omega_e x_k))) \cos(\omega_e x_k) = 0 \end{cases} \quad [5]$$

La solution est donnée par le système [6] suivant, avec conventionnellement $\sum \equiv \sum_{k=1}^n$:

$$\begin{pmatrix} a_0 \\ b_0 \\ c_0 \end{pmatrix} = \begin{pmatrix} n & \sum \sin(\omega_e x_k) & \sum \cos(\omega_e x_k) \\ \sum \sin(\omega_e x_k) & \sum \sin^2(\omega_e x_k) & \sum \sin(\omega_e x_k) \cos(\omega_e x_k) \\ \sum \cos(\omega_e x_k) & \sum \sin(\omega_e x_k) \cos(\omega_e x_k) & \sum \cos^2(\omega_e x_k) \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k \\ \sum y_k \sin(\omega_e x_k) \\ \sum y_k \cos(\omega_e x_k) \end{pmatrix} \quad [6]$$

Le résultat obtenu est présenté en figure 2. On observe en particulier que l'écart quadratique moyen σ_0 est pratiquement le même que σ_e sur la figure 1. Cela signifie que la régression effectuée n'a aucunement augmenté la dispersion des points expérimentaux par rapport à la sinusoïde obtenue en remplacement de la sinusoïde "exacte". Evidemment, ce serait trop beau et trop facile s'il en était toujours ainsi. En effet, après ce bref rappel, il faut aborder la difficulté essentielle : calculer une approximation satisfaisante de ω , lorsque la valeur exacte n'est pas connue a priori comme supposé dans ce qui précède.

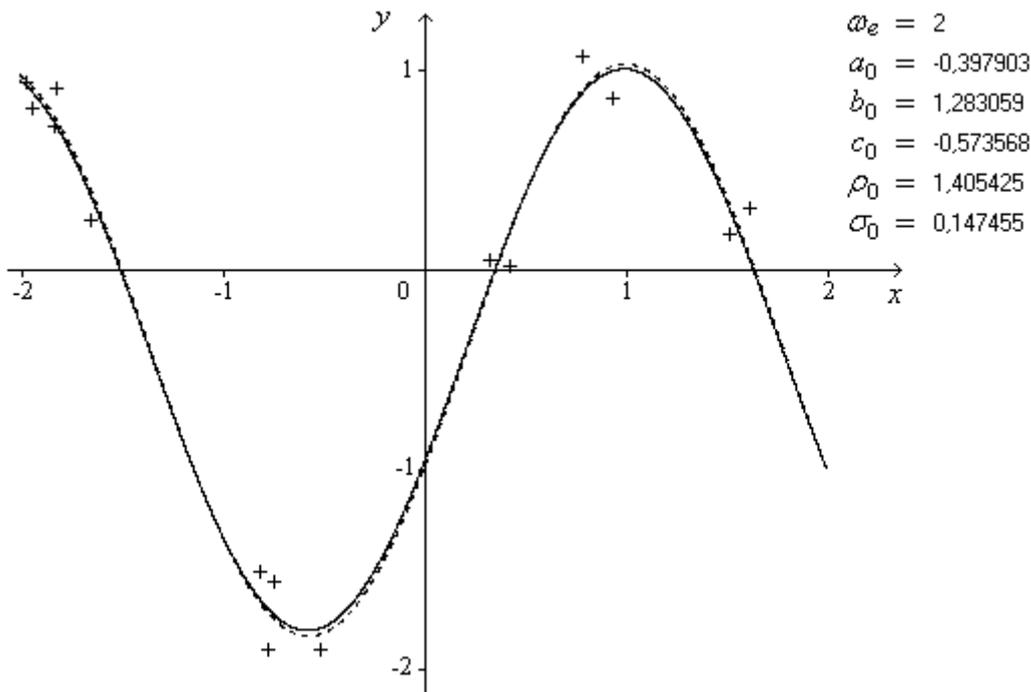


Figure 2 : Cas où ω est a priori exactement connu.

3. Linéarisation par une équation intégrale

L'idéal étant de ramener le problème à une forme de régression linéaire, il est parfois tentant de passer par l'intermédiaire d'une équation différentielle linéaire ayant pour solution la fonction considérée, comme par exemple dans le cas présent, l'équation suivante dont la

$$\text{fonction sinusoïdale [1] est solution : } f(x) = a - \frac{1}{\omega^2} \frac{d^2 f(x)}{dx^2} \quad [7]$$

ce qui conduirait, après les dérivations partielles de [8] à un système linéaire à deux inconnues a et $\beta = 1/\omega^2$

$$\mathcal{E}_{(a,b,c,\omega)}^2 = \sum_{k=1}^n (y_k - f(x_k))^2 = \sum_{k=1}^n (y_k - a - \beta y''(x_k))^2 \quad [8]$$

Malheureusement ce n'est pas viable en pratique (sauf éventuellement si l'on dispose d'un grand nombre de points expérimentaux très bien distribués). En effet la pierre d'achoppement est le calcul des $y''(x_k)$ à partir des n points (x_k, y_k) donnés : les déviations sont beaucoup trop grandes en général (on en verra une illustration sur la figure 3).

Par contre, au lieu des dérivées, le calcul numérique des intégrales est nettement moins problématique. Il n'est donc pas étonnant à ce que l'on s'oriente plutôt vers une équation intégrale ayant pour solution la fonction considérée, comme par exemple dans le cas présent, l'équation [9] dont la fonction sinusoïdale [1] est solution :

$$f(x) = -\omega^2 \int_{x_i}^x \int_{x_j}^v f(u) du dv + P(x) \quad [9]$$

$P(x)$ est un polynôme du second degré dans lequel les coefficients dépendent de a, b, c, ω et des bornes inférieures d'intégration x_i et x_j .

Bien entendu, $P(x)$ peut être complètement explicité, mais cela alourdirait l'exposé sans grand intérêt immédiat. Mieux encore, l'étude complète montre que le choix des bornes inférieures d'intégration n'a pas d'influence sur la régression qui va suivre, en ce qui concerne son résultat essentiel, c'est-à-dire la valeur de ω optimisée (qui sera notée ω_1). Ainsi, pour simplifier, nous posons : $x_i = x_j = x_1$ ce qui conduit à une fonction de la forme suivante :

$$f(x) = A SS(x) + B x^2 + C x + D \quad [10]$$

$$\text{avec : } \begin{cases} SS(x) = \int_{x_1}^x \int_{x_1}^v f(u) du dv & ; & A = -\omega^2 & ; & B = \frac{1}{2} a \omega^2 \\ C = -a \omega^2 x_1 + b \omega \cos(\omega x_1) - c \omega \sin(\omega x_1) & & & & \\ D = a + \frac{1}{2} a \omega^2 x_1^2 + (b + c \omega x_1) \sin(\omega x_1) + (c - b \omega x_1) \cos(\omega x_1) & & & & \end{cases} \quad [11]$$

Les coefficients A , B , C et D ne sont pas connus. Mais ils peuvent être optimisés par régression linéaire, à condition que les $SS(x_k)$ soient préalablement calculés. Pour ce faire nous effectuons successivement deux intégrations numériques :

$$\begin{cases} S(x_1) = S_1 = 0 \\ S(x_k) \approx S_k = S_{k-1} + \frac{1}{2}(y_k + y_{k-1})(x_k - x_{k-1}) \quad k = 2 \text{ à } n \end{cases} \quad [12]$$

$$\begin{cases} SS(x_1) = SS_1 = 0 \\ SS(x_k) \approx SS_k = SS_{k-1} + \frac{1}{2}(S_k + S_{k-1})(x_k - x_{k-1}) \quad k = 2 \text{ à } n \end{cases} \quad [13]$$

Bien entendu, ceci suppose que les (x_k, y_k) aient été préalablement ordonnés selon les x croissants. La somme des carrés des écarts à minimiser est donc la suivante :

$$\mathcal{E}_{(a,b,c,\omega)}^2 = \sum_{k=1}^n \left(y_k - (A SS_k + B x_k^2 + C x_k + D) \right)^2 \quad [14]$$

Inutile de revenir sur la méthode de dérivations partielles déjà vue, relativement à A , B , C et D , pour obtenir l'optimum (A_1, B_1, C_1, D_1) , puis successivement ω_1, a_1, b_1, c_1 (d'après [11]) :

$$\begin{pmatrix} A_1 \\ B_1 \\ C_1 \\ D_1 \end{pmatrix} = \begin{pmatrix} \sum (SS_k)^2 & \sum x_k^2 SS_k & \sum x_k SS_k & \sum SS_k \\ \sum x_k^2 SS_k & \sum x_k^4 & \sum x_k^3 & \sum x_k^2 \\ \sum x_k SS_k & \sum x_k^3 & \sum x_k^2 & \sum x_k \\ \sum SS_k & \sum x_k^2 & \sum x_k & n \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k SS_k \\ \sum y_k x_k^2 \\ \sum y_k x_k \\ \sum y_k \end{pmatrix} \quad [15]$$

$$\begin{cases} \omega_1 = \sqrt{-A_1} & ; & a_1 = \frac{2 B_1}{\omega_1^2} \\ b_1 = (B_1 x_1^2 + C_1 x_1 + D_1 - a_1) \sin(\omega_1 x_1) + \frac{1}{\omega_1} (C_1 + 2 B_1 x_1) \cos(\omega_1 x_1) \\ c_1 = (B_1 x_1^2 + C_1 x_1 + D_1 - a_1) \cos(\omega_1 x_1) - \frac{1}{\omega_1} (C_1 + 2 B_1 x_1) \sin(\omega_1 x_1) \end{cases} \quad [16]$$

L'ensemble des résultats obtenus concernant l'exemple étant regroupés plus loin sur la figure 10, il convient de s'y reporter pour y trouver les valeurs numériques de ω_1, a_1, b_1, c_1 , et voir la représentation graphique de la fonction sinusoidale correspondante, repérée (1), par rapport à la position des points (x_k, y_k) .

Incidentement, il est intéressant de comparer, figure 3, ce que donnent les intégrations numériques [12] et [13], comparées à une dérivation. En voyant les oscillations importantes pour cette dernière (exemple de point noté : f'_k), et encore plus si on effectuait une seconde dérivation (non représentée en raison de l'amplitude excessive des fluctuations), il est patent que la méthode basée sur l'équation intégrale est nettement plus fiable que pourrait l'être celle basée sur l'équation différentielle.

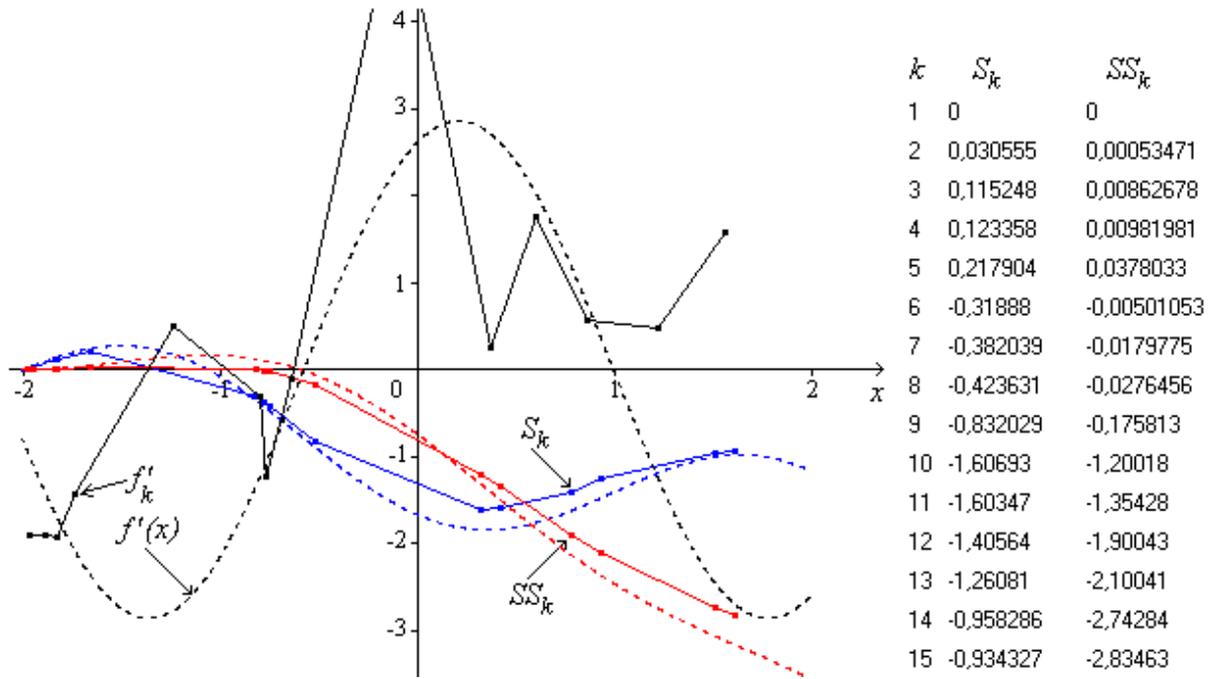


Figure 3 : Intégrations numériques et comparaison avec une dérivation.

Néanmoins on reste conscient que, si les points étaient plus nombreux, les x_k mieux répartis et les y_k moins affectés de dispersion, les fluctuations seraient d'ampleur plus acceptable. Toutefois, n'oublions pas que nous ne cherchons pas ici à traiter des cas faciles, mais au contraire les cas difficiles.

4. Etude succincte des performances

Le paramètre ω est celui dont l'optimisation est primordiale. On comprend bien que si cette optimisation est réussie et quoi qu'il en soit de a , b et c , on pourra toujours se rabattre ensuite sur la régression classique vue au §.2 si nécessaire. On se contentera donc d'une investigation restreinte aux résultats vis-à-vis de ω . Les trois facteurs les plus influents sont :

- le nombre n_p de points par période de la sinusoïde,
- le mode de répartition des points selon les abscisses :
 - soit équidistance : $x_{k+1}-x_k = \text{constante}$
 - soit aléatoire : x_k est tiré au hasard sur la plage allouée aux x .
- la dispersion des ordonnées y_k , caractérisée par (σ_1/ρ_e) c'est-à-dire le rapport entre l'écart quadratique moyen [4] et l'amplitude de la sinusoïde.

4.1. Répartition "équidistante" des abscisses et dispersion nulle des ordonnées :

On constate que le résultat ω_1/ω_e , espéré égal à 1, est affecté d'une déviation qui dépend de n_p et d'autant plus faible que n_p est grand (Figure 4). Il serait envisageable d'en déduire une fonction empirique permettant de la corriger. Mais cela n'aurait pas grand intérêt car la correction ne serait pas satisfaisante dans les cas de répartition aléatoire des points qui sera étudiée plus loin. Une méthode plus générale, dont le principe est exposé au §.5, apparaît mieux appropriée.

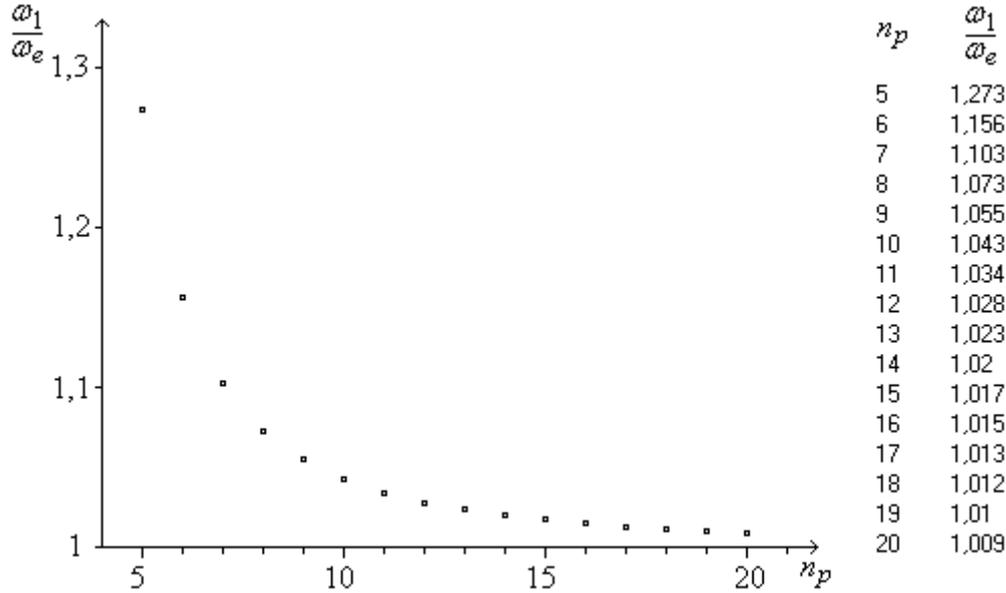


Figure 4 : Influence du nombre de points par période, répartis de façon équidistante.

4.2. Répartition aléatoire des abscisses des points, sans dispersion des ordonnées :

Du fait que les x_k sont donnés aléatoirement, le calcul pour des répartitions successivement tirées au hasard donne des ω_1 d'autant plus dispersés que n_p est petit. Pour une valeur de n_p fixée, la fonction de répartition correspondante (résultant de 10000 simulations) est représentée en figure 5. On constate que le résultat ω_{1m}/ω_e , espéré égal à 1, est affecté d'une déviation encore plus importante que dans les conditions du §.4.1.

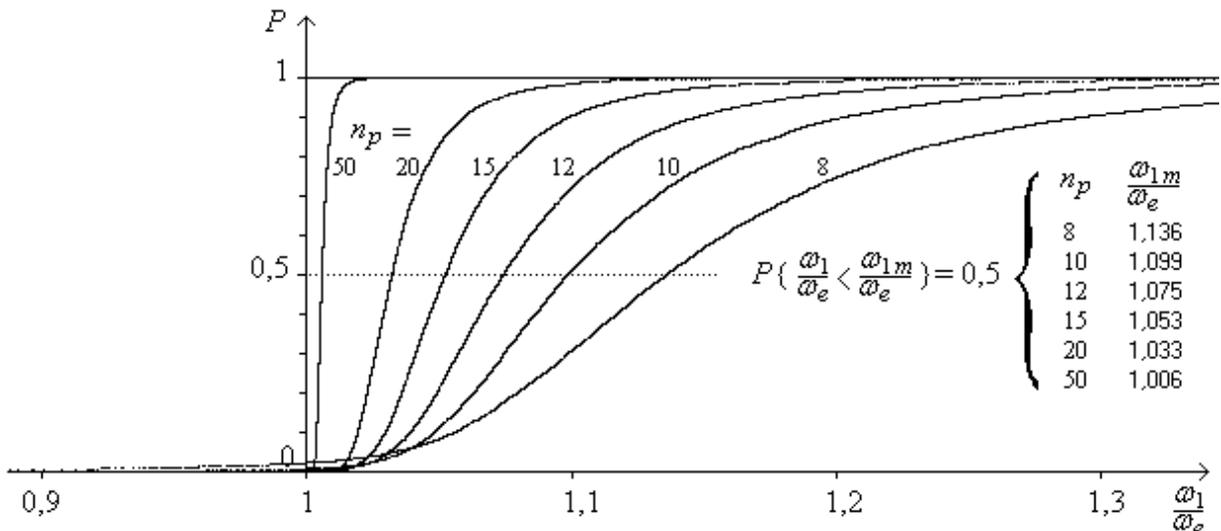


Figure 5 : Fonctions de répartition de ω_1 (répartition aléatoire des x_k , dispersion nulle des y_k)

4.3. Répartition aléatoire des abscisses des points, avec des ordonnées dispersées:

Ainsi que l'on devait s'y attendre, la dispersion des ω_1 est plus étendue que dans le cas précédent, ainsi qu'on le voit nettement sur la figures 6 , dans le cas de $(\sigma_1/\rho_e) = 10\%$, à comparer avec le cas $(\sigma_1/\rho_e) = 0$ représenté en figure 5. Néanmoins, la valeur médiane est peu affectée.

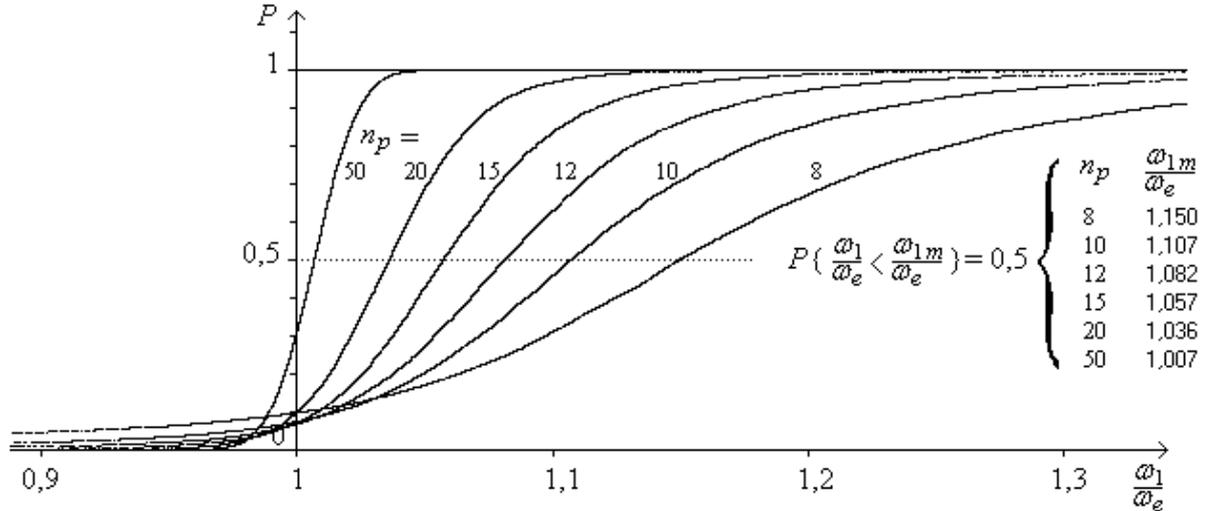


Figure 6 : Fonctions de répartition de ω_1 avec dispersion sur les ordonnées ($\sigma_1/\rho_e = 0,1$)

5. Cas où les paramètres a et ρ sont approximativement connus

Nous nous intéressons maintenant à la fonction $y = f(x)$ exprimée sous sa forme [2], dont la fonction réciproque fait apparaître un arcsin que l'on converti ensuite en arctg :

$$\begin{cases} \Phi(x) = \arctg \left(\frac{f(x) - a}{\sqrt{\rho^2 - (f(x) - a)^2}} \right) \\ \omega x + \varphi = \pm \Phi(x) + \pi K_{(x)} \end{cases} \quad [17]$$

arctg désigne la détermination principale (comprise entre $-\pi/2$ et $+\pi/2$) de la fonction multiforme. Le signe de Φ et l'entier relatif $K_{(x)}$ dépendent de la demi période de la sinusoïde sur laquelle le point (x, y) se trouve, donc dépendent de x d'une façon discontinue. On montre d'ailleurs que le signe est + si $K_{(x)}$ est pair et - s'il est impair.

Si on considère $\Phi(x)$ isolément, il s'agit d'une fonction en dents de scie (figure 7, courbe en pointillés). Les points (x_k, Φ_k) avec $\Phi_k = \Phi(x_k)$ sont repérés par des croix. Présenté ainsi, le problème devient celui d'une régression en dents de scie, qui est largement aussi "cauchemardesque" que celui de la régression sinusoïdale. En effet, lorsque rien d'autre n'est connu que les points (x_k, Φ_k) , la détermination de $K_{(x)}$ est difficile et très empirique, donc d'une fiabilité non assurée dans le cas général. La situation est différente dans notre cas puisqu'on a déjà obtenu les ordres de grandeur des paramètres : a_1, b_1, c_1, ω_1 , ainsi que ρ_1 et φ_1 par les relations [2]. Dans le cas présent, qui sera repéré par l'indice 2, nous posons :

$$a_2 = a_1 \quad ; \quad \rho_2 = \rho_1 = \sqrt{b_1^2 + c_1^2} \quad ; \quad \rho_1 \cos(\varphi_1) = b_1 \quad ; \quad \rho_1 \sin(\varphi_1) = c_1$$

$$\text{Si } b_1 > 0 \rightarrow \varphi_1 = \arctg\left(\frac{c_1}{b_1}\right) \quad ; \quad \text{si } b_1 < 0 \rightarrow \varphi_1 = \arctg\left(\frac{c_1}{b_1}\right) + \pi \quad [18]$$

Il s'agit d'abord de calculer $K_1, K_2, \dots, K_k, \dots, K_n$, ce qui peut être fait de diverses façons, par exemple la suivante, dans laquelle la fonction round consiste à arrondir un réel au plus proche entier :

$$K_k = \text{round}\left(\frac{\omega_1 x_k + \varphi_1}{\pi}\right) \quad [19]$$

Une autre écriture de [17], appliquée aux $x = x_k$, on montre mieux relation de linéarité approximative $\theta_k \approx \omega_2 x_k + \varphi_2$ entre x_k et θ_k défini par :

$$\begin{cases} \theta_k = (-1)^{K_k} \arctg\left(\frac{y_k - a_2}{\sqrt{\rho_2^2 - (y_k - a_2)^2}}\right) + \pi K_k \\ \text{si } \rho_2^2 \leq (y_k - a_2)^2 \rightarrow \arctg = +\frac{\pi}{2} \text{ si } y_k > a_2 \text{ ou } -\frac{\pi}{2} \text{ si } y_k < a_2 \end{cases} \quad [20]$$

On voit, sur la figure 7, que la série de points (x_k, Φ_k) repérée par des croix est transformée en une série de points (x_k, θ_k) repérés par des carrés, dont certains coïncident.

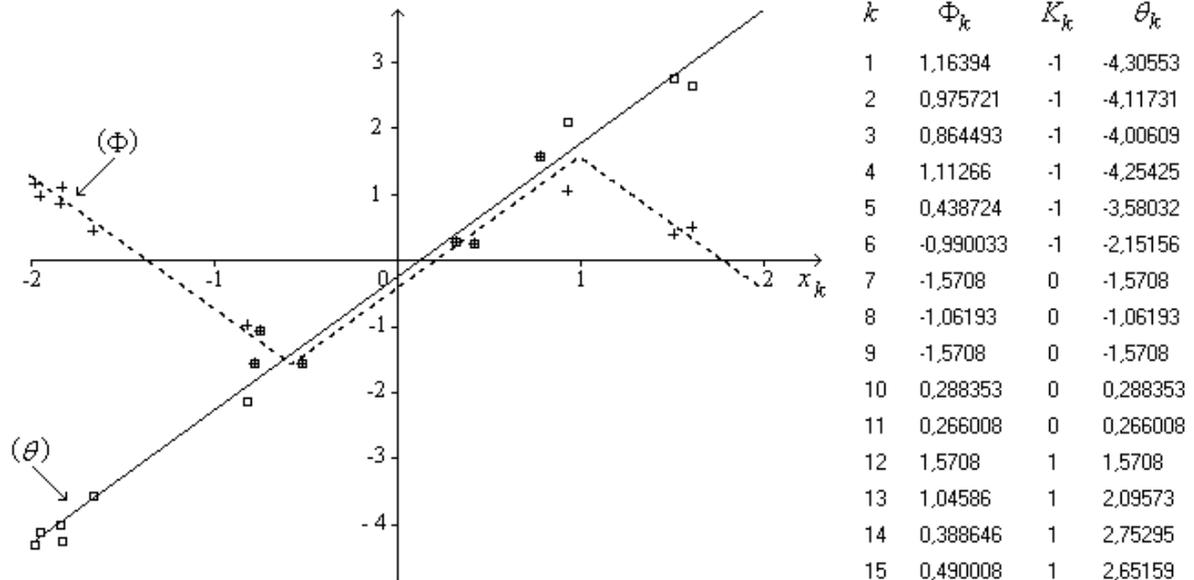


Figure 7 : Transformation de la fonction en dents de scie en vue de régression linéaire

Il est clair que les points (x_k, θ_k) tendent à s'aligner, ce qui rend possible la régression linéaire ayant pour but l'optimisation des paramètres de la droite : ω_2 est son coefficient directeur et φ_2 son abscisse à l'origine. Pour ce faire, on calcule d'abord les θ_k par la relation [20]. Ensuite, le résultat est obtenu classiquement :

$$\begin{pmatrix} \omega_2 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} \sum (x_k)^2 & \sum x_k \\ \sum x_k & n \end{pmatrix}^{-1} \begin{pmatrix} \sum \theta_k x_k \\ \sum \theta_k \end{pmatrix} \quad [21]$$

$$\text{Complété par : } b_2 = \rho_2 \cos(\varphi_2) \quad ; \quad c_2 = \rho_2 \sin(\varphi_2) \quad [22]$$

Les résultats numériques sont récapitulés sur la figure 10 où la sinusoïde correspondante est tracée (repère 2).

A cet état d'avancement des calculs, les performances relativement aux optimisations de ω sont résumées sur les figures 8 et 9. On constate que la déviation sur la valeur médiane ω_m est bien corrigée, ce qui était l'objectif de cette seconde phase du processus d'optimisation.

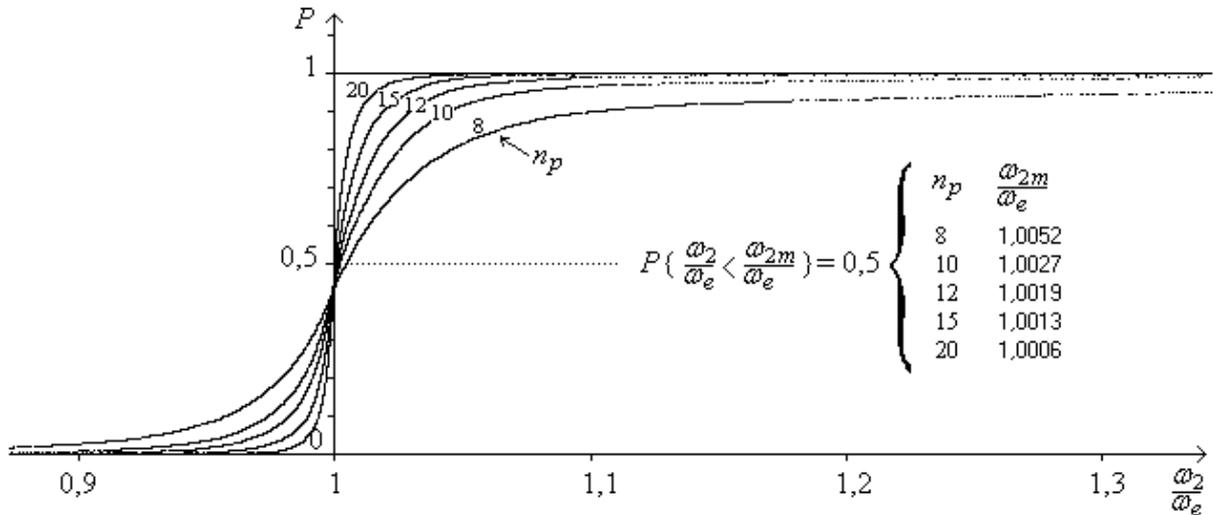


Figure 8 : Fonctions de répartition de ω_2 (répartition aléatoire des x_k , dispersion nulle des y_k)

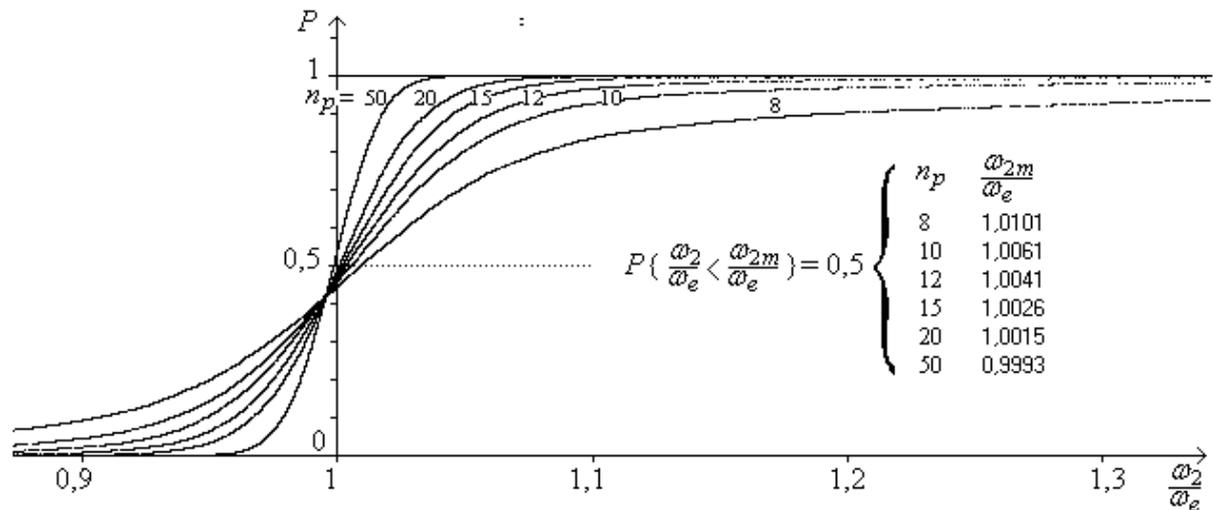


Figure 9 : Fonctions de répartition de ω_2 avec dispersion sur les ordonnées ($\sigma_2/\rho_e = 0,1$)

6. Résultats du processus d'optimisation complet.

A partir du moment où on dispose d'une valeur approchée (ω_2) de ω , la méthode classique de régression rappelée au §.2 peut être appliquée :

$$\begin{pmatrix} a_3 \\ b_3 \\ c_3 \end{pmatrix} = \begin{pmatrix} n & \sum \sin(\omega_2 x_k) & \sum \cos(\omega_2 x_k) \\ \sum \sin(\omega_2 x_k) & \sum \sin^2(\omega_2 x_k) & \sum \sin(\omega_2 x_k) \cos(\omega_2 x_k) \\ \sum \cos(\omega_2 x_k) & \sum \sin(\omega_2 x_k) \cos(\omega_2 x_k) & \sum \cos^2(\omega_2 x_k) \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k \\ \sum y_k \sin(\omega_2 x_k) \\ \sum y_k \cos(\omega_2 x_k) \end{pmatrix} \quad [23]$$

Ceci permet une optimisation finale des paramètres dimensionnels de la sinusoïde, ce qui n'était pas réalisé dans l'étape précédente, puisque a_2 et ρ_2 étaient fixés. Le résultat final, pour l'exemple considéré, est présenté sur la figure 10 (sinusoïde et ses paramètres, repère 3). Les résultats intermédiaires (sinusoïdes 1 et 2 avec leurs paramètres respectifs) sont reportés sur la même figure.

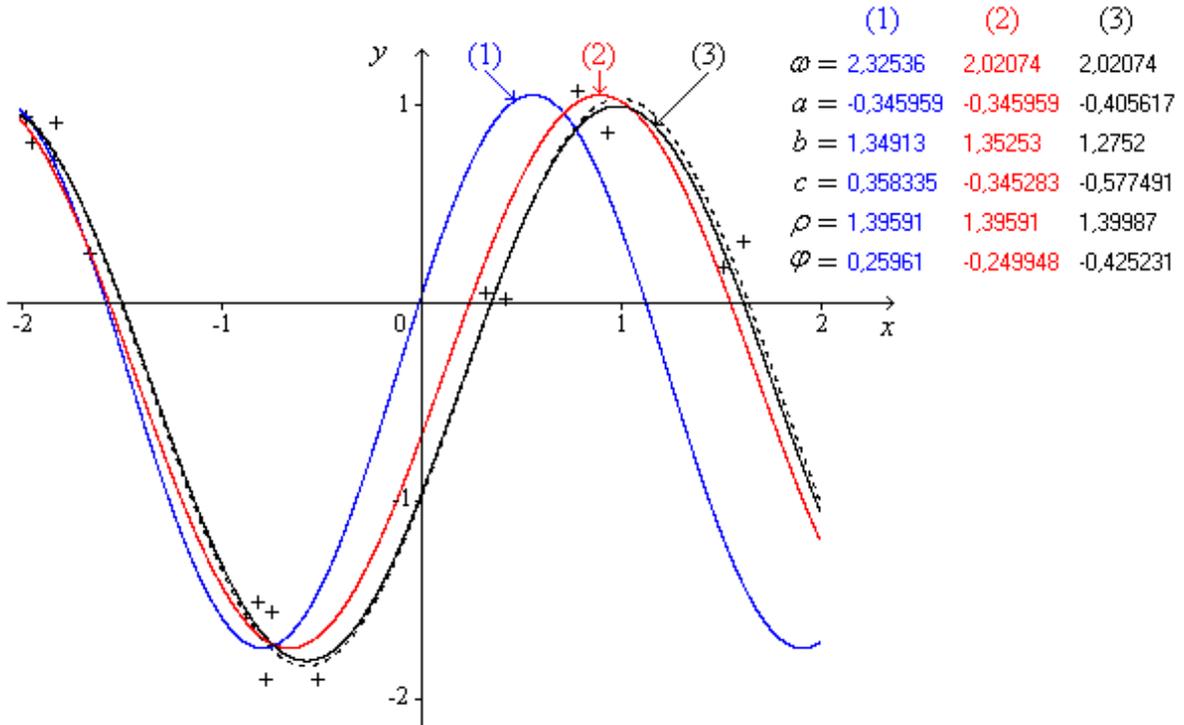


Figure 10 : Exemple de résultat de la régression sinusoïdale.

Cette figure pourrait faire croire que le procédé s'apparente à un calcul par approximations successives tendant à faire converger les courbes en trait plein vers celle en pointillés. Il n'en est rien car l'itération du processus reste sans effet : la répétition du calcul ne modifierait pas le résultat puisqu'elle est basée sur les intégrations numériques initiales, dont les déviations sont toujours présentes et ne diminueraient donc pas d'un cycle d'itération au suivant.

Pour se faire une opinion aussi objective que possible sur les propriétés de la méthode, il faut effectuer un grand nombre de simulations dans différentes conditions. Il convient de se reporter aux figures 8 et 9 pour un résumé des résultats concernant le paramètre ω , dont l'optimisation est primordiale. En effet, il n'y a pas de changement ensuite puisque $\omega_3 = \omega_2$.

Pour chaque simulation dont le résultat est $(\omega_3, a_3, b_3, c_3)$, l'écart quadratique moyen σ_3 est calculé par :

$$\sigma_3 = \sqrt{\frac{1}{n} \sum_{k=1}^n (y_k - (a_3 + b_3 \sin(\omega_3 x_k) + c_3 \cos(\omega_3 x_k)))^2} \quad [24]$$

Pour une valeur de n_p fixée, la fonction de répartition de (σ_3/ρ_e) est tracée à la suite de 10000 simulations, chacune différant des autres par les (x_k, y_k) puisque les ordonnées sont affectées d'une dispersion caractérisée par un écart quadratique moyen σ_e (défini au §.1). Pour

le tracé de la figure 11, les dispersions sont réalisées de telle sorte que le rapport σ_e / ρ_e soit toujours le même soit 0,1 . Par contre, les abscisses des points sont données différemment :

- Figure 11(a) : les abscisses des points successifs sont à intervalles égaux.
- Figure 11(b) : les abscisses sont distribuées au hasard (avec n_p points par période)

Il peut paraître surprenant au premier abord que les écarts quadratiques moyens σ_3 obtenus sont en général nettement inférieurs au σ_e initial. Cela se comprend du fait que l'on remplace la sinusoïde dite "exacte" par une sinusoïde "optimisée" qui, en fin de compte, passe souvent plus près des points lorsque leur dispersion est forte. Ceci est lié à l'optimisation de ω_2 qui peut différer notablement de ω_e comme on l'a vu sur les figures 8 et 9. De ce fait, dans le cas de forte dispersion des ordonnées des points, il n'y a pas d'aggravation aussi importante qu'on aurait pu le craindre lorsque l'on passe du cas (a) au cas (b).

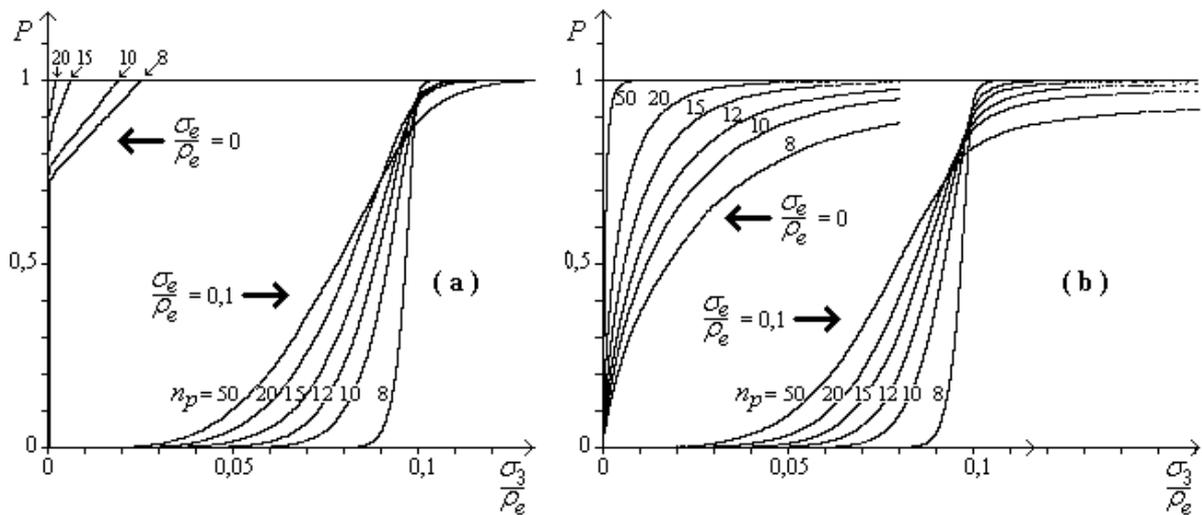
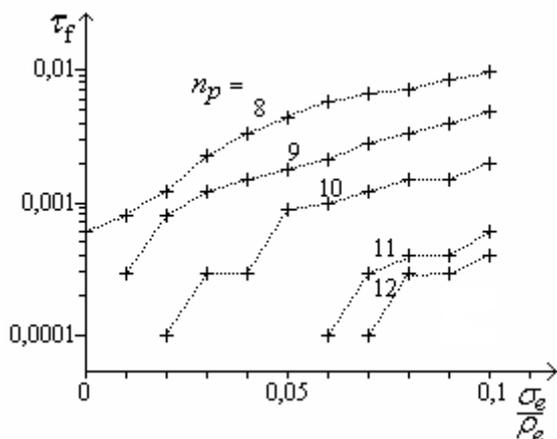


Figure 11 : Fonctions de répartition des écarts quadratiques moyens, avec répartition des abscisses : (a) équidistante ; (b) au hasard.

La plus grande différence entre les cas (a) et (b) se situe au niveau du taux d'échecs du calcul. Il faut bien en parler, car c'est le lot commun de toutes les méthodes, lorsque les points ne sont pas régulièrement distribués. Ceci est inévitable : il y a une probabilité, certes faible mais non nulle, que l'on tombe sur une situation où tous les points sont proches les uns des autres : au quel cas la sinusoïde n'est pas définie et ne peut donc pas être caractérisée, ni même approchée. Cela se traduit de diverses façons selon la méthode de calcul, par exemple une indétermination (division par un nombre trop voisin de zéro), inversion impossible d'une matrice, racine carrée d'un nombre négatif , etc.



Dans la méthode décrite ici, c'est majoritairement au niveau du calcul de ω_1 , dans [16], que l'échec se révèle, heureusement très rarement. Pour des centaines de milliers de simulations effectuées, **aucun échec ne s'est produit dans les cas de type (a)**, ce qui n'est pas étonnant puisque les points ne peuvent jamais être tous groupés. Par contre, dans les cas de type (b), on observe des échecs dont la fréquence, très faible, dépend de la dispersion des données et du nombre de points, ainsi que le montre la figure 12 (qui a nécessité près de cinq cent mille simulations pour qu'une vue à peu près cohérente puisse en ressortir).

Figure 12 : Taux d'échecs (ordres de grandeur).

7. Commentaires

Il serait présomptueux de prétendre que la méthode présentée ici est l'ultime solution au difficile problème de la régression sinusoïdale. Conjecturons plutôt qu'elle offrira des avantages pour certains créneaux d'applications : Sa robustesse surprend parfois dans des situations hasardeuses où l'on ne dispose que d'un relativement faible nombre de points, éventuellement assez mal répartis par exemple. De plus, il est possible que les cas où la régression porte sur un petit nombre de périodes de la sinusoïde, voire même sur une fraction de période, lui soient plus favorables.

Certes, la méthode elle-même n'est pas du genre "calcul par approximations successives" et n'est pas fondamentalement adaptée à évoluer vers un processus itératif, si l'on souhaitait améliorer ses performances. Mais rien n'empêche de la prendre en considération pour obtenir une première approche de qualité, pouvant servir à initialiser un calcul récursif selon une méthode différente.

D'un autre point de vue, la simplicité du principe sur laquelle est basé le processus de calcul et la facilité avec laquelle le programme peut être écrit et mis au point sont certainement attractifs aux yeux de certains utilisateurs potentiels.

En fait, l'écriture d'un programme basé sur les équations qui ont été données est beaucoup plus aisé qu'il n'y paraît à la lecture des pages précédentes. Les nécessaires explications obligent à des développements et à écrire des formules intermédiaires dont on n'a nul besoin dans le programme. Il convient d'élaguer tout ce qui est superflu pour le calcul proprement dit. On en tirera la substantifique moelle, si l'on peut dire, dans l'Annexe suivante.

Appendix 1 : Résumé du processus de régression sinusoïdale

Données : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

Première partie :

[12] : Calcul de : $S_1, S_2, \dots, S_k, \dots, S_n$

[13] : Calcul de : $SS_1, SS_2, \dots, SS_k, \dots, SS_n$

[15] : Résolution du système donnant A_1, B_1, C_1, D_1

[16] : Calcul de ω_1, a_1, b_1, c_1

Deuxième partie :

[18] : $a_2 = a_1$ et calcul de $\rho_1 = \rho_2$ et φ_1

[19] : Calcul de $K_1, K_2, \dots, K_k, \dots, K_n$

[20] : Calcul de $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n$

[21] : Résolution du système donnant ω_2 et φ_2

[22] : Calcul de b_2 et c_2

Troisième partie :

[23] : Avec $\omega_3 = \omega_2$, résolution du système donnant a_3, b_3, c_3

Résultat : ω_3, a_3, b_3, c_3 sont les approximations de ω, a, b, c .

Appendix 2 : PROCEDURE détaillée de REGRESSION SINUSOIDALE

Données : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

Première partie :

- Ordonner les données selon les valeurs croissantes de x_k

- Calcul de : $S_1, S_2, \dots, S_k, \dots, S_n$:

$$S_1 = 0$$

$$S_k = S_{k-1} + \frac{1}{2}(y_k + y_{k-1})(x_k - x_{k-1}) \quad k = 2 \text{ à } n$$

- Calcul de : $SS_1, SS_2, \dots, SS_k, \dots, SS_n$:

$$SS_1 = 0$$

$$SS_k = SS_{k-1} + \frac{1}{2}(S_k + S_{k-1})(x_k - x_{k-1}) \quad k = 2 \text{ à } n$$

- Calcul des sommes :

$$\begin{aligned} & \sum_{k=1}^n x_k \quad ; \quad \sum_{k=1}^n x_k^2 \quad ; \quad \sum_{k=1}^n x_k^3 \quad ; \quad \sum_{k=1}^n x_k^4 \quad ; \\ & \sum_{k=1}^n SS_k \quad ; \quad \sum_{k=1}^n (SS_k)^2 \quad ; \quad \sum_{k=1}^n x_k SS_k \quad ; \quad \sum_{k=1}^n x_k^2 SS_k \quad ; \\ & \sum_{k=1}^n y_k \quad ; \quad \sum_{k=1}^n y_k x_k \quad ; \quad \sum_{k=1}^n y_k x_k^2 \quad ; \quad \sum_{k=1}^n y_k SS_k \end{aligned}$$

- Résolution du système donnant A_1, B_1, C_1, D_1 :

$$\begin{pmatrix} A_1 \\ B_1 \\ C_1 \\ D_1 \end{pmatrix} = \begin{pmatrix} \sum (SS_k)^2 & \sum x_k^2 SS_k & \sum x_k SS_k & \sum SS_k \\ \sum x_k^2 SS_k & \sum x_k^4 & \sum x_k^3 & \sum x_k^2 \\ \sum x_k SS_k & \sum x_k^3 & \sum x_k^2 & \sum x_k \\ \sum SS_k & \sum x_k^2 & \sum x_k & n \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k SS_k \\ \sum y_k x_k^2 \\ \sum y_k x_k \\ \sum y_k \end{pmatrix}$$

- Calcul de ω_1, a_1, b_1, c_1 :

$$\omega_1 = \sqrt{-A_1} \quad ; \quad a_1 = \frac{2 B_1}{\omega_1^2}$$

$$b_1 = \left(B_1 x_1^2 + C_1 x_1 + D_1 - a_1 \right) \sin(\omega_1 x_1) + \frac{1}{\omega_1} (C_1 + 2 B_1 x_1) \cos(\omega_1 x_1)$$

$$c_1 = \left(B_1 x_1^2 + C_1 x_1 + D_1 - a_1 \right) \cos(\omega_1 x_1) - \frac{1}{\omega_1} (C_1 + 2 B_1 x_1) \sin(\omega_1 x_1)$$

Deuxième partie :

$$\begin{cases} a_2 = a_1 \\ \rho_2 = \rho_1 = \sqrt{b_1^2 + c_1^2} \end{cases} \begin{cases} b_1 > 0 \rightarrow \varphi_1 = \arctg\left(\frac{c_1}{b_1}\right) \\ b_1 < 0 \rightarrow \varphi_1 = \arctg\left(\frac{c_1}{b_1}\right) + \pi \end{cases} \begin{cases} b_1 = 0 \\ c_1 > 0 \rightarrow \varphi_1 = \pi / 2 \\ c_1 < 0 \rightarrow \varphi_1 = -\pi / 2 \end{cases}$$

- Calcul de $K_1, K_2, \dots, K_k, \dots, K_n$:
$$K_k = \text{round}\left(\frac{\omega_1 x_k + \varphi_1}{\pi}\right)$$

[La fonction round consiste à arrondir un réel au plus proche entier]

- Calcul de $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n$:

$$\begin{cases} \text{si } \rho_2^2 > (y_k - a_2)^2 \rightarrow \theta_k = (-1)^{K_k} \arctg\left(\frac{y_k - a_2}{\sqrt{\rho_2^2 - (y_k - a_2)^2}}\right) + \pi K_k \\ \text{si } \rho_2^2 \leq (y_k - a_2)^2 \rightarrow \begin{cases} \text{si } y_k > a_2 \rightarrow \theta_k = \frac{\pi}{2}(-1)^{K_k} + \pi K_k \\ \text{si } y_k < a_2 \rightarrow \theta_k = -\frac{\pi}{2}(-1)^{K_k} + \pi K_k \end{cases} \end{cases}$$

- Calcul des sommes :
$$\sum_{k=1}^n x_k \quad ; \quad \sum_{k=1}^n x_k^2 \quad ; \quad \sum_{k=1}^n \theta_k \quad ; \quad \sum_{k=1}^n \theta_k x_k$$

- Résolution du système donnant ω_2 et φ_2 :

$$\begin{pmatrix} \omega_2 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} \sum (x_k)^2 & \sum x_k \\ \sum x_k & n \end{pmatrix}^{-1} \begin{pmatrix} \sum \theta_k x_k \\ \sum \theta_k \end{pmatrix}$$

- Calcul de b_2 et c_2 :
$$b_2 = \rho_2 \cos(\varphi_2) \quad ; \quad c_2 = \rho_2 \sin(\varphi_2)$$

Troisième partie : avec $\omega_3 = \omega_2$

- Calcul des sommes :

$$\begin{aligned} & \sum_{k=1}^n \sin(\omega_3 x_k) \quad ; \quad \sum_{k=1}^n \cos(\omega_3 x_k) \quad ; \quad \sum_{k=1}^n \sin^2(\omega_3 x_k) \quad ; \quad \sum_{k=1}^n \cos^2(\omega_3 x_k) \\ & \sum_{k=1}^n \sin(\omega_3 x_k) \cos(\omega_3 x_k) \quad ; \quad \sum_{k=1}^n y_k \quad ; \quad \sum_{k=1}^n y_k \sin(\omega_3 x_k) \quad ; \quad \sum_{k=1}^n y_k \cos(\omega_3 x_k) \end{aligned}$$

- Résolution du système donnant a_3, b_3, c_3 :

$$\begin{pmatrix} a_3 \\ b_3 \\ c_3 \end{pmatrix} = \begin{pmatrix} n & \sum \sin(\omega_3 x_k) & \sum \cos(\omega_3 x_k) \\ \sum \sin(\omega_3 x_k) & \sum \sin^2(\omega_3 x_k) & \sum \sin(\omega_3 x_k) \cos(\omega_3 x_k) \\ \sum \cos(\omega_3 x_k) & \sum \sin(\omega_3 x_k) \cos(\omega_3 x_k) & \sum \cos^2(\omega_3 x_k) \end{pmatrix}^{-1} \begin{pmatrix} \sum y_k \\ \sum y_k \sin(\omega_3 x_k) \\ \sum y_k \cos(\omega_3 x_k) \end{pmatrix}$$

Résultat : ω_3, a_3, b_3, c_3 pour l'équation optimisée : $y = a_3 + b_3 \sin(\omega_3 x) + c_3 \cos(\omega_3 x)$

Application to the logistic distribution (three parameters)

Jean Jacquelin

The logistic laws of probability $P(x)$ and distribution $D(x)$ are :

$$P(x) = \frac{\exp\left(-\frac{x-\mu}{\sigma}\right)}{\sigma\left(1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right)} ; D(x) = \frac{1}{1 + \exp\left(-\frac{x-\mu}{\sigma}\right)} \quad \{\mu > 0 ; \sigma > 0\}$$

In practice, the experimental data is related to a three parameters law :

$$y(x) = y_{\max} D(x) = \frac{\lambda}{1 + \exp\left(-\frac{x-\mu}{\sigma}\right)} \quad \text{with } \lambda = y_{\max}$$

Search for an adapted integral equation :

$$\int y(x) dx = \int \frac{\lambda}{1 + \exp\left(-\frac{x-\mu}{\sigma}\right)} dx = \lambda \sigma \ln\left(1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right) + \lambda(x-\mu)$$

$$\ln\left(1 + \exp\left(-\frac{x-\mu}{\sigma}\right)\right) = \ln\left(\frac{\lambda}{y}\right) \rightarrow \int y(x) dx = \lambda \sigma (\ln(\lambda) - \ln(y)) + \lambda x - \lambda \mu$$

$$\int_0^x y(x) dx = (\lambda \sigma (\ln(\lambda) - \ln(y(x))) + \lambda x - \lambda \mu) - (\lambda \sigma (\ln(\lambda) - \ln(y_{(x=0)})) - \lambda \mu)$$

$$\int_0^x y(x) dx = -\lambda \sigma \ln(y(x)) + \lambda x + \lambda \sigma \ln(y_{(x=0)}) \quad \text{where } y_{(x=0)} = \frac{\lambda}{1 + \exp\left(\frac{\mu}{\sigma}\right)}$$

$$\int_0^x y(x) dx = -\lambda \sigma \ln(y(x)) + \lambda x + \lambda \sigma \ln(\lambda) - \lambda \sigma \ln\left(1 + \exp\left(\frac{\mu}{\sigma}\right)\right)$$

$$\ln(y) = -\frac{1}{\lambda \sigma} \int_0^x y(x) dx + \frac{\lambda}{\lambda \sigma} x + \ln(\lambda) - \ln\left(1 + \exp\left(\frac{\mu}{\sigma}\right)\right)$$

But two difficulties arise in numerical computation with this integral equation.

First : If some values of y are close to 0, or even worse if $y=0$, then $\ln(y)$ is big or infinite. In order to avoid a failure of the computation process, we will use :

$$y \ln(y) = -\frac{1}{\lambda \sigma} y \int_0^x y(x) dx + \frac{\lambda}{\lambda \sigma} x y + \left(\ln(\lambda) - \ln\left(1 + \exp\left(\frac{\mu}{\sigma}\right)\right) \right) y$$

because $y \ln(y)$ tends to 0 when y tends to 0. So, we can set $0 \ln(0) = 0$.

Second : The numerical integration process of $y(x)$ from $x=0$ to $x=x_n$ suppose to know $y(0)$. If x_1 (i.e. the lower known value of x) is not 0 or not close to 0, the starting value S_1 cannot be set to 0 and requires to be approximated, which is a cause of deviation. In fact, it is more convenient to translate all the points (x,y) so that the first one be located at $x=0$.

The point (x_k, y_k) is replaced by (X_k, y_k) where $X_k = x_k - x_1$. Of course, in the equation $(x-\mu)$ is replaced by $((x-x_1) - (\mu-x_1)) = (X - (\mu-x_1))$

The linear relationship is : $Y = A F_1(X) + B F_2(X) + C F_3(X)$ where :

$$\begin{cases} Y = y \ln(y) & ; & F_1 = y \int_0^X y dx & ; & F_2 = X y & ; & F_3 = y \\ A = -\frac{1}{\lambda \sigma} & ; & B = \frac{1}{\sigma} & ; & C = \ln(\lambda) - \ln\left(1 + \exp\left(\frac{\mu - x_1}{\sigma}\right)\right) \end{cases}$$

a linear regression leads to the approximations of A, B, C and then, to those of λ, σ, μ

Algorithm :

Starting from Data : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

Rank Data from lower x to higher x

Compute: $X_k = x_k - x_1$ from $k = 1$ to $k = n$

$$\text{Compute } S_k : \begin{cases} S_1 = 0 \\ S_k = \frac{1}{2}(y_k + y_{k-1})(X_k - X_{k-1}) \end{cases}$$

Linear regression : Compute the components of the matrix M and vector V :

$$M_{1,1} = \sum_{k=1}^n (y_k S_k)^2 \quad ; \quad M_{2,2} = \sum_{k=1}^n (X_k y_k)^2 \quad ; \quad M_{1,2} = \sum_{k=1}^n (y_k)^2$$

$$M_{1,2} = \sum_{k=1}^n (y_k S_k)(X_k y_k) \quad ; \quad M_{1,3} = \sum_{k=1}^n (y_k S_k)(y_k) \quad ; \quad M_{2,3} = \sum_{k=1}^n (X_k y_k)(y_k)$$

$$V_1 = \sum_{k=1}^n (S_k)(y_k \ln(y_k)) \quad ; \quad V_2 = \sum_{k=1}^n (X_k y_k)(y_k \ln(y_k)) \quad ; \quad V_3 = \sum_{k=1}^n (y_k)(y_k \ln(y_k))$$

$$[M] = \begin{bmatrix} M_{1,1} & M_{1,2} & M_{1,3} \\ M_{1,2} & M_{2,2} & M_{2,3} \\ M_{1,3} & M_{2,3} & M_{3,3} \end{bmatrix} \quad [V] = \begin{bmatrix} V_1 \\ V_2 \\ V_3 \end{bmatrix}$$

$$\text{Compute } A, B, C : \begin{bmatrix} A \\ B \\ C \end{bmatrix} = [M]^{-1} [V]$$

$$\text{Compute } s, L, m : \quad s = \frac{1}{B} \quad ; \quad L = -\frac{B}{A} \quad ; \quad m = x_1 + \frac{1}{B} \ln\left(-\frac{B}{A} e^{-C} - 1\right)$$

s, L, m are the approximations of σ, λ, μ

$$\text{Theoretical : } y(x) = \frac{\lambda}{1 + \exp\left(-\frac{x - \mu}{\sigma}\right)} \quad ; \quad \text{Computed : } y(x) = \frac{L}{1 + \exp\left(-\frac{x - m}{s}\right)}$$

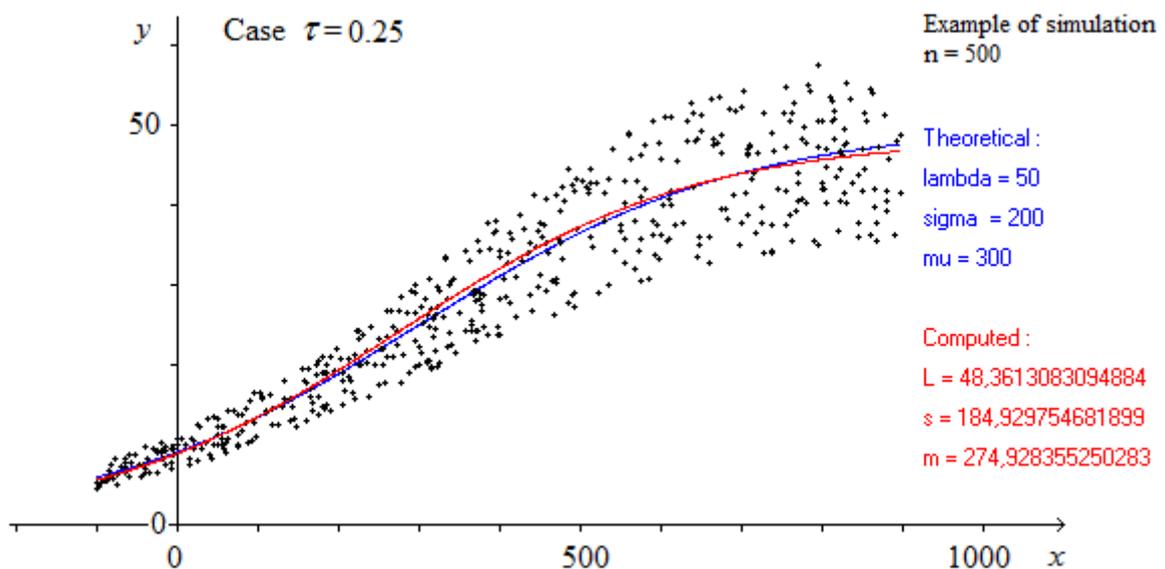
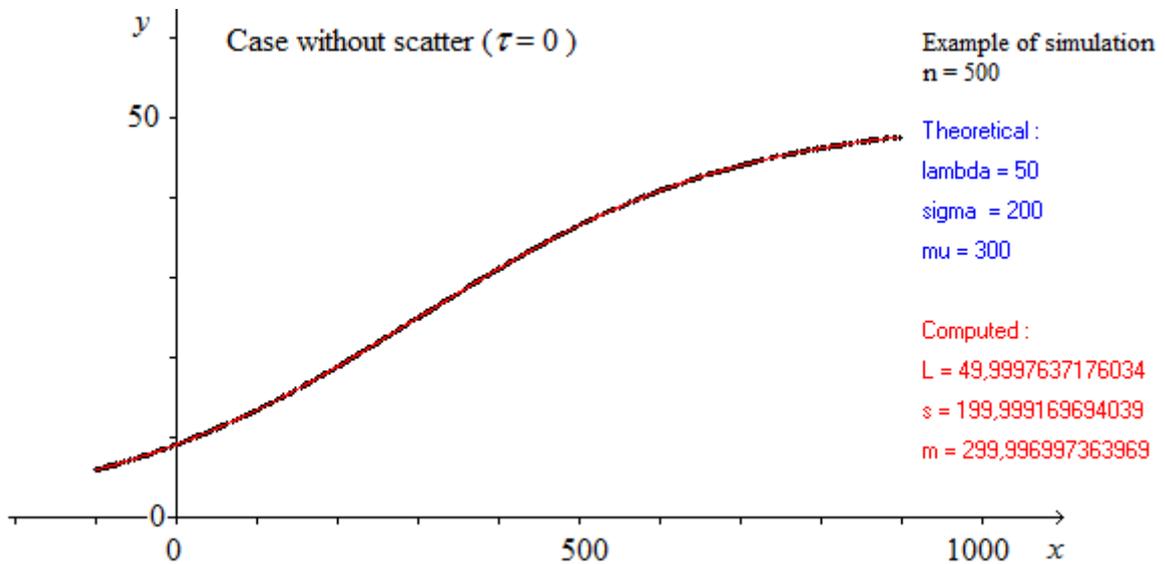
Examples showing the effect of scatter

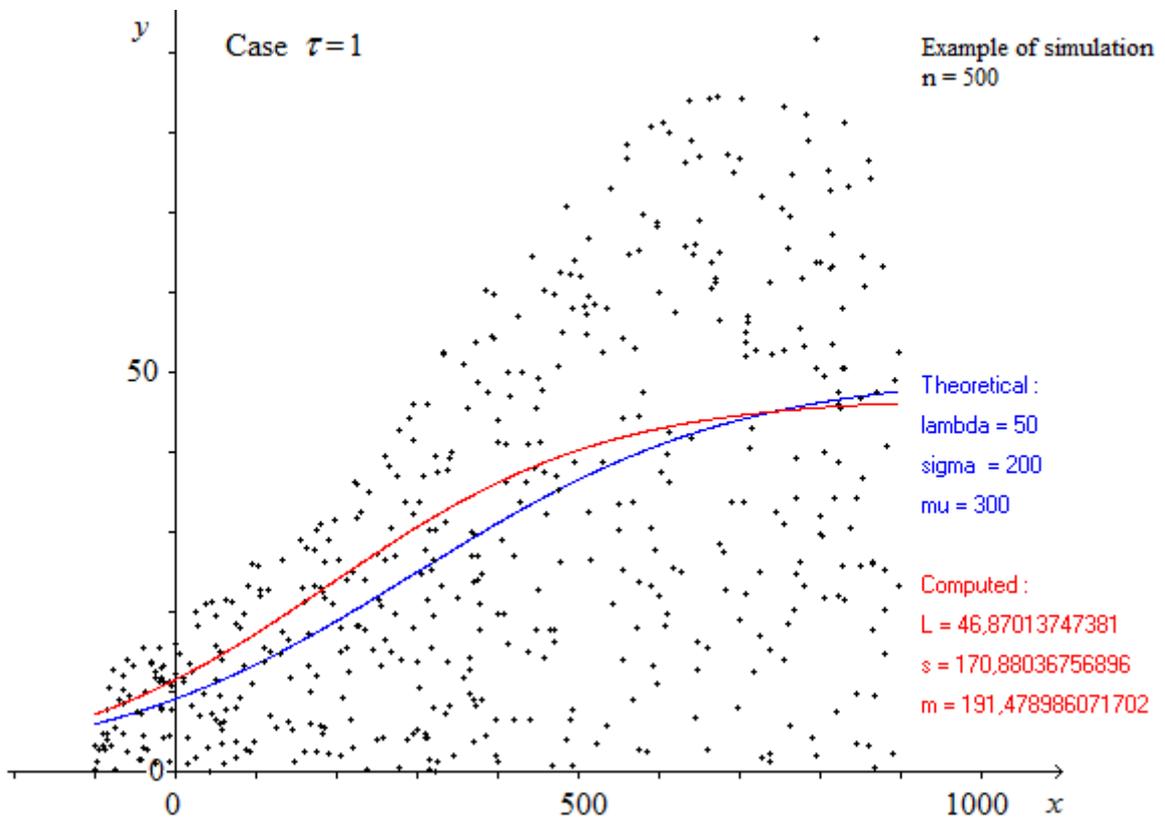
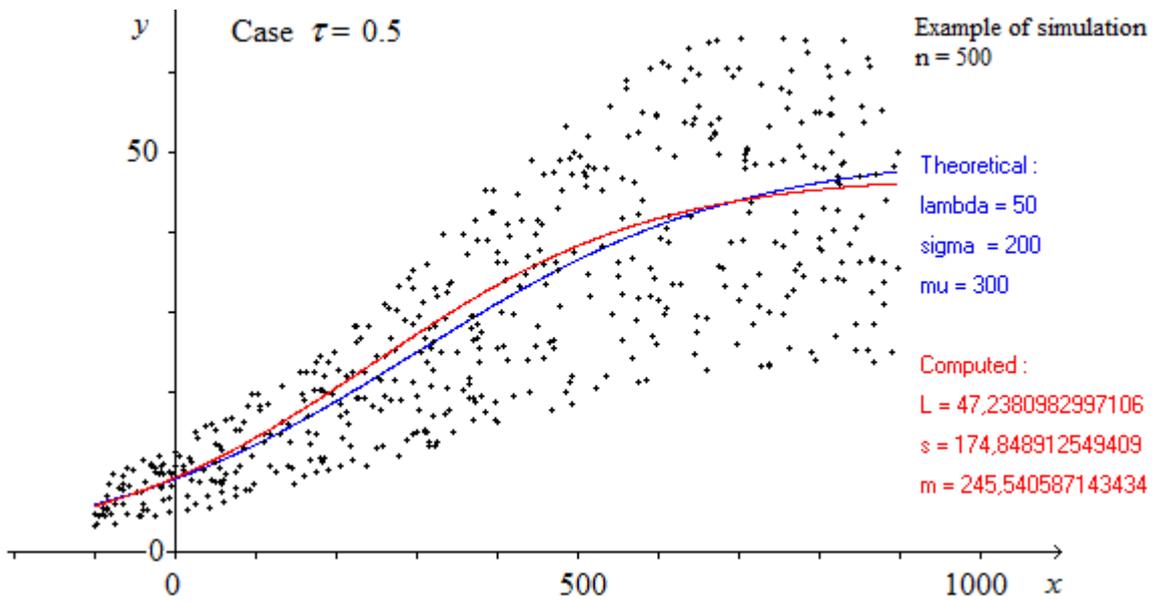
The simulated data is generated with this function :

$$\begin{cases} x_k = \text{random}\{x_{\min} \text{ to } x_{\max}\} \\ y_k = \frac{\lambda}{1 + \exp\left(-\frac{(x_k - \mu)}{\sigma}\right)} (1 + \tau \text{ random}\{-1 \text{ to } +1\}) \end{cases}$$

For example : $x_{\min} = -100$ and $x_{\max} = 800$. The parameter τ defines the range of scatter.

Blue curves : $y(x) = \frac{\lambda}{1 + \exp\left(-\frac{(x - \mu)}{\sigma}\right)}$ Red curves : $y(x) = \frac{L}{1 + \exp\left(-\frac{(x - m)}{s}\right)}$





Remark : In case of a so large scatter , the red curve has to be considered as a purely mathematical outcome, probably with no meaning on the statistical viewpoint.

Application à la distribution logistique à quatre paramètres

Application to the logistic distribution (Four parameters)

$$y(x) = \gamma + \frac{\lambda}{1 + \exp\left(-\frac{x - \mu}{\sigma}\right)}$$

Dans le cas à trois paramètres (paragraphe précédent), on a vu que le calcul de $\int y(x)dx$ conduit à une équation intégrale linéaire contenant le terme $\ln(y)$. Le même calcul, dans le cas à quatre paramètres, conduit à une équation intégrale contenant $\ln(y - \gamma)$. L'équation intégrale n'est pas linéaire relativement au paramètre γ . Il en est de même avec $\int (y(x))^2 dx$

En combinant les deux équations intégrales qui contiennent chacune $\ln(y - \gamma)$, on peut éliminer ce terme et obtenir une équation intégrale linéaire.

Au lieu de ce calcul un peu lourd, une méthode plus élégante conduit au même résultat :

$$\frac{dy}{dx} = \frac{\lambda}{\sigma} \frac{\exp\left(-\frac{x - \mu}{\sigma}\right)}{\left(1 + \exp\left(-\frac{x - \mu}{\sigma}\right)\right)^2} = \frac{\lambda}{\sigma} \frac{\left(\frac{\lambda}{y - \gamma} - 1\right)}{\left(\frac{\lambda}{y - \gamma}\right)^2} = \frac{1}{\lambda\sigma} (\lambda + \gamma - y)(y - \gamma)$$

$$\frac{dy}{dx} = -\frac{1}{\lambda\sigma} y^2 + \frac{\lambda + 2\gamma}{\lambda\sigma} y - \frac{(\lambda + \gamma)\gamma}{\lambda\sigma}$$

Cette équation différentielle pourrait être utilisée pour la régression linéaire. Mais il est en général préférable d'utiliser une équation intégrale. Son intégration conduit à :

$$y(x) = -\frac{1}{\lambda\sigma} \int_{x_1}^x (y(x))^2 dx + \frac{(\lambda + 2\gamma)}{\lambda\sigma} \int_{x_1}^x y(x) dx - \frac{(\lambda + \gamma)\gamma}{\lambda\sigma} (x - x_1) + y(x_1)$$

$$y(x_1) = \gamma + \frac{\lambda}{1 + \exp\left(-\frac{x_1 - \mu}{\sigma}\right)} \quad \text{ne doit pas être confondu avec l'ordonnée du point } (x_1, y_1)$$

Il peut paraître surprenant que, d'abord dériver, puis intégrer, conduise à une expression différente de la fonction de départ. Il ne faut pas mal interpréter cette constatation car, entre temps, la fonction a été transformée en une équation, ce qui n'est pas la même chose.

Nous disposons maintenant d'une équation intégrale linéaire :

$$y = AF_1 + BF_2 + CF_3 + DF_4 \quad \begin{cases} F_1(x) = \int_{x_1}^x (y(x))^2 dx \\ F_2(x) = \int_{x_1}^x y(x) dx \\ F_3(x) = x - x_1 \\ F_4(x) = 1 \end{cases} \quad \begin{cases} A = -\frac{1}{\lambda\sigma} \\ B = \frac{(\lambda + 2\gamma)}{\lambda\sigma} \\ C = -\frac{(\lambda + \gamma)\gamma}{\lambda\sigma} \\ D = y(x_1) \end{cases}$$

La régression linéaire donne A , B , C et D , d'où les valeurs approchées des paramètres :

$$\left\{ \begin{array}{l} \lambda + 2\gamma = -\frac{B}{A} \\ (\lambda + \gamma)\gamma = -\frac{C}{A} \\ \sigma = -\frac{1}{\lambda A} \\ y_{(x_1)} = D \end{array} \right. \rightarrow \left\{ \begin{array}{l} \lambda = -\frac{1}{A}\sqrt{B^2 + 4AC} \\ \gamma = \frac{1}{2A}\left(-B + \sqrt{B^2 + 4AC}\right) \\ \sigma = \frac{1}{\sqrt{B^2 + 4AC}} \\ \mu = x_1 + \sigma \ln\left(\frac{\lambda}{y_{(x_1)} - \gamma} - 1\right) = x_1 + \sigma \ln\left(\frac{\lambda}{D - \gamma} - 1\right) \end{array} \right. \quad \left[A < 0 \ ; \ C > -\frac{B^2}{4A} \right]$$

Algorithme :

Données : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

Réordonner les données en x croissants (minimum : x_1 , maximum : x_n)

$$\text{Calculer } S1_k : \left\{ \begin{array}{l} S1_1 = 0 \\ S1_k = S1_k + \frac{1}{2}(y_k + y_{k-1})(x_k - x_{k-1}) \end{array} \right.$$

$$\text{Calculer } S2_k : \left\{ \begin{array}{l} S2_1 = 0 \\ S2_k = S2_k + \frac{1}{2}(y_k^2 + y_{k-1}^2)(x_k - x_{k-1}) \end{array} \right.$$

Régression linéaire : Calculer les composants de la matrice M et du vecteur V :

$$[M] = \begin{bmatrix} \sum_{k=1}^n (S2_k)^2 & \sum_{k=1}^n S2_k S1_k & \sum_{k=1}^n S2_k (x_k - x_1) & \sum_{k=1}^n S2_k \\ \sum_{k=1}^n S2_k S1_k & \sum_{k=1}^n (S1_k)^2 & \sum_{k=1}^n S1_k (x_k - x_1) & \sum_{k=1}^n S1_k \\ \sum_{k=1}^n S2_k (x_k - x_1) & \sum_{k=1}^n S1_k (x_k - x_1) & \sum_{k=1}^n (x_k - x_1)^2 & \sum_{k=1}^n (x_k - x_1) \\ \sum_{k=1}^n S2_k & \sum_{k=1}^n S1_k & \sum_{k=1}^n (x_k - x_1) & n \end{bmatrix}$$

$$[V] = \begin{bmatrix} \sum_{k=1}^n S2_k y_k \\ \sum_{k=1}^n S1_k y_k \\ \sum_{k=1}^n (x_k - x_1) y_k \\ \sum_{k=1}^n y_k \end{bmatrix} \quad \text{Calculer } A, B, C, D : \quad \begin{bmatrix} A \\ B \\ C \\ D \end{bmatrix} = [M]^{-1} [V]$$

Calculer L, g, s, m qui sont les approximations de $\lambda, \gamma, \sigma, \mu$

$$\begin{cases} L = -\frac{1}{A} \sqrt{B^2 + 4AC} \\ g = -\frac{1}{2} \left(\frac{B}{A} + L \right) \\ s = -\frac{1}{AL} \\ m = x_1 + s \ln \left(\frac{L}{D - g} - 1 \right) \end{cases}$$

Résultat : La fonction calculée est : $y(x) = g + \frac{L}{1 + \exp\left(-\frac{x-m}{s}\right)}$ qui est une

approximation de la fonction théorique : $y(x) = \gamma + \frac{\lambda}{1 + \exp\left(-\frac{x-\mu}{\sigma}\right)}$

La régression fonction à quatre paramètres est moins robuste que celle à trois paramètres relativement à la dispersion des données. Si nécessaire, pour améliorer le résultat, il est suggéré d'utiliser la valeur obtenue g (approximation de γ) pour calculer de nouvelles données : (x_k, Y_k) avec $Y_k = y_k - g$ et d'effectuer la régression à trois paramètres, ce qui conduira à L, s, m comme nouvelles approximations de λ, σ, μ , complétées par l'approximation g de γ déjà obtenue.

Une illustration très sommaire en est faite par l'exemple suivant.

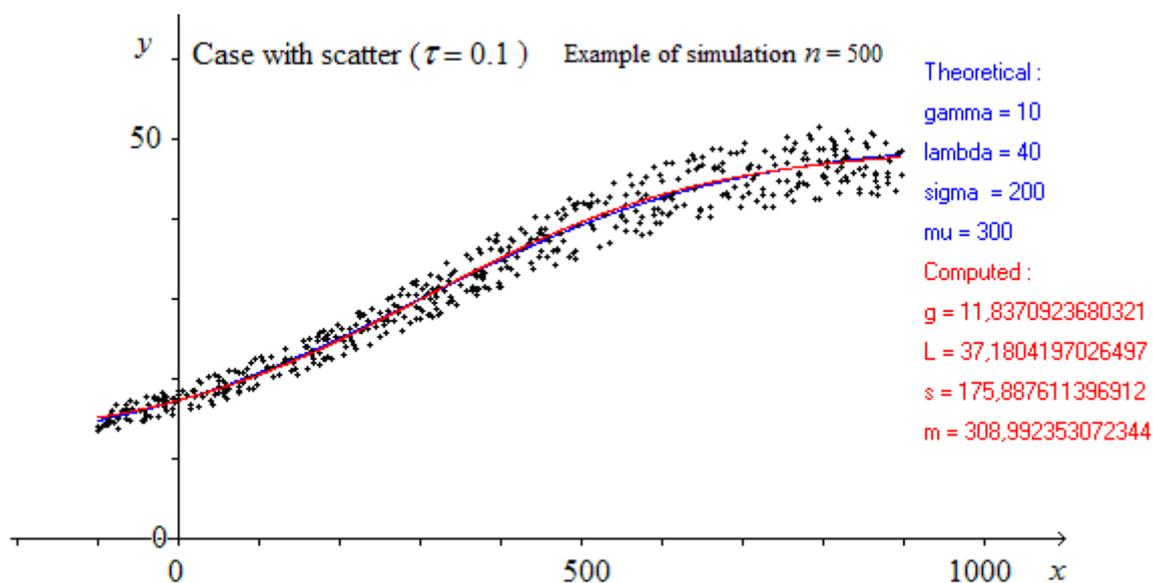
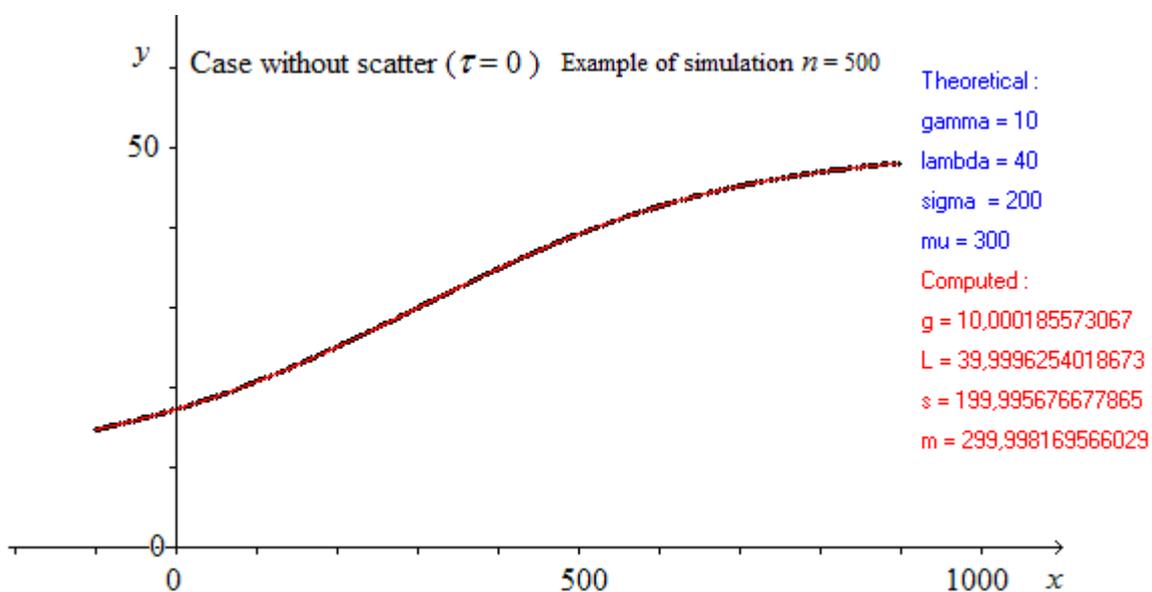
Exemple montrant l'effet de la dispersion des points donnés.

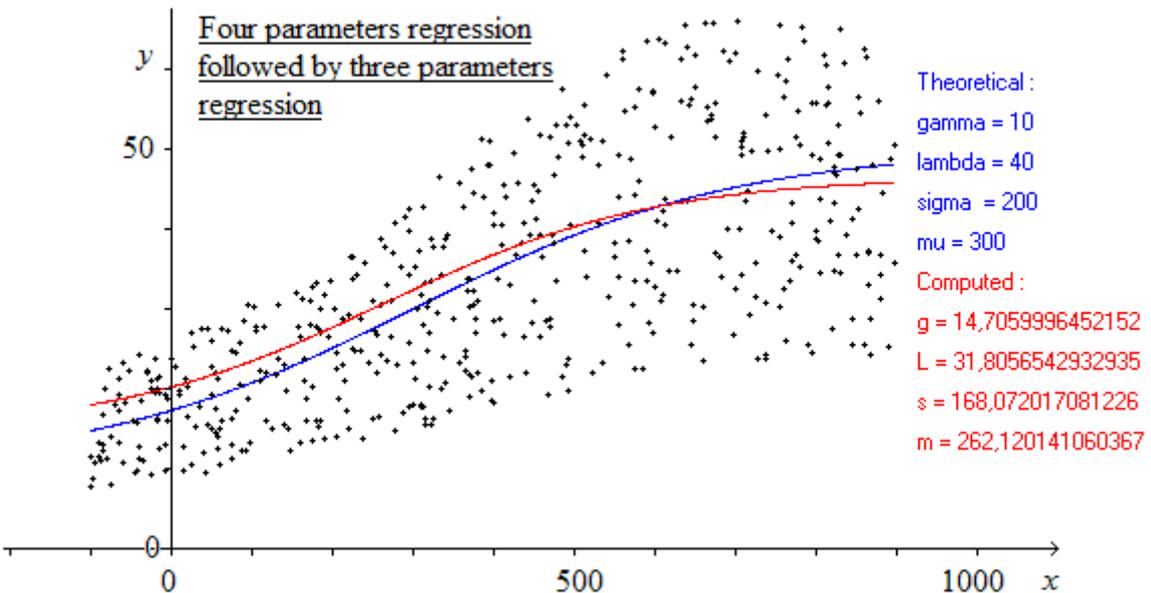
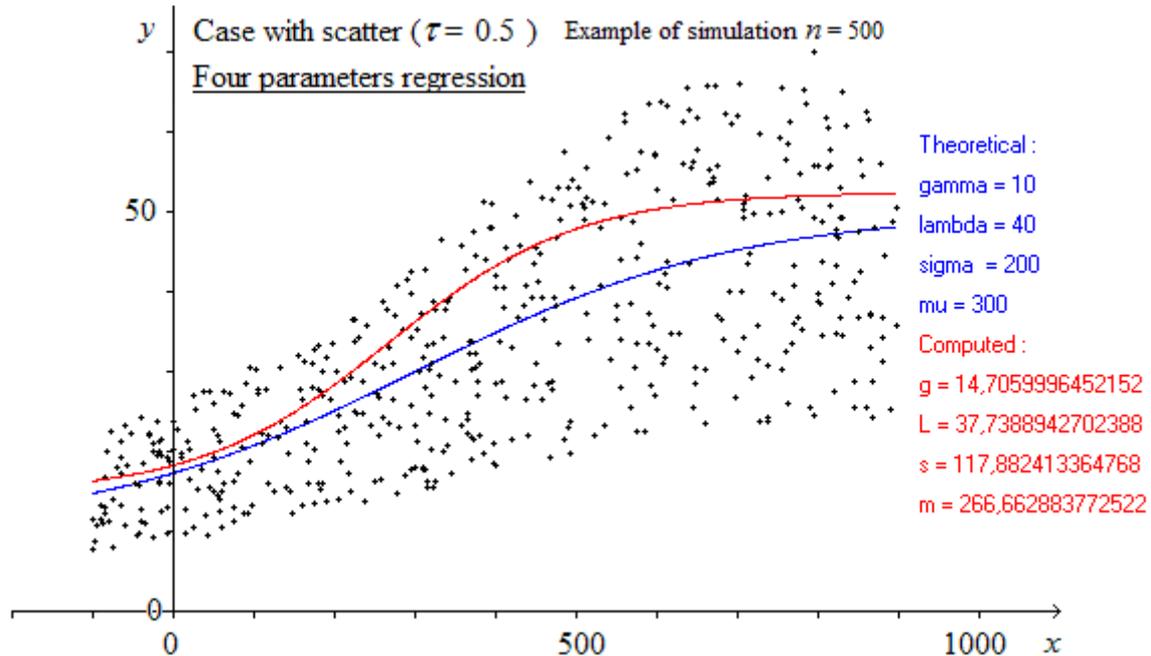
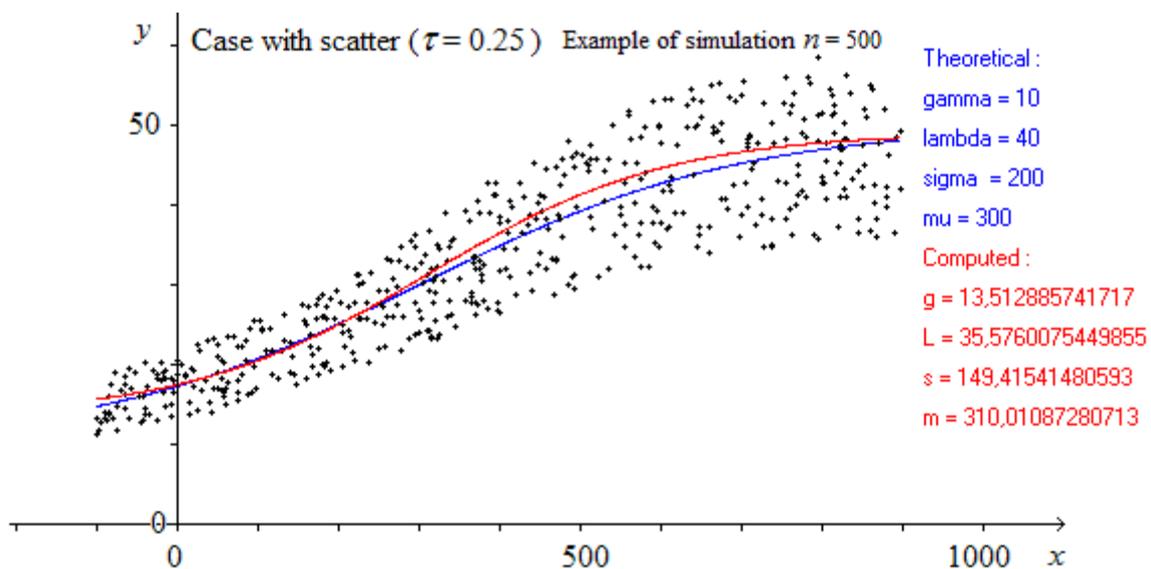
La simulation est faite avec la fonction suivante :

$$\begin{cases} x_k = \text{random}\{x_{\min} \text{ to } x_{\max}\} \\ y_k = \left(\gamma + \frac{\lambda}{1 + \exp\left(-\frac{(x_k - \mu)}{\sigma}\right)} \right) (1 + \tau \text{ random}\{-1 \text{ to } +1\}) \end{cases}$$

Par exemple : $x_{\min} = -100$ and $x_{\max} = 900$. Le paramètre τ défines la largeur de la dispersion.

Courbe bleue : $y(x) = \gamma + \frac{\lambda}{1 + \exp\left(-\frac{(x - \mu)}{\sigma}\right)}$ Courbe rouge : $y(x) = g + \frac{L}{1 + \exp\left(-\frac{(x - m)}{s}\right)}$

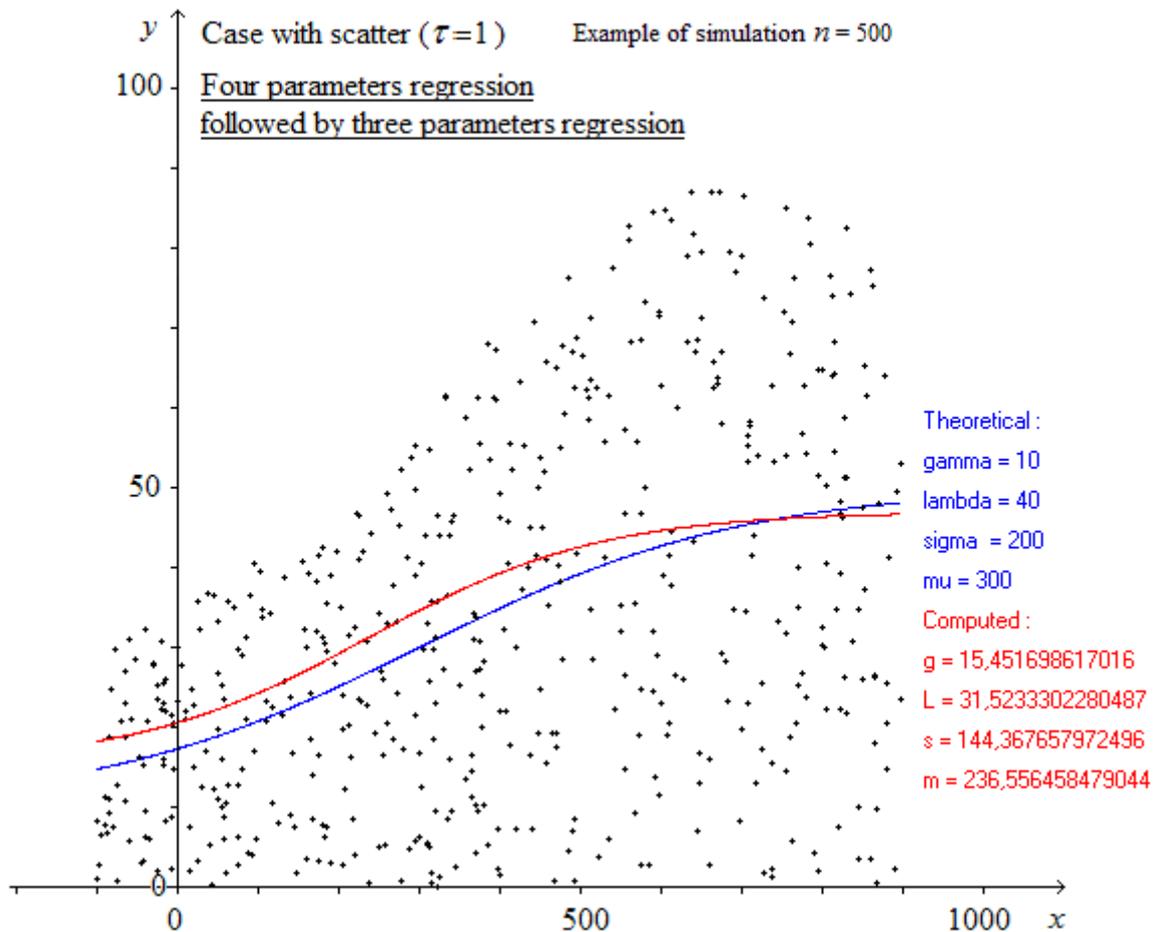




La figure précédente montre l'amélioration obtenue en effectuant une régression à trois paramètres après la régression à quatre paramètres.

La régression à quatre paramètres, selon l'algorithme donné dans le présent chapitre, produit une valeur g approximative de γ . Cette valeur sert à calculer de nouvelles données : (x_k, Y_k) avec $Y_k = y_k - g$. Finalement, la régression à trois paramètres appliquée à ces nouvelles données, selon l'algorithme donné au chapitre précédent, donne L, s, m approximations de λ, σ, μ .

Exemple avec une excessivement large dispersion :



Remarque : Dans le cas de dispersion aussi large, la courbe rouge devrait être considérée comme un résultat purement mathématique, sans vraie signification du point de vue statistique.

MIXED LINEAR AND SINUSOIDAL REGRESSION

Jean Jacquelin

In a preceding chapter, named "REGRESSION SINUSOIDALE", a process is described in order to fit the function $y(x) = a + b \sin(\omega x) + c \cos(\omega x)$ to a data set :

$$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$$

The process will be extended to fit the function

$$y(x) = a + p x + b \sin(\omega x) + c \cos(\omega x)$$

The parameters to be optimized are : ω, a, b, c, p

The problem is the same as the fitting of the functions :

$$y(x) = a + p x + \rho \sin(\omega x + \varphi)$$

where the parameters to be optimized are : $\omega, \varphi, \rho, a, p$

$$\text{because } \rho \sin(\omega x + \varphi) = b \sin(\omega x) + c \cos(\omega x) \quad \begin{cases} b = \rho \cos(\varphi) \\ c = \rho \sin(\varphi) \end{cases}$$

Integral equation :

Two successive integrations of $y(x)$ leads to :

$$\int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv = \frac{a}{2} x^2 + \frac{p}{6} x^3 - \frac{1}{\omega^2} \rho \sin(\omega x + \varphi) + C x + D$$

C and D are constants which depend on the lower bound of the integrals.

Of course, lower bounds other than x_1 could be chosen, but this would introduce more complicated terms and more complicated numerical integrations.

The elimination of $\rho \sin(\omega x + \varphi)$ by the combination of the equations of $y(x)$ and the double integral leads to the integral equation :

$$y(x) = -\omega^2 \int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv + \omega^2 \frac{p}{6} x^3 + \omega^2 \frac{a}{2} x^2 + C x + D$$

The new constants C and D are not the same as the preceding ones. Of course, it is possible to analytically express them. The formulas are complicated (similar to eq.11 in chapter "Régression sinusoidale", but with more terms). This arduous calculus will be avoid : Instead of using C and D to compute the approximates of b and c , we will see that they can be obtained much more easily, thanks to a complementary linear regression.

The integral equation above is on the linear kind :

$$y(x) = A SS(x) + E x^3 + B x^2 + Cx + D$$

$$\text{where } \begin{cases} A = -\omega^2 & ; & E = \frac{\omega^2 p}{6} & ; & B = \omega^2 \frac{a}{2} \\ SS(x) = \int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv \end{cases}$$

The values of $SS(x_k)$ are approximated by numerical integration.

According to the integral equation, a linear regression gives A_0, E_0, B_0, C_0, D_0 which are approximates of A, E, B, C, D . (Details in next section "Short way ")

So, the approximates of ω is obtained : $\omega_1 = \sqrt{-A_0}$

Then, ω_1 is the starting value for a linear regression according to the approximate function :

$y(x) \simeq a + p x + b \sin(\omega_1 x) + c \cos(\omega_1 x)$, which leads directly to the approximates of a, p, b, c and then to the approximates of ρ and φ , namely $a_1, p_1, b_1, c_1, \rho_1, \varphi_1$. This is a short way to obtain the approximate equation of $y(x)$:

$$y(x) \simeq a_1 + p_1 x + b_1 \sin(\omega_1 x) + c_1 \cos(\omega_1 x)$$

$$\text{or } y(x) \simeq a_1 + p_1 x + \rho_1 \sin(\omega_1 x + \varphi_1)$$

The "short way" above might be not sufficient in case of small data set (n small), or in case of low number of periods of the sinusoidal part of the function, or in case of large scatter, or in case of bad distribution of the x_k on the range of x , as it is pointed out in the chapter "Régression sinusoidale".

The "**Full way**" starts with the values α_1, p_1 and ρ_1 above.

$$\text{Let } \theta = \omega x + \varphi = \arcsin \left(\frac{y(x) - a_1 - p_1 x}{\rho_1} \right)$$

The inverse functions of sin, cos or tan have an infinity of determinations. The right one must be determined. This is shown in the next section "Details of the process", where arctan is the inverse function used.

$$\text{The computation of } \theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n \text{ with } \theta_k = \arcsin \left(\frac{y_k - a_1 - p_1 x_k}{\rho_1} \right)$$

(with the correct respective determinations) allows a linear regression relatively to the function $\theta = \omega x + \varphi$. This leads to the approximate ω_2 of ω .

Then, $\omega_3 = \omega_2$ is used as starting value for a linear regression relatively to the approximate function: $y(x) \simeq a + p x + b \sin(\omega_3 x) + c \cos(\omega_3 x)$, which leads to the approximates of $a, p, b, c, \rho, \varphi$, respectively named $a_3, p_3, b_3, c_3, \rho_3, \varphi_3$. Finally, the approximate of the equation $y(x)$ is :

$$y(x) \simeq a_3 + p_3 x + b_3 \sin(\omega_3 x) + c_3 \cos(\omega_3 x)$$

$$\text{or } y(x) \simeq a_3 + p_3 x + \rho_3 \sin(\omega_3 x + \varphi_3)$$

The subscripts 1, 2, 3 for the approximates of the fitted parameters were chosen in order to be consistent with the notations used in the previous chapter "Sinusoidal Regression".

DETAILS OF THE PROCESS

$$y(x) = a + p x + b \sin(\omega x) + c \cos(\omega x)$$

$$\text{or } y(x) = a + p x + \rho \sin(\omega x + \varphi)$$

Set of data : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$
They must be written by ascending values of x .

Short way :

- Computation of $S_1, S_2, \dots, S_k, \dots, S_n$:

$$S_1 = 0 \quad ; \quad S_k = S_{k-1} + \frac{1}{2}(y_k + y_{k-1})(x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Computation of $SS_1, SS_2, \dots, SS_k, \dots, SS_n$:

$$SS_1 = 0 \quad ; \quad SS_k = SS_{k-1} + \frac{1}{2}(S_k + S_{k-1})(x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Solving of the regression system for A_0, E_0, B_0, C_0 and D_0 :

$$\begin{pmatrix} A_0 \\ E_0 \\ B_0 \\ C_0 \\ D_0 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n SS_k^2 & \sum_{k=1}^n SS_k x_k^3 & \sum_{k=1}^n SS_k x_k^2 & \sum_{k=1}^n SS_k x_k & \sum_{k=1}^n SS_k \\ \sum_{k=1}^n SS_k x_k^3 & \sum_{k=1}^n x_k^6 & \sum_{k=1}^n x_k^5 & \sum_{k=1}^n x_k^4 & \sum_{k=1}^n x_k^3 \\ \sum_{k=1}^n SS_k x_k^2 & \sum_{k=1}^n x_k^5 & \sum_{k=1}^n x_k^4 & \sum_{k=1}^n x_k^3 & \sum_{k=1}^n x_k^2 \\ \sum_{k=1}^n SS_k x_k & \sum_{k=1}^n x_k^4 & \sum_{k=1}^n x_k^3 & \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \\ \sum_{k=1}^n SS_k & \sum_{k=1}^n x_k^3 & \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k & n \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n SS_k y_k \\ \sum_{k=1}^n x_k^3 y_k \\ \sum_{k=1}^n x_k^2 y_k \\ \sum_{k=1}^n x_k y_k \\ \sum_{k=1}^n y_k \end{pmatrix}$$

- Computation of : $\omega_1 = \sqrt{-A_0}$

- Computation of $\beta_1, \dots, \beta_k, \dots, \beta_n, \eta_1, \dots, \eta_k, \dots, \eta_n$:

$$\beta_k = \sin(\omega_1 x_k) \quad ; \quad \eta_k = \cos(\omega_1 x_k)$$

- Solving of the regression system for α_1, p_1, b_1, c_1 :

$$\begin{pmatrix} a_1 \\ p_1 \\ b_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} n & \sum_{k=1}^n x_k & \sum_{k=1}^n \beta_k & \sum_{k=1}^n \eta_k \\ \sum_{k=1}^n x_k & \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \beta_k & \sum_{k=1}^n x_k \eta_k \\ \sum_{k=1}^n \beta_k & \sum_{k=1}^n x_k \beta_k & \sum_{k=1}^n \beta_k^2 & \sum_{k=1}^n \beta_k \eta_k \\ \sum_{k=1}^n \eta_k & \sum_{k=1}^n x_k \eta_k & \sum_{k=1}^n \beta_k \eta_k & \sum_{k=1}^n \eta_k^2 \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n y_k \\ \sum_{k=1}^n x_k y_k \\ \sum_{k=1}^n \beta_k y_k \\ \sum_{k=1}^n \eta_k y_k \end{pmatrix}$$

$$\text{- Computation of : } \rho_1 = \sqrt{b_1^2 + c_1^2} \text{ and of } \varphi_1 : \begin{cases} \text{if } b_1 > 0 \rightarrow \varphi_1 = \arctan\left(\frac{c_1}{b_1}\right) \\ \text{if } b_1 < 0 \rightarrow \varphi_1 = \arctan\left(\frac{c_1}{b_1}\right) + \pi \end{cases}$$

Full way (continuation of the process above) :

$$\text{- Computation of } K_1, K_2, \dots, K_k, \dots, K_n : \quad K_k = \text{round}\left(\frac{\omega_1 x_k + \varphi_1}{\pi}\right)$$

[The real argument is rounded to the nearest integer]

- Computation of $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n$:

$$\begin{cases} r_k = y_k - a_1 - p_1 x_k \\ \text{if } \rho_1^2 > r_k^2 \rightarrow \theta_k = (-1)^{K_k} \text{atan}\left(\frac{r_k}{\sqrt{\rho_1^2 - r_k^2}}\right) + \pi K_k \\ \text{if } \rho_1^2 \leq r_k^2 \rightarrow \begin{cases} \text{if } r_k > 0 \rightarrow \theta_k = \frac{\pi}{2} (-1)^{K_k} + \pi K_k \\ \text{if } r_k < 0 \rightarrow \theta_k = -\frac{\pi}{2} (-1)^{K_k} + \pi K_k \end{cases} \end{cases}$$

- Solving of the regresson system for ω_2 and φ_2 :

$$\begin{pmatrix} \omega_2 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \\ \sum_{k=1}^n x_k & n \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \theta_k x_k \\ \sum_{k=1}^n \theta_k \end{pmatrix}$$

$$\text{- Computation of } b_2 \text{ and } c_2 : \quad b_2 = \rho_2 \cos(\varphi_2) \quad ; \quad c_2 = \rho_2 \sin(\varphi_2)$$

$$\text{- with } \omega_3 = \omega_2 \text{ computation of : } \beta_k = \sin(\omega_3 x_k) \quad ; \quad \eta_k = \cos(\omega_3 x_k)$$

- Solving of the regression system for a_3, p_3, b_3, c_3 :

$$\begin{pmatrix} a_3 \\ p_3 \\ b_3 \\ c_3 \end{pmatrix} = \begin{pmatrix} n & \sum_{k=1}^n x_k & \sum_{k=1}^n \beta_k & \sum_{k=1}^n \eta_k \\ \sum_{k=1}^n x_k & \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \beta_k & \sum_{k=1}^n x_k \eta_k \\ \sum_{k=1}^n \beta_k & \sum_{k=1}^n x_k \beta_k & \sum_{k=1}^n \beta_k^2 & \sum_{k=1}^n \beta_k \eta_k \\ \sum_{k=1}^n \eta_k & \sum_{k=1}^n x_k \eta_k & \sum_{k=1}^n \beta_k \eta_k & \sum_{k=1}^n \eta_k^2 \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n y_k \\ \sum_{k=1}^n x_k y_k \\ \sum_{k=1}^n \beta_k y_k \\ \sum_{k=1}^n \eta_k y_k \end{pmatrix}$$

Computation of ρ_3 and φ_3 : $\rho_3 = \sqrt{b_3^2 + c_3^2}$; $\begin{cases} \text{if } b_3 > 0 \rightarrow \varphi_3 = \arctan\left(\frac{c_3}{b_3}\right) \\ \text{if } b_3 < 0 \rightarrow \varphi_3 = \arctan\left(\frac{c_3}{b_3}\right) + \pi \end{cases}$

Result : the fitted function is $\begin{cases} y(x) \approx a_3 + p_3 x + b_3 \sin(\omega_3 x) + c_3 \cos(\omega_3 x) \\ \text{or : } y(x) \approx a_3 + p_3 x + \rho_3 \sin(\omega_3 x + \varphi_3) \end{cases}$

EXAMPLE

Generation of the data set : ($n = 20$)

n real numbers are randomly taken on the range x_{min} to x_{max} ($x_{min} = 0$; $x_{max} = 10$ for example).. They are rounded to two decimal places. They are recorded by ascending values as the data set ($x_1, \dots, x_k, \dots, x_n$ on the next table).

"Theretical" values of parameters $a, p, b, c, \rho, \varphi$ are chosen, for example : $\omega=0.8, a=0.4, p=0.6, b=1.2, c=0.75$

"Theretical" values of $y(x_k) = a + p x_k + \rho \sin(\omega x_k + \varphi)$ are computed. They are scattered by adding random real numbers taken on the range $-\epsilon$ to $+\epsilon$ (for example $\epsilon = 0.5$). They are rounded to two decimal places (next table). The points (x_k, y_k) are drawn on the figure at the end of the current section. On the same graph, the dotted curve represents the "theoretical" function $y(x) = a + p x + \rho \sin(\omega x + \varphi)$

Computation in details :

Short way :

| k | x_k | y_k | S_k | SS_k | k | x_k | y_k | S_k | SS_k |
|-----|-------|-------|-------------|-------------|-----|-------|-------|-------------|-------------|
| 1 | 0 | 1.35 | 0.00000E+00 | 0.00000E+00 | 11 | 3.28 | 2.29 | 7.54390E+00 | 7.54390E+00 |
| 2 | 0.31 | 1.95 | 5.11500E-01 | 5.11500E-01 | 12 | 3.68 | 1.76 | 8.35390E+00 | 8.35390E+00 |
| 3 | 0.6 | 2.20 | 1.11325E+00 | 1.11325E+00 | 13 | 3.72 | 2.46 | 8.43830E+00 | 8.43830E+00 |
| 4 | 0.71 | 1.91 | 1.33930E+00 | 1.33930E+00 | 14 | 4.26 | 1.70 | 9.56150E+00 | 9.56150E+00 |
| 5 | 0.82 | 1.88 | 1.54775E+00 | 1.54775E+00 | 15 | 4.75 | 2.20 | 1.05170E+01 | 1.05170E+01 |
| 6 | 1.62 | 2.56 | 3.32375E+00 | 3.32375E+00 | 16 | 6.72 | 4.43 | 1.70476E+01 | 1.70476E+01 |
| 7 | 2.03 | 2.74 | 4.41025E+00 | 4.41025E+00 | 17 | 7.75 | 5.69 | 2.22594E+01 | 2.22594E+01 |
| 8 | 2.73 | 2.33 | 6.18475E+00 | 6.18475E+00 | 18 | 8.41 | 7.03 | 2.64570E+01 | 2.64570E+01 |
| 9 | 2.93 | 2.82 | 6.69975E+00 | 6.69975E+00 | 19 | 8.61 | 7.19 | 2.78790E+01 | 2.78790E+01 |
| 10 | 3.19 | 2.14 | 7.34455E+00 | 7.34455E+00 | 20 | 9.17 | 6.84 | 3.18074E+01 | 3.18074E+01 |

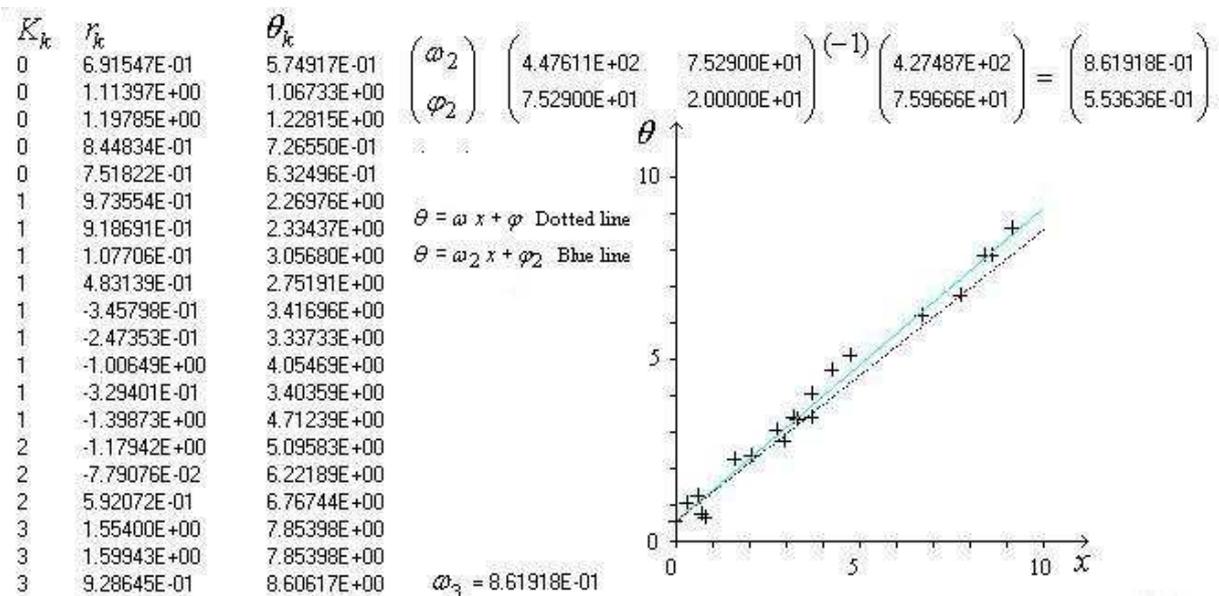
$$\begin{pmatrix} A_0 \\ E_0 \\ B_0 \\ C_0 \\ D_0 \end{pmatrix} = \begin{pmatrix} 3.86789E+04 & 2.54125E+05 & 3.09341E+04 & 3.92503E+03 & 5.39832E+02 \\ 2.54125E+05 & 1.69060E+06 & 2.02222E+05 & 2.48023E+04 & 3.18596E+03 \\ 3.09341E+04 & 2.02222E+05 & 2.48023E+04 & 3.18596E+03 & 4.47611E+02 \\ 3.92503E+03 & 2.48023E+04 & 3.18596E+03 & 4.47611E+02 & 7.52900E+01 \\ 5.39832E+02 & 3.18596E+03 & 4.47611E+02 & 7.52900E+01 & 2.00000E+01 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 2.96162E+03 \\ 1.89243E+04 \\ 2.39477E+03 \\ 3.34431E+02 \\ 6.34700E+01 \end{pmatrix} = \begin{pmatrix} -8.95247E-01 \\ 8.37812E-02 \\ 3.09791E-01 \\ 1.53450E+00 \\ 1.28151E+00 \end{pmatrix}$$

$$\omega_1 = 9.46175E-01$$

$$\begin{pmatrix} a_1 \\ p_1 \\ b_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} 2.00000E+01 & 7.52900E+01 & 6.25927E+00 & -2.12798E+00 \\ 7.52900E+01 & 4.47611E+02 & 2.75029E+01 & -2.03310E+01 \\ 6.25927E+00 & 2.75029E+01 & 8.49012E+00 & 1.47437E+00 \\ -2.12798E+00 & -2.03310E+01 & 1.47437E+00 & 1.15099E+01 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 6.34700E+01 \\ 3.34431E+02 \\ 3.08218E+01 \\ -8.00493E+00 \end{pmatrix} = \begin{pmatrix} 6.58453E-01 \\ 5.72836E-01 \\ 1.24076E+00 \\ 2.79172E-01 \end{pmatrix}$$

$$\rho_1 = 1.27178E+00 \quad \varphi_1 = 2.21316E-01$$

Full way :



Drawing θ as a function of x is very usefull to check if there is no mistake in the determinations of arctan.

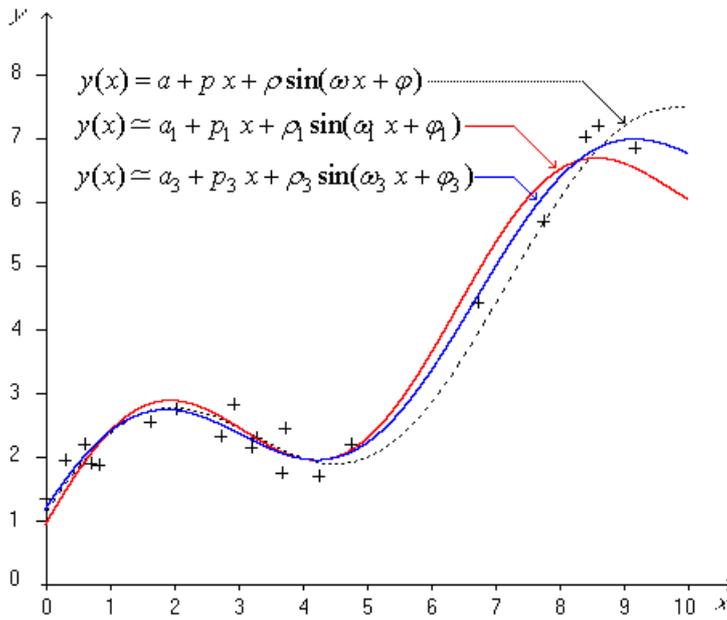
$$\begin{pmatrix} a_3 \\ p_3 \\ b_3 \\ c_3 \end{pmatrix} = \begin{pmatrix} 2.00000E+01 & 7.52900E+01 & 7.15797E+00 & 3.15374E-01 \\ 7.52900E+01 & 4.47611E+02 & 2.81140E+01 & -1.24206E+00 \\ 7.15797E+00 & 2.81140E+01 & 7.87726E+00 & 1.76828E+00 \\ 3.15374E-01 & -1.24206E+00 & 1.76828E+00 & 1.21227E+01 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 6.34700E+01 \\ 3.34431E+02 \\ 3.03306E+01 \\ 8.86804E+00 \end{pmatrix} = \begin{pmatrix} 5.71315E-01 \\ 5.82827E-01 \\ 1.11331E+00 \\ 6.13981E-01 \end{pmatrix}$$

$$\rho_3 = 1.27139E+00 \quad \varphi_3 = 1.06681E+00$$

The results are summarized on the next figure where the short and full ways are compared.

Note : This first example is caricatural because a very large scatter was applied in order to make clear the differences between the curves and the points. Generally the distribution of the points is better and the scatter less important. A more realistic example is shown below, on the second figure :

Example with large scatter (range ± 0.5)

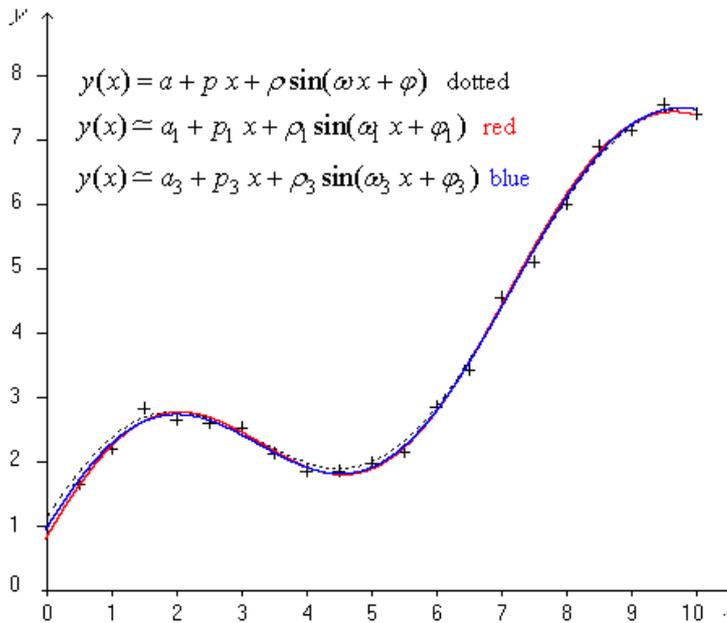


| Theoretical | Computed | Short way | Full way |
|----------------|----------|-----------|----------|
| w = 0.8 | w = | 0.946175 | 0.861918 |
| a = 0.4 | a = | 0.658453 | 0.571315 |
| p = 0.6 | p = | 0.572836 | 0.582827 |
| b = 1.2 | b = | 1.240756 | 1.113310 |
| c = 0.75 | c = | 0.279172 | 0.613981 |
| ro = 1.415097 | ro = | 1.271775 | 1.271389 |
| phi = 0.558599 | phi = | 0.221316 | 0.503988 |

| k | x(k) | y(k) | Fitted values y(k) | |
|----|------|------|--------------------|----------|
| 1 | 0 | 1.35 | 0.937625 | 1.185296 |
| 2 | 0.31 | 1.95 | 1.462016 | 1.638129 |
| 3 | 0.6 | 2.20 | 1.904687 | 2.005129 |
| 4 | 0.71 | 1.91 | 2.055904 | 2.127247 |
| 5 | 0.82 | 1.88 | 2.196398 | 2.239106 |
| 6 | 1.62 | 2.56 | 2.836911 | 2.718489 |
| 7 | 2.03 | 2.74 | 2.891156 | 2.740740 |
| 8 | 2.73 | 2.33 | 2.643079 | 2.519366 |
| 9 | 2.93 | 2.82 | 2.524377 | 2.421329 |
| 10 | 3.19 | 2.14 | 2.361336 | 2.288543 |
| 11 | 3.28 | 2.29 | 2.305694 | 2.243514 |
| 12 | 3.68 | 1.76 | 2.089164 | 2.068727 |
| 13 | 3.72 | 2.46 | 2.071834 | 2.054707 |
| 14 | 4.26 | 1.70 | 1.959362 | 1.961478 |
| 15 | 4.75 | 2.20 | 2.107653 | 2.076647 |
| 16 | 6.72 | 4.43 | 4.879398 | 4.504302 |
| 17 | 7.75 | 5.69 | 6.312973 | 6.084664 |
| 18 | 8.41 | 7.03 | 6.681334 | 6.737764 |
| 19 | 8.61 | 7.19 | 6.698072 | 6.857628 |
| 20 | 9.17 | 6.84 | 6.551027 | 6.997195 |

Mean square Deviation = 0.341 0.256

Example with good distribution and small scatter (range ± 0.2)



| Theoretical | Computed | Short way | Full way |
|----------------|----------|-----------|----------|
| w = 0.8 | w = | 0.827014 | 0.809132 |
| a = 0.4 | a = | 0.249471 | 0.266603 |
| p = 0.6 | p = | 0.613961 | 0.613464 |
| b = 1.2 | b = | 1.347289 | 1.278916 |
| c = 0.75 | c = | 0.560551 | 0.680126 |
| ro = 1.415097 | ro = | 1.459249 | 1.448515 |
| phi = 0.558599 | phi = | 0.394273 | 0.488762 |

| k | x(k) | y(k) | Fitted values y(k) | |
|----|------|------|--------------------|----------|
| 1 | 0.5 | 1.66 | 1.611130 | 1.701963 |
| 2 | 1 | 2.20 | 2.234455 | 2.274977 |
| 3 | 1.5 | 2.83 | 2.626674 | 2.622777 |
| 4 | 2 | 2.66 | 2.773417 | 2.738735 |
| 5 | 2.5 | 2.59 | 2.701695 | 2.653650 |
| 6 | 3 | 2.53 | 2.475345 | 2.430783 |
| 7 | 3.5 | 2.12 | 2.184267 | 2.155639 |
| 8 | 4 | 1.85 | 1.929273 | 1.922165 |
| 9 | 4.5 | 1.85 | 1.805093 | 1.817578 |
| 10 | 5 | 1.97 | 1.884405 | 1.908287 |
| 11 | 5.5 | 2.16 | 2.205587 | 2.229169 |
| 12 | 6 | 2.86 | 2.766244 | 2.777941 |
| 13 | 6.5 | 3.42 | 3.523615 | 3.515523 |
| 14 | 7 | 4.56 | 4.401781 | 4.372354 |
| 15 | 7.5 | 5.11 | 5.304458 | 5.259619 |
| 16 | 8 | 6.00 | 6.131234 | 6.083588 |
| 17 | 8.5 | 6.91 | 6.794488 | 6.760753 |
| 18 | 9 | 7.16 | 7.234166 | 7.231304 |
| 19 | 9.5 | 7.56 | 7.427898 | 7.468794 |
| 20 | 10 | 7.41 | 7.394775 | 7.484402 |

Mean square Deviation = 0.108 0.102

GENERALIZED SINUSOIDAL REGRESSION

Jean Jacquelin

In the preceeding chapter " **Mixed linear and sinusoidal regression**", a process is described in order to fit the function $y(x) = a + p x + b \sin(\omega x) + c \cos(\omega x)$ to a data set :

$$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$$

The process will be extended to fit the function

$$y(x) = b \sin(\omega x) + c \cos(\omega x) + \lambda_1 f_1(x) + \dots + \lambda_j f_j(x) + \dots + \lambda_m f_m(x)$$

The parameters to be optimized are : $\omega, b, c, \lambda_1, \dots, \lambda_j, \dots, \lambda_m$

$f_1(x), f_2(x), \dots, f_j(x), \dots, f_m(x)$ are different known functions. They doesn't contain a parameter belonging to the list of parameters to be optimized.

The problem is the same as the fitting of the functions :

$$y(x) = \rho \sin(\omega x + \varphi) + \lambda_1 f_1(x) + \dots + \lambda_j f_j(x) + \dots + \lambda_m f_m(x)$$

where the parameters to be optimized are : $\omega, \varphi, \rho, \lambda_1, \dots, \lambda_j, \dots, \lambda_m$

$$\text{because } \rho \sin(\omega x + \varphi) = b \sin(\omega x) + c \cos(\omega x) \quad \begin{cases} b = \rho \cos(\varphi) \\ c = \rho \sin(\varphi) \end{cases}$$

In fact, the preceeding chapter "Mixed linear and sinusoidal regression" deals with the particular case : $f_1(x) = 1, f_2(x) = x; \lambda_1 = a, \lambda_2 = p$

Integral equation :

Two successive integrations of $y(x)$ leads to :

$$SS(x) = \int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv = -\frac{1}{\omega^2} \rho \sin(\omega x + \varphi) + C x + D + \sum_{j=1}^m \lambda_j FF_j(x)$$

C and D are constants which depend on the lower bound of the integrals.

$$FF_j(x) = \int_{x_1}^x \left(\int_{x_1}^v f_j(u) du \right) dv \quad \text{The values corresponding to } x_k \text{ are computed by}$$

analytical or numerical integration.

$$FF_{j,k} = FF_j(x_k) = \int_{x_1}^x \left(\int_{x_1}^v f_j(u) du \right) dv \quad [1 \leq j \leq m]$$

Of course, $SS_k \approx SS(x_k)$ cannot be computed by analytical integration since $y(x)$ is not known. But approximate values SS_k are computed by numerical integration from the data set (x_k, y_k) .

Lower bounds other than x_1 could be chosen, but this would introduce more complicated terms and more complicated numerical integrations. The linear part of the integrals are merged in the term $Cx+D$.

The elimination of $\rho \sin(\omega x + \varphi)$ by the combination of the equations of $y(x)$ and the double integral leads to the integral equation :

$$y(x) = -\omega^2 SS(x) + C\omega^2 x + D\omega^2 + \sum_{j=1}^m \lambda_j \omega^2 FF_j(x) + \sum_{j=1}^m \lambda_j f_j(x)$$

The new constants C and D are not the same as the preceding ones. The arduous calculus of the analytic formulas will be avoid : Instead of using C and D for further developments, we will see that it is much more easy to do a complementary linear regression.

The integral equation above is on the linear kind relatively to the parameters :

$A, C, D, \mu_j, \lambda_j$

$$y(x) = A SS(x) + Cx + D + \sum_{j=1}^m \mu_j FF_j(x) + \sum_{j=1}^m \lambda_j f_j(x)$$

where $A = -\omega^2$

Eventually, among the functions involved, somme could be identical. Then, the corresponding terms must be merged. For example, if $f_1(x)=1/x^2$ and $f_2(x)=\ln(x)$ then $FF_1(x)=-\ln(x)+c_1x+c_2$. We see that the same function $\ln(x)$ appears in $\lambda_1 \omega^2 FF_1(x)$ and in $\lambda_2 f_2(x)$. So, the two terms must be merged. Also c_1x+c_2 is merged with $Cx+D$.

A linear regression gives A_0 which is the approximate of A (Details in next section "Short way")

So, the approximates of ω is obtained : $\omega_1 = \sqrt{-A_0}$

Then, ω_1 is the starting value for a linear regression according to the approximate function :

$$y(x) \simeq b \sin(\omega_1 x) + c \cos(\omega_1 x) + \lambda_1 f_1(x) + \dots + \lambda_j f_j(x) + \dots + \lambda_m f_m(x),$$

which leads directly to the approximates of $b, c, \lambda_1, \dots, \lambda_l, \dots, \lambda_m$ and then to the approximates of ρ and φ , namely $b_1, c_1, \lambda_{1,1}, \dots, \lambda_{l,1}, \dots, \lambda_{m,1}$ and ρ_1, φ_1

This is a short way to obtain the approximate equation of $y(x)$:

$$y(x) \simeq b_1 \sin(\omega_1 x) + c_1 \cos(\omega_1 x) + \lambda_{1,1} f_1(x) + \dots + \lambda_{j,1} f_j(x) + \dots + \lambda_{m,1} f_m(x)$$

$$\text{or } y(x) \simeq \rho_1 \sin(\omega_1 x + \varphi_1) + \lambda_{1,1} f_1(x) + \dots + \lambda_{j,1} f_j(x) + \dots + \lambda_{m,1} f_m(x)$$

The "**short way**" above might be not sufficient in case of small data set (n small), or in case of low number of periods of the sinusoidal part of the function, or in case of large scatter, or in case of bad distribution of the x_k on the range of x , as it is pointed out in the chapter "Régression sinusoidale".

The "**Full way**" starts with the values ρ_1 and $\lambda_{j,1}$ above.

$$\text{Let } \theta = \omega x + \varphi = \arcsin \left(\frac{1}{\rho_1} \left(y(x) - \sum_{j=1}^m \lambda_{j,1} f_j(x) \right) \right)$$

The inverse functions of sin, cos or tan have an infinity of determinations. The right one must be determined. This is shown in the next section "Details of the process", where arctan is the inverse function used.

The computation of $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n$ with

$$\theta_k = \arcsin \left(\frac{1}{\rho_1} \left(y(x_k) - \sum_{j=1}^m \lambda_{j,1} f_j(x_k) \right) \right)$$

(with the correct respective determinations) allows a linear regression relatively to the function $\theta = \omega x + \varphi$. This leads to the approximate ω_2 of ω .

Then, $\omega_3 = \omega_2$ is used as starting value for a linear regression according to the approximate function: $y(x) \simeq b \sin(\omega_3 x) + c \cos(\omega_3 x) + \lambda_1 f_1(x) + \dots + \lambda_j f_j(x) + \dots + \lambda_m f_m(x)$,

which leads to the approximates of $b, c, \lambda_1, \dots, \lambda_l, \dots, \lambda_m, \rho, \varphi$, namely $b_3, c_3, \lambda_{1,3}, \dots, \lambda_{l,3}, \dots, \lambda_{m,3}, \rho_3, \varphi_3$. Finally, the approximate of the equation $y(x)$ is :

$$y(x) \simeq b_3 \sin(\omega_3 x) + c_3 \cos(\omega_3 x) + \lambda_{1,3} f_1(x) + \dots + \lambda_{j,3} f_j(x) + \dots + \lambda_{m,3} f_m(x)$$

$$\text{or } y(x) \simeq \rho_3 \sin(\omega_3 x + \varphi_3) + \lambda_{1,3} f_1(x) + \dots + \lambda_{j,3} f_j(x) + \dots + \lambda_{m,3} f_m(x)$$

The subscripts 1, 2, 3 for the approximates of the fitted parameters where chosen in order to be consistent with the notations used in the previous chapters "Sinusoidal Regression" and "Mixed linear and sinusoidal regression".

DETAILS OF THE PROCESS

$$y(x) = b \sin(\omega x) + c \cos(\omega x) + \lambda_1 f_1(x) + \dots + \lambda_j f_j(x) + \dots + \lambda_m f_m(x)$$

$$\text{or } y(x) = \rho \sin(\omega x + \varphi) + \lambda_1 f_1(x) + \dots + \lambda_j f_j(x) + \dots + \lambda_m f_m(x)$$

Set of data : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

They must be written by ascending values of x .

Short way :

- Computation of $S_1, S_2, \dots, S_k, \dots, S_n$:

$$S_1 = 0 \quad ; \quad S_k = S_{k-1} + \frac{1}{2} (y_k + y_{k-1}) (x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Computation of $SS_1, SS_2, \dots, SS_k, \dots, SS_n$:

$$SS_1 = 0 \quad ; \quad SS_k = SS_{k-1} + \frac{1}{2} (S_k + S_{k-1}) (x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Computation of $f_{j,k} = f_j(x_k)$ for $j=1$ to m and $k=1$ to n

The computation of $FF_{j,k} = \int_{x_1}^x \left(\int_{x_1}^v f_j(u) du \right) dv \quad [1 \leq j \leq m \quad ; \quad 1 \leq k \leq n]$

is preferably done by analytical integration of the known functions $f_j(x)$.

Alternatively, approximates can be obtained by numerical integrations in this way :

Computation of $F_{j,1}, F_{j,2}, \dots, F_{j,k}, \dots, F_{j,n}$ for $j=1$ to m

$$F_{j,1} = 0 \quad ; \quad F_{j,k} = F_{j,k-1} + \frac{1}{2} (f_{j,k} + f_{j,k-1}) (x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

Computation of $FF_{j,1}, FF_{j,2}, \dots, FF_{j,k}, \dots, FF_{j,n}$ for $j=1$ to m :

$$FF_{j,1} = 0 \quad ; \quad FF_{j,k} = FF_{j,k-1} + \frac{1}{2}(F_{j,k} + F_{j,k-1})(x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

The preceding numerical methods is useful if the analytical integration of an integral is too complicated, for example if some special functions are involved.

- Solving of the regression system for A_0, C_0, D_0 , and (V_{00}) :

$$\begin{pmatrix} A_0 \\ C_0 \\ D_0 \\ (V_{00})^T \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n SS_k^2 & \sum_{k=1}^n SS_k & \sum_{k=1}^n SS_k x_k & (V_{14}) & (V_{15}) \\ \sum_{k=1}^n SS_k & n & \sum_{k=1}^n x_k & (V_{24}) & (V_{25}) \\ \sum_{k=1}^n SS_k x_k & \sum_{k=1}^n x_k & \sum_{k=1}^n x_k^2 & (V_{34}) & (V_{35}) \\ (V_{14})^T & (V_{24})^T & (V_{34})^T & (V_{44}) & (V_{45}) \\ (V_{15})^T & (V_{25})^T & (V_{35})^T & (V_{45})^T & (V_{55}) \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n SS_k y_k \\ \sum_{k=1}^n y_k \\ \sum_{k=1}^n x_k y_k \\ (V_{04})^T \\ (V_{05})^T \end{pmatrix}$$

The superscript T indicates that the row of m elements is transposed to a column.

In interest of space, the big matrix is written on a condensed manner, thanks to the following notations :

$$(V_{00}) = (\mu_1, \dots, \mu_j, \dots, \mu_m, \dots, \lambda_1, \dots, \lambda_j, \dots, \lambda_m)$$

$$(V_{04}) = \left(\sum_{k=1}^n y_k FF_{1,k}, \dots, \sum_{k=1}^n y_k FF_{j,k}, \dots, \sum_{k=1}^n y_k FF_{m,k} \right)$$

$$(V_{05}) = \left(\sum_{k=1}^n y_k f_{1,k}, \dots, \sum_{k=1}^n y_k f_{j,k}, \dots, \sum_{k=1}^n y_k f_{m,k} \right)$$

$$(V_{14}) = \left(\sum_{k=1}^n SS_k FF_{1,k}, \dots, \sum_{k=1}^n SS_k FF_{j,k}, \dots, \sum_{k=1}^n SS_k FF_{m,k} \right)$$

$$(V_{15}) = \left(\sum_{k=1}^n SS_k f_{1,k}, \dots, \sum_{k=1}^n SS_k f_{j,k}, \dots, \sum_{k=1}^n SS_k f_{m,k} \right)$$

$$(V_{24}) = \left(\sum_{k=1}^n FF_{1,k}, \dots, \sum_{k=1}^n FF_{j,k}, \dots, \sum_{k=1}^n FF_{m,k} \right)$$

$$(V_{25}) = \left(\sum_{k=1}^n f_{1,k}, \dots, \sum_{k=1}^n f_{j,k}, \dots, \sum_{k=1}^n f_{m,k} \right)$$

$$(V_{34}) = \left(\sum_{k=1}^n x_k FF_{1,k}, \dots, \sum_{k=1}^n x_k FF_{j,k}, \dots, \sum_{k=1}^n x_k FF_{m,k} \right)$$

$$(V_{35}) = \left(\sum_{k=1}^n x_k f_{1,k}, \dots, \sum_{k=1}^n x_k f_{j,k}, \dots, \sum_{k=1}^n x_k f_{m,k} \right)$$

$$\begin{aligned}
(V_{44}) &= \begin{pmatrix} \sum_{k=1}^n (FF_{1,k})^2 & \dots & \sum_{k=1}^n FF_{1,k} FF_{j,k} & \dots & \sum_{k=1}^n FF_{1,k} FF_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n FF_{j,k} FF_{1,k} & \dots & \sum_{k=1}^n (FF_{j,k})^2 & \dots & \sum_{k=1}^n FF_{j,k} FF_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n FF_{m,k} FF_{1,k} & \dots & \sum_{k=1}^n FF_{m,k} FF_{j,k} & \dots & \sum_{k=1}^n (FF_{m,k})^2 \end{pmatrix} \\
(V_{45}) &= \begin{pmatrix} \sum_{k=1}^n FF_{1,k} f_{1,k} & \dots & \sum_{k=1}^n FF_{1,k} f_{j,k} & \dots & \sum_{k=1}^n FF_{1,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n FF_{j,k} f_{1,k} & \dots & \sum_{k=1}^n FF_{j,k} f_{j,k} & \dots & \sum_{k=1}^n FF_{j,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n FF_{m,k} f_{1,k} & \dots & \sum_{k=1}^n FF_{m,k} f_{j,k} & \dots & \sum_{k=1}^n FF_{m,k} f_{m,k} \end{pmatrix} \\
(V_{55}) &= \begin{pmatrix} \sum_{k=1}^n (f_{1,k})^2 & \dots & \sum_{k=1}^n f_{1,k} f_{j,k} & \dots & \sum_{k=1}^n f_{1,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{j,k} f_{1,k} & \dots & \sum_{k=1}^n (f_{j,k})^2 & \dots & \sum_{k=1}^n f_{j,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{m,k} f_{1,k} & \dots & \sum_{k=1}^n f_{m,k} f_{j,k} & \dots & \sum_{k=1}^n (f_{m,k})^2 \end{pmatrix}
\end{aligned}$$

Note : The maximum size of the equations system is $3+2m$. But in some casses, among the functions involved, somme are identical. Then, the corresponding terms must be merged. This reduces the size of the matrix and vectors. Accordingly, the coefficients (C, D, μ_j, λ_j) are not the same as before. Only the first coefficient (A) becomes the same, allowing to compute the approximate of ω in all cases.

- Computation of : $\omega_1 = \sqrt{-A_0}$
- Computation of $\beta_1, \dots, \beta_k, \dots, \beta_n, \gamma_1, \dots, \gamma_k, \dots, \gamma_n$:
 $\beta_k = \sin(\omega_1 x_k) \quad ; \quad \eta_k = \cos(\omega_1 x_k)$
- Solving of the regression system for $b_1, c_1, \lambda_{j,1}$:

$$\begin{aligned}
(V_{03}) &= \left(\sum_{k=1}^n y_k f_{1,k}, \dots, \sum_{k=1}^n y_k f_{j,k}, \dots, \sum_{k=1}^n y_k f_{m,k} \right) \\
(V_{13}) &= \left(\sum_{k=1}^n \beta_k f_{1,k}, \dots, \sum_{k=1}^n \beta_k f_{j,k}, \dots, \sum_{k=1}^n \beta_k f_{m,k} \right) \\
(V_{23}) &= \left(\sum_{k=1}^n \eta_k f_{1,k}, \dots, \sum_{k=1}^n \eta_k f_{j,k}, \dots, \sum_{k=1}^n \eta_k f_{m,k} \right) \\
(V_{33}) &= \begin{pmatrix} \sum_{k=1}^n (f_{1,k})^2 & \dots & \sum_{k=1}^n f_{1,k} f_{j,k} & \dots & \sum_{k=1}^n f_{1,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{j,k} f_{1,k} & \dots & \sum_{k=1}^n (f_{j,k})^2 & \dots & \sum_{k=1}^n f_{j,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{m,k} f_{1,k} & \dots & \sum_{k=1}^n f_{m,k} f_{j,k} & \dots & \sum_{k=1}^n (f_{m,k})^2 \end{pmatrix} \\
\begin{pmatrix} b_1 \\ c_1 \\ \lambda_{1,1} \\ \dots \\ \lambda_{j,1} \\ \dots \\ \lambda_{m,1} \end{pmatrix} &= \begin{pmatrix} \sum_{k=1}^n \beta_k^2 & \sum_{k=1}^n \beta_k \eta_k & (V_{13}) \\ \sum_{k=1}^n \eta_k \beta_k & \sum_{k=1}^n \eta_k^2 & (V_{23}) \\ (V_{13})^T & (V_{23})^T & (V_{33}) \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \beta_k y_k \\ \sum_{k=1}^n \eta_k y_k \\ (V_{03})^T \end{pmatrix}
\end{aligned}$$

- Computation of : $\rho_1 = \sqrt{b_1^2 + c_1^2}$ and of φ_1 :

$$\begin{cases} \text{if } b_1 > 0 \rightarrow \varphi_1 = \arctan\left(\frac{c_1}{b_1}\right) \\ \text{if } b_1 < 0 \rightarrow \varphi_1 = \arctan\left(\frac{c_1}{b_1}\right) + \pi \end{cases}$$

Full way (continuation of the process above)

:

- Computation of $K_1, K_2, \dots, K_k, \dots, K_n$: $K_k = \text{round}\left(\frac{\omega_1 x_k + \varphi_1}{\pi}\right)$

[The real argument is rounded to the nearest integer]

- Computation of $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n$:

$$\left\{ \begin{array}{l} r_k = y_k - \sum_{k=1}^n \lambda_k f_{j,k} \\ \text{if } \rho_1^2 > r_k^2 \rightarrow \theta_k = (-1)^{K_k} \operatorname{atan} \left(\frac{r_k}{\sqrt{\rho_1^2 - r_k^2}} \right) + \pi K_k \\ \text{if } \rho_1^2 \leq r_k^2 \rightarrow \left\{ \begin{array}{l} \text{if } r_k > 0 \rightarrow \theta_k = \frac{\pi}{2} (-1)^{K_k} + \pi K_k \\ \text{if } r_k < 0 \rightarrow \theta_k = -\frac{\pi}{2} (-1)^{K_k} + \pi K_k \end{array} \right. \end{array} \right.$$

- Solving of the regression system for ω_2 and φ_2 :

$$\begin{pmatrix} \omega_2 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \\ \sum_{k=1}^n x_k & n \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \theta_k x_k \\ \sum_{k=1}^n \theta_k \end{pmatrix}$$

- Computation of b_2 and c_2 : $b_2 = \rho_2 \cos(\varphi_2)$; $c_2 = \rho_2 \sin(\varphi_2)$

- with $\omega_3 = \omega_2$ computation of : $\beta_k = \sin(\omega_3 x_k)$; $\eta_k = \cos(\omega_3 x_k)$

- Solving of the regression system for $b_3, c_3, \lambda_{j,3}$:

$$\begin{aligned} (V_{03}) &= \left(\sum_{k=1}^n y_k f_{1,k}, \dots, \sum_{k=1}^n y_k f_{j,k}, \dots, \sum_{k=1}^n y_k f_{m,k} \right) \\ (V_{13}) &= \left(\sum_{k=1}^n \beta_k f_{1,k}, \dots, \sum_{k=1}^n \beta_k f_{j,k}, \dots, \sum_{k=1}^n \beta_k f_{m,k} \right) \\ (V_{23}) &= \left(\sum_{k=1}^n \eta_k f_{1,k}, \dots, \sum_{k=1}^n \eta_k f_{j,k}, \dots, \sum_{k=1}^n \eta_k f_{m,k} \right) \\ (V_{33}) &= \begin{pmatrix} \sum_{k=1}^n (f_{1,k})^2 & \dots & \sum_{k=1}^n f_{1,k} f_{j,k} & \dots & \sum_{k=1}^n f_{1,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{j,k} f_{1,k} & \dots & \sum_{k=1}^n (f_{j,k})^2 & \dots & \sum_{k=1}^n f_{j,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{m,k} f_{1,k} & \dots & \sum_{k=1}^n f_{m,k} f_{j,k} & \dots & \sum_{k=1}^n (f_{m,k})^2 \end{pmatrix} \end{aligned}$$

$$\begin{pmatrix} b_3 \\ c_3 \\ \lambda_{1,3} \\ \dots \\ \lambda_{j,3} \\ \dots \\ \lambda_{m,3} \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n \beta_k^2 & \sum_{k=1}^n \beta_k \eta_k & (V_{13}) \\ \sum_{k=1}^n \eta_k \beta_k & \sum_{k=1}^n \eta_k^2 & (V_{23}) \\ (V_{13})^T & (V_{23})^T & (V_{33}) \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \beta_k y_k \\ \sum_{k=1}^n \eta_k y_k \\ (V_{03})^T \end{pmatrix}$$

- Computation of ρ_3 and φ_3 : $\rho_3 = \sqrt{b_3^2 + c_3^2}$; $\begin{cases} \text{if } b_3 > 0 \rightarrow \varphi_3 = \arctan\left(\frac{c_3}{b_3}\right) \\ \text{if } b_3 < 0 \rightarrow \varphi_3 = \arctan\left(\frac{c_3}{b_3}\right) + \pi \end{cases}$

Result : the fitted function is

$$y(x) = b_3 \sin(\omega_3 x) + c_3 \cos(\omega_3 x) + \lambda_{1,3} f_1(x) + \dots + \lambda_{j,3} f_j(x) + \dots + \lambda_{m,3} f_m(x)$$

or $y(x) = \rho_3 \sin(\omega_3 x + \varphi_3) + \lambda_{1,3} f_1(x) + \dots + \lambda_{j,3} f_j(x) + \dots + \lambda_{m,3} f_m(x)$

EXAMPLE : case of the function $y(x) = \rho \sin(\omega x + \varphi) + \lambda_1 \frac{1}{x} + \lambda_2 \ln(x)$

Generation of the data set : ($n = 20$)

For this example, the abscisses $x_1, \dots, x_k, \dots, x_n$ are linearly taken on the range x_{min} to x_{max}
 $: x_k = x_{min} + (x_{max} - x_{min}) \frac{k}{n}$ ($x_{min} = 0.1$; $x_{max} = 1.1$)

"Theoretical" values of parameters are chosen, for example :

$$\omega = 25, b = 0.2, c = -0.15, \lambda_1 = 2, \lambda_2 = 5$$

The given functions are : $f_1(x) = \frac{1}{x}$; $f_2(x) = \ln(x)$

"Theoretical" values of $y(x_k) = \rho \sin(\omega x_k + \varphi) + \lambda_1 \frac{1}{x_k} + \lambda_2 \ln(x_k)$ are computed. They are

scattered by adding random real numbers taken on the range $-\epsilon$ to $+\epsilon$ (for example $\epsilon = 0.05$). They are rounded to two decimal places, which generate $y_1, \dots, y_k, \dots, y_n$. The points (x_k, y_k) are drawn on the figure at the end of the current section. On the same graph, the dotted curve represents the "theoretical" function $y(x) = \rho \sin(\omega x + \varphi) + \lambda_1 \frac{1}{x} + \lambda_2 \ln(x)$

Analytic integration of $f_1(x)$ and $f_2(x)$ leads to:

$$\int_{x_1}^x \left(\int_{x_1}^v \frac{1}{u} du \right) dv = x \ln(x) - (1 + \ln(x_1))x + x_1 \quad \text{The non linear part is : } FF_1(x) = x \ln(x)$$

$$\int_{x_1}^x \left(\int_{x_1}^v \ln(u) du \right) dv = \frac{x^2}{4} (2 \ln(x) - 3) + (x_1 - \ln(x_1)x_1)x + \frac{x_1^2}{4} (2 \ln(x_1) - 1)$$

which non linear part is : $FF_2(x) = \frac{x^2}{4} (2 \ln(x) - 3)$

Short way :

| k | x_k | y_k | S_k | SS_k | $f_{1,k}$ | $f_{2,k}$ | $FF_{1,k}$ | $FF_{2,k}$ |
|-----|-------|-------|-------------|-------------|-------------|--------------|--------------|--------------|
| 1 | 0.15 | 3.81 | 0.00000E+00 | 0.00000E+00 | 6.66667E+00 | -1.89712E+00 | -2.84568E-01 | -4.46760E-03 |
| 2 | 0.2 | 1.67 | 1.37000E-01 | 3.42500E-03 | 5.00000E+00 | -1.60944E+00 | -3.21888E-01 | -2.18876E-03 |
| 3 | 0.25 | 0.95 | 2.02500E-01 | 1.19125E-02 | 4.00000E+00 | -1.38629E+00 | -3.46574E-01 | 3.55330E-03 |
| 4 | 0.3 | 0.75 | 2.45000E-01 | 2.31000E-02 | 3.33333E+00 | -1.20397E+00 | -3.61192E-01 | 1.33212E-02 |
| 5 | 0.35 | 0.68 | 2.80750E-01 | 3.62438E-02 | 2.85714E+00 | -1.04982E+00 | -3.67438E-01 | 2.75734E-02 |
| 6 | 0.4 | 0.45 | 3.09000E-01 | 5.09875E-02 | 2.50000E+00 | -9.16291E-01 | -3.66516E-01 | 4.66967E-02 |
| 7 | 0.45 | 0.20 | 3.25250E-01 | 6.68438E-02 | 2.22222E+00 | -7.98508E-01 | -3.59328E-01 | 7.10261E-02 |
| 8 | 0.5 | 0.34 | 3.38750E-01 | 8.34438E-02 | 2.00000E+00 | -6.93147E-01 | -3.46574E-01 | 1.00857E-01 |
| 9 | 0.55 | 0.76 | 3.66250E-01 | 1.01069E-01 | 1.81818E+00 | -5.97837E-01 | -3.28810E-01 | 1.36452E-01 |
| 10 | 0.6 | 1.02 | 4.10750E-01 | 1.20494E-01 | 1.66667E+00 | -5.10826E-01 | -3.06495E-01 | 1.78051E-01 |
| 11 | 0.65 | 0.91 | 4.59000E-01 | 1.42238E-01 | 1.53846E+00 | -4.30783E-01 | -2.80009E-01 | 2.25872E-01 |
| 12 | 0.7 | 0.84 | 5.02750E-01 | 1.66281E-01 | 1.42857E+00 | -3.56675E-01 | -2.49672E-01 | 2.80115E-01 |
| 13 | 0.75 | 1.02 | 5.49250E-01 | 1.92581E-01 | 1.33333E+00 | -2.87682E-01 | -2.15762E-01 | 3.40964E-01 |
| 14 | 0.8 | 1.54 | 6.13250E-01 | 2.21644E-01 | 1.25000E+00 | -2.23144E-01 | -1.78515E-01 | 4.08594E-01 |
| 15 | 0.85 | 1.74 | 6.95250E-01 | 2.54356E-01 | 1.17647E+00 | -1.62519E-01 | -1.38141E-01 | 4.83165E-01 |
| 16 | 0.9 | 1.71 | 7.81500E-01 | 2.91275E-01 | 1.11111E+00 | -1.05361E-01 | -9.48245E-02 | 5.64829E-01 |
| 17 | 0.95 | 1.67 | 8.66000E-01 | 3.32463E-01 | 1.05263E+00 | -5.12933E-02 | -4.87286E-02 | 6.53729E-01 |
| 18 | 1 | 1.81 | 9.53000E-01 | 3.77938E-01 | 1.00000E+00 | 0.00000E+00 | 0.00000E+00 | 7.50000E-01 |
| 19 | 1.05 | 2.29 | 1.05550E+00 | 4.28150E-01 | 9.52381E-01 | 4.87902E-02 | 5.12297E-02 | 8.53771E-01 |
| 20 | 1.1 | 2.52 | 1.17575E+00 | 4.83931E-01 | 9.09091E-01 | 9.53102E-02 | 1.04841E-01 | 9.65163E-01 |

$$\begin{pmatrix} 9.95268E-01 & 3.38838E+00 & 2.92797E+00 & -3.37431E-01 & 1.90789E+00 & 4.26149E+00 & -6.15304E-01 \\ 3.38838E+00 & 2.00000E+01 & 1.25000E+01 & -4.43896E+00 & 6.09708E+00 & 4.38163E+01 & -1.21366E+01 \\ 2.92797E+00 & 1.25000E+01 & 9.47500E+00 & -2.01835E+00 & 5.47009E+00 & 2.00000E+01 & -4.43896E+00 \\ -3.37431E-01 & -4.43896E+00 & -2.01835E+00 & 1.41902E+00 & -4.78151E-01 & -1.21366E+01 & 3.93412E+00 \\ 1.90789E+00 & 6.09708E+00 & 5.47009E+00 & -4.78151E-01 & 3.68419E+00 & 7.15552E+00 & -8.04251E-01 \\ 4.26149E+00 & 4.38163E+01 & 2.00000E+01 & -1.21366E+01 & 7.15552E+00 & 1.40199E+02 & -4.27063E+01 \\ -6.15304E-01 & -1.21366E+01 & -4.43896E+00 & 3.93412E+00 & -8.04251E-01 & -4.27063E+01 & 1.37358E+01 \end{pmatrix}^{(-1)} \begin{pmatrix} 5.51019E+00 \\ 2.66800E+01 \\ 1.78435E+01 \\ -4.45081E+00 \\ 1.04273E+01 \\ 6.51959E+01 \\ -1.61581E+01 \end{pmatrix} = \begin{pmatrix} -6.88958E+02 \\ -2.17851E+03 \\ 1.75465E+03 \\ -2.24940E+03 \\ 3.58098E+02 \\ -3.23426E+01 \\ -7.90075E+02 \end{pmatrix}$$

$A_0 = -6.88958E+02$ $\omega_1 = 2.62480E+01$

$$\begin{pmatrix} b_1 \\ c_1 \\ (\lambda_1)_1 \\ (\lambda_2)_1 \end{pmatrix} = \begin{pmatrix} 9.91835E+00 & 4.57199E-01 & -5.73921E+00 & 1.66881E+00 \\ 4.57199E-01 & 1.00816E+01 & -1.08156E+00 & -1.43637E-01 \\ -5.73921E+00 & -1.08156E+00 & 1.40199E+02 & -4.27063E+01 \\ 1.66881E+00 & -1.43637E-01 & -4.27063E+01 & 1.37358E+01 \end{pmatrix}^{(-1)} \begin{pmatrix} -2.75478E+00 \\ -5.13170E+00 \\ 6.51959E+01 \\ -1.61581E+01 \end{pmatrix} = \begin{pmatrix} 4.14675E-02 \\ -2.28845E-01 \\ 1.97204E+00 \\ 4.94753E+00 \end{pmatrix}$$

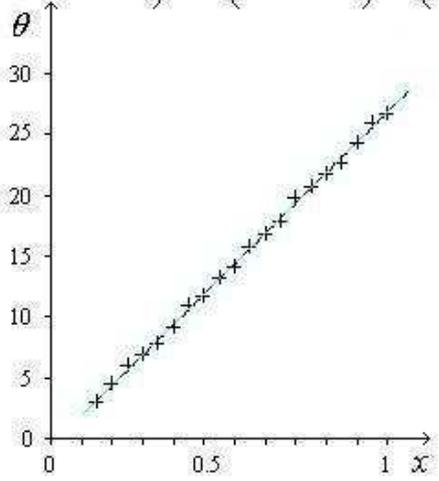
$\rho_1 = 2.32571E-01$ $\varphi_1 = -1.39154E+00$

Full way :

| K_k | r_k | θ_k |
|-------|--------------|-------------|
| 1 | 4.91390E-02 | 2.92870E+00 |
| 1 | -2.27447E-01 | 4.50207E+00 |
| 2 | -7.94180E-02 | 5.93470E+00 |
| 2 | 1.33233E-01 | 6.89319E+00 |
| 2 | 2.39633E-01 | 7.85398E+00 |
| 3 | 5.32817E-02 | 9.19363E+00 |
| 3 | -2.31665E-01 | 1.09073E+01 |
| 4 | -1.74709E-01 | 1.17165E+01 |
| 4 | 1.32293E-01 | 1.31715E+01 |
| 5 | 2.60595E-01 | 1.41372E+01 |
| 5 | 7.40699E-03 | 1.56761E+01 |
| 5 | -2.12537E-01 | 1.68606E+01 |
| 6 | -1.86068E-01 | 1.79222E+01 |
| 6 | 1.78962E-01 | 1.97276E+01 |
| 7 | 2.24022E-01 | 2.06923E+01 |
| 7 | 4.01207E-02 | 2.18178E+01 |
| 7 | -1.52055E-01 | 2.27037E+01 |
| 8 | -1.62038E-01 | 2.43619E+01 |
| 8 | 1.70477E-01 | 2.59555E+01 |
| 9 | 2.55688E-01 | 2.67035E+01 |

$$\begin{pmatrix} \omega_2 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} 9.47500E+00 & 1.25000E+01 \\ 1.25000E+01 & 2.00000E+01 \end{pmatrix}^{(-1)} \begin{pmatrix} 2.28809E+02 \\ 2.99660E+02 \end{pmatrix} = \begin{pmatrix} 2.49751E+01 \\ -6.26427E-01 \end{pmatrix}$$

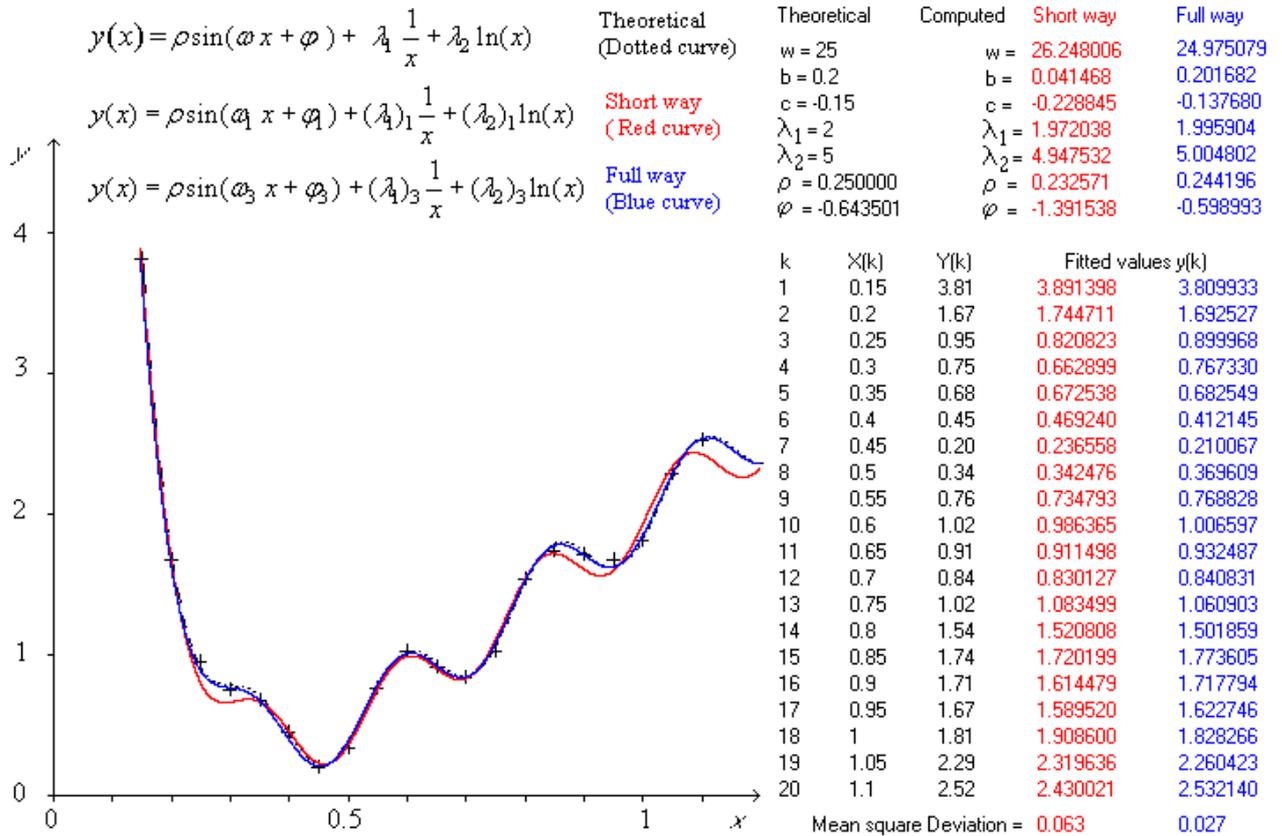
$\theta = \omega x + \varphi$ Dotted line
 $\theta = \omega_2 x + \varphi_2$ Blue line



$\omega_3 = 2.49751E+01$

Drawing θ as a function of x is very useful to check if there is no mistake in the determinations of arctan.

The results are summarized on the next figure where the short and full ways are compared.



DAMPED SINUSOIDAL REGRESSION

Jean Jacquelin

The goal is to fit the function $y(x) = \rho \sin(\omega x + \varphi) \exp(\alpha x)$ to a data set :

$$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$$

with $\alpha < 0$ in case of damped sinusoidal function.

The parameters to be optimized are : $\omega, \varphi, \alpha, \rho$

The problem is the same as the fitting of the functions :

$$y(x) = b \sin(\omega x) \exp(\alpha x) + c \cos(\omega x) \exp(\alpha x)$$

where the parameters to be optimized are : ω, α, b, c

$$\text{because } \rho \sin(\omega x + \varphi) = b \sin(\omega x) + c \cos(\omega x) \quad \begin{cases} b = \rho \cos(\varphi) \\ c = \rho \sin(\varphi) \end{cases}$$

Integral equation :

Two successive integrations of $y(x)$ leads to two formulas, which can be combined with the $y(x)$ formula in order to eliminate the terms $\rho \sin(\omega x + \varphi) \exp(\alpha x)$ and $\rho \cos(\omega x + \varphi) \exp(\alpha x)$:

$$y = -(\alpha^2 + \omega^2) \int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv + 2\alpha \int_{x_1}^x y(u) du + Cx + D$$

C and D are constants which depend on the lower bound of the integrals.

Of course, lower bounds other than x_1 could be chosen, but this would introduce more complicated terms and more complicated numerical integrations.

Of course, it is possible to analytically express C and D . The formulas are complicated (similar to eq.11 in chapter "Régression sinusoidale", but with more terms). This arduous calculus will be avoid : Instead of using C and D to compute the approximates of ρ and φ , or of b and c , we will see that they can be obtained much more easily, thanks to a complementary linear regression.

The integral equation above is on the linear kind :

$$y(x) = A S(x) + B SS(x) + Cx + D$$

$$\text{where } \begin{cases} A = -(\alpha^2 + \omega^2) & ; & B = 2\alpha \\ S(x) = \int_{x_1}^x y(u) du & ; & SS(x) = \int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv \end{cases}$$

The values of $S(x_k)$ and $SS(x_k)$ are approximated by numerical integration.

According to the integral equation, a linear regression gives A_0, B_0, C_0, D_0 which are approximates of A, B, C, D . (Details in next section "Short way ")

$$\text{So, the approximates of } \alpha \text{ and } \omega \text{ are obtained : } \alpha_1 = \frac{B_0}{2} & ; & \omega_1 = \sqrt{-(A_0 + \alpha_1^2)}$$

Then, ω_1 and α_1 are the starting values for a linear regression according to the approximate function : $y(x) \simeq b \sin(\omega_1 x) \exp(\alpha_1 x) + c \cos(\omega_1 x) \exp(\alpha_1 x)$, which leads directly to the approximates of b , c and then to the approximates of ρ and φ , namely b_1 , c_1 , ρ_1 , φ_1 . This is a short way to obtain the approximate equation of $y(x)$:

$$y(x) \simeq b_1 \sin(\omega_1 x) \exp(\alpha_1 x) + c_1 \cos(\omega_1 x) \exp(\alpha_1 x)$$

or : $y(x) \simeq \rho_1 \sin(\omega_1 x + \varphi_1) \exp(\alpha_1 x)$

The "**short way**" above might be not sufficient in case of small data set (n small), or in case of low number of periods of the sinusoidal part of the function, or in case of large scatter, or in case of bad distribution of the x_k on the range of x , as it is pointed out in the chapter "Régression sinusoidale".

The "**Full way**" starts with the values α_1 and ρ_1 above.

Let $\theta = \omega x + \varphi = \arcsin\left(\frac{y(x)}{\rho_1 \exp(\alpha_1 x)}\right)$

The inverse functions of sin, cos or tan have an infinity of determinations. The right one must be determined. This is shown in the next section "Details of the process", where arctan is the inverse function used.

The computation of $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n$ with $\theta_k = \arcsin\left(\frac{y_k}{\rho_1 \exp(\alpha_1 x_k)}\right)$

(with the correct respective determinations) allows a linear regression relatively to the function $\theta = \omega x + \varphi$. This leads to the approximates ω_2 and φ_2 of ω and φ respectively.

Then, $\omega_3 = \omega_2$, $\varphi_3 = \varphi_2$, and $\alpha_3 = \alpha_1$ are used as starting values for a linear regression relatively to the approximate function :

$y(x) \simeq b \sin(\omega_3 x) \exp(\alpha_3 x) + c \cos(\omega_3 x) \exp(\alpha_3 x)$, which leads to the approximates of b , c , respectively named b_3 , c_3 . Finally, the approximate of the equation $y(x)$ is :

$$y(x) \simeq b_3 \sin(\omega_3 x) \exp(\alpha_3 x) + c_3 \cos(\omega_3 x) \exp(\alpha_3 x)$$

or $y(x) \simeq \rho_3 \sin(\omega_3 x + \varphi_3) \exp(\alpha_3 x)$

The subscripts 1, 2, 3 for the approximates of the fitted parameters were chosen in order to be consistent with the notations used in the main chapter "Sinusoidal Regression".

DETAILS OF THE PROCESS

$$y(x) = \rho \sin(\omega x + \varphi) \exp(\alpha x)$$

$$\text{or } y(x) = b \sin(\omega x) \exp(\alpha x) + c \cos(\omega x) \exp(\alpha x)$$

Set of data : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

They must be written by ascending values of x .

Short way :

- Computation of $S_1, S_2, \dots, S_k, \dots, S_n$:

$$S_1 = 0 \quad ; \quad S_k = S_{k-1} + \frac{1}{2}(y_k + y_{k-1})(x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Computation of $SS_1, SS_2, \dots, SS_k, \dots, SS_n$:

$$SS_1 = 0 \quad ; \quad SS_k = SS_{k-1} + \frac{1}{2}(S_k + S_{k-1})(x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Solving of the regression system for A_0, B_0, C_0 and D_0 :

$$\begin{pmatrix} A_0 \\ B_0 \\ C_0 \\ D_0 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n SS_k^2 & \sum_{k=1}^n SS_k S_k & \sum_{k=1}^n SS_k x_k & \sum_{k=1}^n SS_k \\ \sum_{k=1}^n SS_k S_k & \sum_{k=1}^n S_k^2 & \sum_{k=1}^n S_k x_k & \sum_{k=1}^n S_k \\ \sum_{k=1}^n SS_k x_k & \sum_{k=1}^n S_k x_k & \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \\ \sum_{k=1}^n SS_k & \sum_{k=1}^n S_k & \sum_{k=1}^n x_k & n \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n SS_k y_k \\ \sum_{k=1}^n S_k y_k \\ \sum_{k=1}^n x_k y_k \\ \sum_{k=1}^n y_k \end{pmatrix}$$

- Computation of : $\alpha_1 = \frac{B_0}{2}$; $\omega_1 = \sqrt{-(A_0 + \alpha_1^2)}$

- Computation of $\beta_1, \dots, \beta_k, \dots, \beta_m$ and $\eta_1, \dots, \eta_k, \dots, \eta_m$:

$$\beta_k = \sin(\omega_1 x_k) \exp(\alpha_1 x_k) \quad ; \quad \eta_k = \cos(\omega_1 x_k) \exp(\alpha_1 x_k)$$

- Solving of the regression system for b_1, c_1 :

$$\begin{pmatrix} b_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n \beta_k^2 & \sum_{k=1}^n \beta_k \eta_k \\ \sum_{k=1}^n \beta_k \eta_k & \sum_{k=1}^n \eta_k^2 \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \beta_k y_k \\ \sum_{k=1}^n \eta_k y_k \end{pmatrix}$$

Computation of : $\rho_1 = \sqrt{b_1^2 + c_1^2}$ and of φ_1 : $\begin{cases} \text{if } b_1 > 0 & \rightarrow \varphi_1 = \arctan\left(\frac{c_1}{b_1}\right) \\ \text{if } b_1 < 0 & \rightarrow \varphi_1 = \arctan\left(\frac{c_1}{b_1}\right) + \pi \end{cases}$

Full way (continuation of the process above) :

- Computation of $K_1, K_2, \dots, K_k, \dots, K_n$: $K_k = \text{round}\left(\frac{1}{\pi}(\omega_1 x_k + \varphi_1)\right)$

[The real argument is rounded to the nearest integer]

- Computation of $\theta_1, \theta_2, \dots, \theta_k, \dots, \theta_n$:

$$\left\{ \begin{array}{l} r_k = y_k \exp(-\alpha_1 x_k) \\ \text{if } \rho_1^2 > r_k^2 \rightarrow \theta_k = (-1)^{K_k} \text{atan}\left(\frac{r_k}{\sqrt{\rho_1^2 - r_k^2}}\right) + \pi K_k \\ \text{if } \rho_1^2 \leq r_k^2 \rightarrow \left\{ \begin{array}{l} \text{if } r_k > 0 \rightarrow \theta_k = \frac{\pi}{2}(-1)^{K_k} + \pi K_k \\ \text{if } r_k < 0 \rightarrow \theta_k = -\frac{\pi}{2}(-1)^{K_k} + \pi K_k \end{array} \right. \end{array} \right.$$

- Solving of the regression system for ω_2 and φ_2 :

$$\begin{pmatrix} \omega_2 \\ \varphi_2 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \\ \sum_{k=1}^n x_k & n \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \theta_k x_k \\ \sum_{k=1}^n \theta_k \end{pmatrix}$$

- Computation of b_2 and c_2 : $b_2 = \rho_2 \cos(\varphi_2)$; $c_2 = \rho_2 \sin(\varphi_2)$

- with $\omega_3 = \omega_2$ and $\alpha_3 = \alpha_2 = \alpha_1$ computation of :

$$\beta_k = \sin(\omega_3 x_k) \exp(\alpha_3 x_k) \quad ; \quad \eta_k = \cos(\omega_3 x_k) \exp(\alpha_3 x_k)$$

- Solving of the regression system for b_3, c_3 :

$$\begin{pmatrix} b_3 \\ c_3 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n \beta_k^2 & \sum_{k=1}^n \beta_k \eta_k \\ \sum_{k=1}^n \beta_k \eta_k & \sum_{k=1}^n \eta_k^2 \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \beta_k y_k \\ \sum_{k=1}^n \eta_k y_k \end{pmatrix}$$

Computation of ρ_3 and φ_3 : $\rho_3 = \sqrt{b_3^2 + c_3^2}$; $\left\{ \begin{array}{l} \text{if } b_3 > 0 \rightarrow \varphi_3 = \arctan\left(\frac{c_3}{b_3}\right) \\ \text{if } b_3 < 0 \rightarrow \varphi_3 = \arctan\left(\frac{c_3}{b_3}\right) + \pi \end{array} \right.$

Result : the fitted function is

$$y(x) \simeq \rho_3 \sin(\omega_3 x + \varphi_3) \exp(\alpha_3 x)$$

$$\text{or : } y(x) \simeq b_3 \sin(\omega_3 x) \exp(\alpha_3 x) + c_3 \cos(\omega_3 x) \exp(\alpha_3 x)$$

:
EXAMPLE

Generation of the data set : (n = 20)

"Theretical" values of ω , φ , ρ , α are chosen, for example : $\omega=1.5$, $\varphi=0.8$, $\rho=4$, $\alpha= -0.2$
 "Theretical" values of $y(x_k) = \rho \sin(\omega x_k + \varphi) \exp(\alpha x_k)$ are computed. They are scattered by adding random real numbers taken on the range $-\epsilon$ to $+\epsilon$ (for example $\epsilon = 0.1$). They are rounded to two decimal places (next table). The points (x_k, y_k) are drawn on the figure at the end of the current section. On the same graph, the dotted curve represents the "theoretical" function $y(x) = \rho \sin(\omega x + \varphi) \exp(\alpha x)$

Short way :

| k | x_k | y_k | S_k | SS_k |
|-----|-------|-------|-------------|-------------|
| 1 | 0 | 2.58 | 0.00000E+00 | 0.00000E+00 |
| 2 | 0.5 | 3.28 | 1.46500E+00 | 1.46500E+00 |
| 3 | 1 | 2.62 | 2.94000E+00 | 2.94000E+00 |
| 4 | 1.5 | 0.25 | 3.65750E+00 | 3.65750E+00 |
| 5 | 2 | -1.57 | 3.32750E+00 | 3.32750E+00 |
| 6 | 2.5 | -2.48 | 2.31500E+00 | 2.31500E+00 |
| 7 | 3 | -1.76 | 1.25500E+00 | 1.25500E+00 |
| 8 | 3.5 | -0.43 | 7.07500E-01 | 7.07500E-01 |
| 9 | 4 | 0.87 | 8.17500E-01 | 8.17500E-01 |
| 10 | 4.5 | 1.53 | 1.41750E+00 | 1.41750E+00 |
| 11 | 5 | 1.22 | 2.10500E+00 | 2.10500E+00 |
| 12 | 5.5 | 0.48 | 2.53000E+00 | 2.53000E+00 |
| 13 | 6 | -0.40 | 2.55000E+00 | 2.55000E+00 |
| 14 | 6.5 | -1.05 | 2.18750E+00 | 2.18750E+00 |
| 15 | 7 | -0.86 | 1.71000E+00 | 1.71000E+00 |
| 16 | 7.5 | -0.42 | 1.39000E+00 | 1.39000E+00 |
| 17 | 8 | 0.20 | 1.33500E+00 | 1.33500E+00 |
| 18 | 8.5 | 0.59 | 1.53250E+00 | 1.53250E+00 |
| 19 | 9 | 0.69 | 1.85250E+00 | 1.85250E+00 |
| 20 | 9.5 | 0.35 | 2.11250E+00 | 2.11250E+00 |

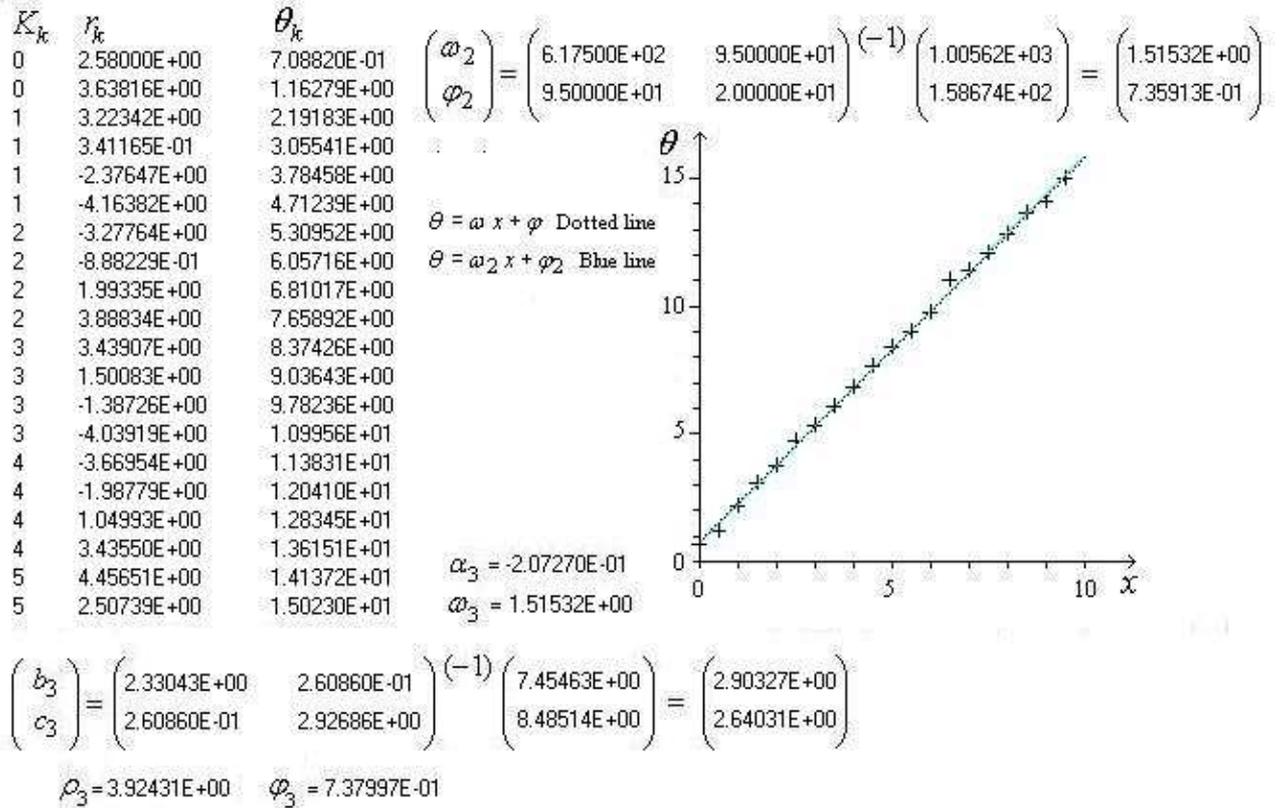
$$\begin{pmatrix} A_0 \\ B_0 \\ C_0 \\ D_0 \end{pmatrix} = \begin{pmatrix} 2.38083E+03 & 3.46100E+02 & 1.21071E+03 & 1.88481E+02 \\ 3.46100E+02 & 8.43528E+01 & 1.74293E+02 & 3.72075E+01 \\ 1.21071E+03 & 1.74293E+02 & 6.17500E+02 & 9.50000E+01 \\ 1.88481E+02 & 3.72075E+01 & 9.50000E+01 & 2.00000E+01 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \\ -1 \\ -1 \end{pmatrix} = \begin{pmatrix} -2.49628E+00 \\ 4.42398E+00 \\ 5.37000E+00 \\ 5.69000E+00 \end{pmatrix} = \begin{pmatrix} -2.45404E+00 \\ -4.14540E-01 \\ 4.51974E+00 \\ 2.71385E+00 \end{pmatrix}$$

$$\omega_1 = 1.55276E+00 \quad \alpha_1 = -2.07270E-01$$

$$\begin{pmatrix} b_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} 2.33251E+00 & 2.47078E-01 \\ 2.47078E-01 & 2.92478E+00 \end{pmatrix} \begin{pmatrix} -1 \\ -1 \end{pmatrix} = \begin{pmatrix} 7.93227E+00 \\ 7.83101E+00 \end{pmatrix} = \begin{pmatrix} 3.14528E+00 \\ 2.41176E+00 \end{pmatrix}$$

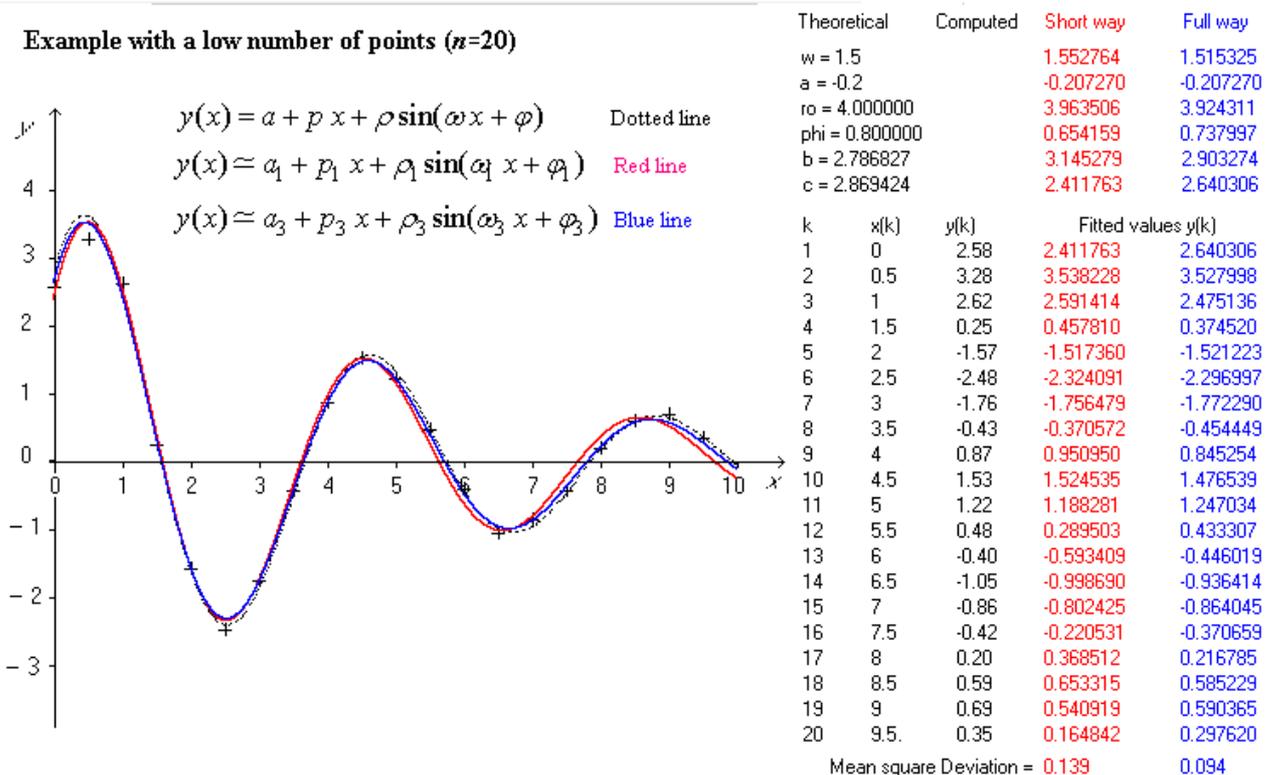
$$\rho_1 = 3.96351E+00 \quad \varphi_1 = 6.54159E-01$$

Full way :

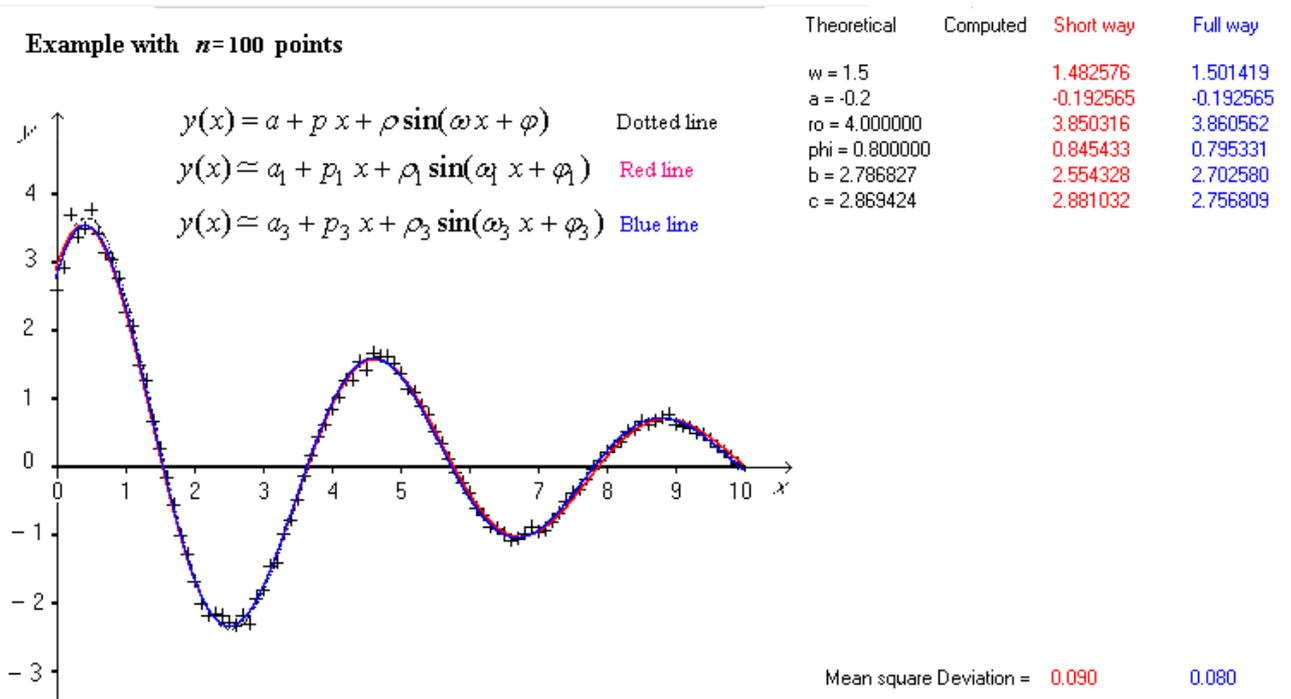


Drawing θ as a function of x is very useful to check if there is no mistake in the determinations of arctan.

The results are summarized on the next figure where the short and full ways are compared.



In order to make clear the differences between the curves, the data has been limited to a small number of points. In fact, 20 points distributed on about three periods is not enough. A more realistic example, with more points per period, is shown on the next figure ($n=100$ instead of 20, with the same scatter).



We can see that the "short way" yields to almost the same result than the "full way" which is more complicated. If the number of points per period is large enough, the "short way" is sufficient.

DOUBLE EXPONENTIAL REGRESSION

DOUBLE POWER REGRESSION

The goal is to fit the function $y(x) = b \exp(px) + c \exp(qx)$ to a data set :

$$(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$$

or, to fit the function $Y(X) = b X^p + c X^q$ to a data set :

$$(X_1, Y_1), (X_2, Y_2), \dots, (X_k, Y_k), \dots, (X_n, Y_n)$$

In both cases, the parameters to be optimized are : p, q, b, c

The method is the same, thanks to the conversion of data : $x_k = \ln(X_k)$; $y_k = Y_k$

Two successive integrations of $y(x)$ leads to two formulas, which can be combined with the $y(x)$ formula in order to eliminate the terms $\exp(px)$ and $\exp(qx)$:

$$y = -pq \int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv + (p+q) \int_{x_1}^x y(u) du + Cx + D$$

C and D are constants which depend on the lower bound of the integrals.

Of course, lower bounds other than x_1 could be chosen, but this would introduce more complicated terms and more complicated numerical integrations.

Of course, it is possible to analytically express C and D . The formulas are complicated. This calculus will be avoid : Instead of using C and D to compute the approximates of b and c , we will see that they can be obtained much more easily, thanks to a complementary linear regression.

The integral equation above is on the linear kind :

$$y(x) = A SS(x) + B S(x) + Cx + D$$

where
$$\begin{cases} A = pq & ; & B = -(p+q) \\ S(x) = \int_{x_1}^x y(u) du & ; & SS(x) = \int_{x_1}^x \left(\int_{x_1}^v y(u) du \right) dv \end{cases}$$

The values of $S(x_k)$ and $SS(x_k)$ are appoximated by numerical integration.

According to the integral equation, a linear regression gives A_1, B_1 which are approximates of A, B as shown latter.

So, the approximates of p and q are obtained :

$$p_1 = \frac{1}{2} \left(B_1 + \sqrt{B_1^2 + 4A_1} \right) & ; & q_1 = \frac{1}{2} \left(B_1 - \sqrt{B_1^2 + 4A_1} \right)$$

Then, p_1 and q_1 are the starting values for a linear regression according to the approximate function : $y(x) \simeq b \exp(p_1 x) + c \exp(q_1 x)$, which leads directly to the approximates b_1 and c_1 of b and c . So, an approximate of the equation of $y(x)$ is obtained:

$$y(x) \simeq b_1 \exp(p_1 x) + c_1 \exp(q_1 x)$$

The process is shown in more details on the next page. Then, two numerical examples are presented.

DETAILS OF THE PROCESS

$$Y(x) = b X^p + c X^q$$

Set of data : $(X_1, Y_1), (X_2, Y_2), \dots, (X_k, Y_k), \dots, (X_n, Y_n)$

- In this case, compute : $x_k = \ln(X_k)$; $y_k = Y_k$ which leads to the next form :

$$y(x) = b \exp(px) + c \exp(qx)$$

Set of data : $(x_1, y_1), (x_2, y_2), \dots, (x_k, y_k), \dots, (x_n, y_n)$

They must be written by ascending values of x .

- Computation of $S_1, S_2, \dots, S_k, \dots, S_n$:

$$S_1 = 0 \quad ; \quad S_k = S_{k-1} + \frac{1}{2}(y_k + y_{k-1})(x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Computation of $SS_1, SS_2, \dots, SS_k, \dots, SS_n$:

$$SS_1 = 0 \quad ; \quad SS_k = SS_{k-1} + \frac{1}{2}(S_k + S_{k-1})(x_k - x_{k-1}) \quad \text{from } k=2 \text{ to } n$$

- Solving of the regression system for A_1, B_1, C_1 and D_1 :

$$\begin{pmatrix} A_1 \\ B_1 \\ C_1 \\ D_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n SS_k^2 & \sum_{k=1}^n SS_k S_k & \sum_{k=1}^n SS_k x_k & \sum_{k=1}^n SS_k \\ \sum_{k=1}^n SS_k S_k & \sum_{k=1}^n S_k^2 & \sum_{k=1}^n S_k x_k & \sum_{k=1}^n S_k \\ \sum_{k=1}^n SS_k x_k & \sum_{k=1}^n S_k x_k & \sum_{k=1}^n x_k^2 & \sum_{k=1}^n x_k \\ \sum_{k=1}^n SS_k & \sum_{k=1}^n S_k & \sum_{k=1}^n x_k & n \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n SS_k y_k \\ \sum_{k=1}^n S_k y_k \\ \sum_{k=1}^n x_k y_k \\ \sum_{k=1}^n y_k \end{pmatrix}$$

- Computation of : $p_1 = \frac{1}{2} \left(B_1 + \sqrt{B_1^2 + 4A_1} \right)$; $q_1 = \frac{1}{2} \left(B_1 - \sqrt{B_1^2 + 4A_1} \right)$

- Computation of $\beta_1, \dots, \beta_k, \dots, \beta_m$ and $\eta_1, \dots, \eta_k, \dots, \eta_m$:

$$\beta_k = \exp(p_1 x_k) \quad ; \quad \eta_k = \exp(q_1 x_k)$$

- Solving of the regression system for b_1, c_1 :

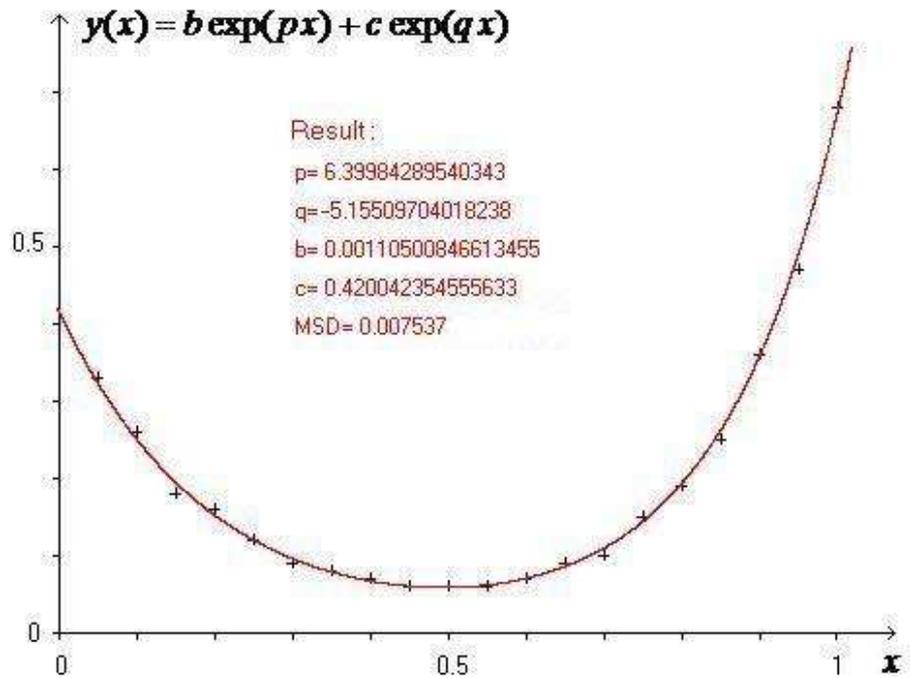
$$\begin{pmatrix} b_1 \\ c_1 \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n \beta_k^2 & \sum_{k=1}^n \beta_k \eta_k \\ \sum_{k=1}^n \beta_k \eta_k & \sum_{k=1}^n \eta_k^2 \end{pmatrix}^{(-1)} \begin{pmatrix} \sum_{k=1}^n \beta_k y_k \\ \sum_{k=1}^n \eta_k y_k \end{pmatrix}$$

Result : the fitted function is $y(x) \simeq b_1 \exp(p_1 x) + c_1 \exp(q_1 x)$

$$\text{or } Y(X) \simeq b_1 X^{p_1} + c_1 X^{q_1}$$

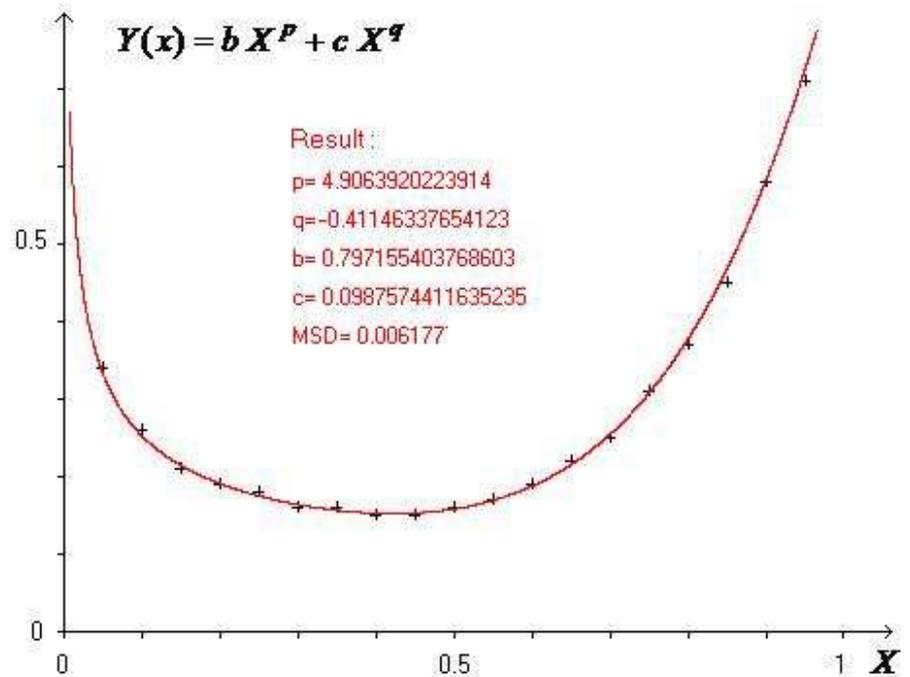
EXAMPLE of fitting, case of double exponential function :

| k | x | y |
|----|------|------|
| 1 | 0,05 | 0,33 |
| 2 | 0,1 | 0,26 |
| 3 | 0,15 | 0,18 |
| 4 | 0,2 | 0,16 |
| 5 | 0,25 | 0,12 |
| 6 | 0,3 | 0,09 |
| 7 | 0,35 | 0,08 |
| 8 | 0,4 | 0,07 |
| 9 | 0,45 | 0,06 |
| 10 | 0,5 | 0,06 |
| 11 | 0,55 | 0,06 |
| 12 | 0,6 | 0,07 |
| 13 | 0,65 | 0,09 |
| 14 | 0,7 | 0,1 |
| 15 | 0,75 | 0,15 |
| 16 | 0,8 | 0,19 |
| 17 | 0,85 | 0,25 |
| 18 | 0,9 | 0,36 |
| 19 | 0,95 | 0,47 |
| 20 | 1 | 0,68 |



EXAMPLE of fitting, case of double power function :

| k | X | Y |
|----|------|------|
| 1 | 0,05 | 0,34 |
| 2 | 0,1 | 0,26 |
| 3 | 0,15 | 0,21 |
| 4 | 0,2 | 0,19 |
| 5 | 0,25 | 0,18 |
| 6 | 0,3 | 0,16 |
| 7 | 0,35 | 0,16 |
| 8 | 0,4 | 0,15 |
| 9 | 0,45 | 0,15 |
| 10 | 0,5 | 0,16 |
| 11 | 0,55 | 0,17 |
| 12 | 0,6 | 0,19 |
| 13 | 0,65 | 0,22 |
| 14 | 0,7 | 0,25 |
| 15 | 0,75 | 0,31 |
| 16 | 0,8 | 0,37 |
| 17 | 0,85 | 0,45 |
| 18 | 0,9 | 0,58 |
| 19 | 0,95 | 0,71 |
| 20 | 1 | 0,91 |



GENERALISATION

Case of more exponential terms :

One understand that fitting a sum of different exponential terms (or power terms) can be carried out on the same principle. If the number of exponential terms is N , we need N successive integrals. Each of those integrals is a linear sum of the exponentials. Solving this linear system leads to each exponential term expressed as a sum of the N integrals. Bringing them back into the function to be fitted leads to a linear relationship between $y(x)$ and the integrals. This relationship includes in addition a $(N-1)$ degree polynomial coming from the lower bounds of the integrals. Altogether, $y(x)$ is equal to a sum of $(2N)$ terms.

Then, a linear regression gives the approximates of the $2N$ coefficients of those terms. In fact, only N will be used (the coefficients of the integrals) to built the system of N equations where the N unknowns are the shape parameters (i.e. the coefficient of x in the argument of each exponential). This system of N equations is non-linear. For large N , the numerical solving requires a software to compute the N roots which are the approximates of the N shape parameters.

The relationship between $y(x)$ and the sum of exponential terms is now linear since each exponential can be computed knowing its shape parameter. Then, a linear regression gives the approximate values of the corresponding N magnitude parameters (i.e. the coefficient of each exponential).

Finally, the approximate values of the $2N$ parameters of the fitted function are obtained (i.e. for each exponential term : a magnitude coefficient and a shape coefficient)

Case of a sum of exponential terms and other given functions with unknown coefficients (without unknown parameter inside the functions). This general case includes the case of an unknown constant added to the sum of the exponential terms.

In addition to the above comments, (number of successive integrations depending on the number of exponentials), the treatment of the additional functions is explained in the preceeding chapter "GENERALIZED SINUSOIDAL REGRESSION" .

Of course, the more the number of functions considered is large, the more the theoretical calculus leading to the convenient integral equation is arduous and the more the practical application is complicated. But this is anyways a straightforward method. No initial guessed values are required. In case of a sum of several exponential terms and other functions, it becommes almost impossible to guess some convenient initial values. That is why the method described here is more robust.

REGRESSIONS MULTI-VARIABLES

Au lieu d'une seule variable x , la fonction à ajuster comporte plusieurs variables x, t, \dots

Rappel du cas linéaire :

Dans ce cas simple, la fonction $y(x, t, \dots)$ est linéaire relativement aux coefficients $\lambda_1, \lambda_2, \dots, \lambda_j, \dots, \lambda_m$ à optimiser :

$$y(x, t, \dots) = \lambda_1 f_1(x, t, \dots) + \dots + \lambda_j f_j(x, t, \dots) + \dots + \lambda_m f_m(x, t, \dots)$$

Les fonctions données $f_j(x, t, \dots)$ ne contiennent pas de paramètre ajustable.

Les n points données étant $(x_1, t_1, \dots; y_1), (x_2, t_2, \dots; y_2), \dots, (x_k, t_k, \dots; y_k), \dots, (x_n, t_n, \dots; y_n)$ les paramètres se calculent par la méthode des moindres carrés :

Avec : $f_{j,k} = f_j(x_k, t_k, \dots)$

$$\begin{pmatrix} \lambda_1 \\ \dots \\ \lambda_j \\ \dots \\ \lambda_m \end{pmatrix} = \begin{pmatrix} \sum_{k=1}^n (f_{1,k})^2 & \dots & \sum_{k=1}^n f_{1,k} f_{j,k} & \dots & \sum_{k=1}^n f_{1,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{1,k} f_{j,k} & \dots & \sum_{k=1}^n (f_{j,k})^2 & \dots & \sum_{k=1}^n f_{j,k} f_{m,k} \\ \dots & \dots & \dots & \dots & \dots \\ \sum_{k=1}^n f_{1,k} f_{m,k} & \dots & \sum_{k=1}^n f_{j,k} f_{m,k} & \dots & \sum_{k=1}^n (f_{m,k})^2 \end{pmatrix}^{-1} \begin{pmatrix} \sum_{k=1}^n y_k f_{1,k} \\ \dots \\ \sum_{k=1}^n y_k f_{j,k} \\ \dots \\ \sum_{k=1}^n y_k f_{m,k} \end{pmatrix}$$

Cas non linéaire :

Si une ou plusieurs fonctions contiennent un ou plusieurs paramètres ajustables, $y(x, t, \dots)$ n'est pas linéaire relativement à l'ensemble des paramètres à ajuster/

$$y(x, t, \dots) = b_1 \varphi_1(p_1, p_2, \dots; x, t, \dots) + b_2 \varphi_2(p_1, p_2, \dots; x, t, \dots) + \dots + \lambda_1 f_1(x, t, \dots) + \dots + \lambda_j f_j(x, t, \dots) + \dots + \lambda_m f_m(x, t, \dots)$$

Les fonctions données $\varphi_1, \varphi_2, \dots$ contiennent les paramètres ajustables p_1, p_2, \dots

L'ensemble des paramètres à optimiser est alors : $p_1, p_2, \dots, b_1, b_2, \dots, \lambda_1, \dots, \lambda_j, \dots, \lambda_m$

La méthode des moindres carrés ne peut pas être appliquée directement. De nombreuses variantes sont utilisées, consistant généralement à supposer des valeurs initiales pour les paramètres intervenant non linéairement (p_1, p_2, \dots) et ensuite, à procéder par calculs récursifs pour corriger les valeurs initialement supposées en se rapprochant progressivement de valeurs optimum.

La méthode étudiée ici est d'un principe très différent. On effectue une ou plusieurs intégrations de $y(x, t, \dots)$, telles que par exemple : $\int y(x, t, \dots) dx$, ou $\int y(x, t, \dots) dt$, ou $\iint y(x, t, \dots) dx dt$, ou d'autres multiples intégrations.

On peut également introduire des fonctions nouvelles $g(x, t, \dots)$ lorsque cela permet des intégrations plus simples : $\int g(x, t, \dots) y(x, t, \dots) dx$, ou $\int g(x, t, \dots) y(x, t, \dots) dt$, ou $\iint g(x, t, \dots) y(x, t, \dots) dx dt$, ou d'autres multiples intégrations

On voit donc que les possibilités sont extrêmement nombreuses. Il est parfois possible de combiner linéairement entre elles certaines des relations obtenues, de telle sorte que l'on fait disparaître les termes non linéaires relativement aux paramètres p_1, p_2, \dots

Par exemple, considérons la fonction suivante :

$$y(x, t) = b \exp(p x t) + \lambda (x - t)^2 \quad \left\{ \begin{array}{l} b_1 = b \quad ; \quad b_2 = b_3 = \dots = 0 \\ p_1 = p \quad ; \quad p_2 = p_3 = \dots = 0 \\ \varphi(p_1, p_2, \dots; x, t, \dots) = \exp(p x t) \\ \lambda_1 = \lambda \quad ; \quad \lambda_2 = \lambda_3 = \dots = 0 \\ f_1(x, t, \dots) = (x - t)^2 \end{array} \right.$$

$$\left\{ \begin{array}{l} \text{avec :} \\ g(x, t) = t \end{array} \right\} \left\{ \begin{array}{l} \int g(x, t) y(x, t) dx = \int t b \exp(p x t) dx + \int t \lambda (x - t)^2 dx \\ = \frac{b}{p} \exp(p x t) + \frac{\lambda}{3} t (x - t)^3 + \dots \end{array} \right.$$

On voit qu'une combinaison linéaire entre cette équation et l'équation définissant $y(x, t)$ permet de faire disparaître le terme $\exp(p x t)$:

$$y(x, t) = p \int t y(x, t) dx - \frac{1}{3} \lambda p t (x - t)^3 + \lambda (x - t)^2 + \dots$$

$$y(x, t) = p S(x, t) + \lambda_1 f_1(x, t) + \lambda_2 f_2(x, t) + \dots \quad \left\{ \begin{array}{l} S(x, t) = \int t y(x, t) dx \\ f_1(x, t) = t (x - t)^3 \\ f_2(x, t) = (x - t)^2 \end{array} \right.$$

$S(x, t)$ sera calculé approximativement à partir des données numériques.

Cette relation est linéaire relativement aux paramètres p, λ_1, λ_2 . La régression par les moindres carrés donne une valeur approchée de p , ce qui était recherché. Ce qui ramène ensuite l'équation $y(x, t)$ sous forme d'une relation linéaire relativement à b et λ .

Néanmoins, les équations de la page précédente ne sont pas correctes car elles ont été volontairement écrites incomplètement, pour rendre l'explication plus aisée à suivre. En effet, les intégrales doivent ne pas rester indéfinies, en particulier pour rendre possible l'intégration numérique. Il est indispensable de définir la borne inférieure d'intégration. Cette obligation ne posait pas de difficulté fondamentale dans le cas de fonction $y(x)$ à une seule variable. Cela devient un problème délicat dans le cas de fonctions $y(x,t,\dots)$ à plusieurs variables. Il convient d'examiner attentivement cette question de l'intégration numérique dans le cas de la présence de plusieurs variables (x, t, \dots) , même si l'intégrale ne porte que sur une seule d'entre elles.

Remarque : Dans de nombreux cas, la linéarisation peut aussi être obtenue par des dérivations

partielles de la fonction $y(x,t,\dots)$: $\frac{\partial y}{\partial x}$, ou $\frac{\partial y}{\partial t}$, ou $\frac{\partial^2 y}{\partial t^2}$, ou d'autres dérivations

successives. Il est même souvent plus facile de trouver une équation différentielle ou aux dérivées partielles qui convienne que de trouver une équation intégrale. Mais il a été observé et rendu compte, dès le début de l'étude, que le calcul numérique des dérivées est sujet à des déviations beaucoup plus importantes que le calcul numérique des intégrales, si les données numériques sont affectées de dispersion ou si elles ne sont pas assez régulièrement distribuées.

Intégration numérique :

Le cas est le plus simple lorsque les données sont fournies sous forme de tableau complet :

$$\text{Nombre de points : } n = n_1 n_2 \left\{ \begin{array}{ccccccc} & t_1 & t_2 & \dots & t_{k_2} & \dots & t_{n_2} \\ x_1 & y_{1,1} & y_{1,2} & \dots & y_{1,k_2} & \dots & y_{1,n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{k_1} & y_{k_1,1} & y_{k_1,2} & \dots & y_{k_1,k_2} & \dots & y_{k_1,n_2} \\ \vdots & \vdots & \vdots & \ddots & \vdots & \ddots & \vdots \\ x_{n_1} & y_{n_1,1} & y_{n_1,2} & \dots & y_{n_1,k_2} & \dots & y_{n_1,n_2} \end{array} \right.$$

Les t sont ordonnés par valeurs croissantes de droite à gauche. Les x sont ordonnés par valeurs croissantes de haut en bas.

La méthode d'intégration numérique qui va être décrite est très élémentaire. Considérer des méthodes plus sophistiquées serait certes intéressant, mais risquerait de faire perdre de vue ce qui est essentiel dans la compréhension du principe de linéarisation par équation intégrale.

Intégration de $y(x,t)$ relativement à la variable x :

$$\text{Notation : } \int_{x_1}^{x_{k_1}} y(x, t_{k_2}) dx \simeq S_1(x_{k_1}, t_{k_2})$$

$$\left\{ \begin{array}{l} S_1(x_1, t_{k_2}) = 0 \\ S_1(x_{k_1}, t_{k_2}) = S_1(x_{k_1-1}, t_{k_2}) + \frac{y_{k_1,k_2} + y_{k_1-1,k_2}}{2} (x_{k_1} - x_{k_1-1}) \end{array} \right.$$

Intégration de $y(x,t)$ relativement à la variable t :

$$\text{Notation : } \int_{t_1}^{t_{k_2}} y(x_{k_1}, t) dt \simeq S_2(x_{k_1}, t_{k_2})$$

$$\begin{cases} S_2(x_{k_1}, t_1) = 0 \\ S_2(x_{k_1}, t_{k_2}) = S_2(x_{k_1}, t_{k_2-1}) + \frac{y_{k_1, k_2} + y_{k_1, k_2-1}}{2} (t_{k_2} - t_{k_2-1}) \end{cases}$$

Intégration de $y(x,t)$ relativement aux deux variables :

$$\text{Notation : } \int_{x_1}^{x_{k_1}} \int_{t_1}^{t_{k_2}} y(x_{k_1}, t) dt dx \simeq SS(x_{k_1}, t_{k_2})$$

$$\begin{cases} SS(x_1, t_{k_2}) = 0 \\ SS(x_{k_1}, t_{k_2}) = SS(x_{k_1-1}, t_{k_2}) + \frac{S_2(x_{k_1}, t_{k_2}) + S_2(x_{k_1-1}, t_{k_2})}{2} (x_{k_1} - x_{k_1-1}) \end{cases}$$

Et de même si l'on est amené à utiliser des intégrales multiples.

Lorsqu'on introduit en facteur une fonction $g(x,t)$, les calculs sont similaires, par exemple lorsqu'on intègre $g(x,t)y(x,t)$ relativement à la variable x :

$$\text{Notations : } \begin{cases} \int_{x_1}^{x_{k_1}} g(x, t_{k_2}) y(x, t_{k_2}) dx \simeq S_1(x_{k_1}, t_{k_2}) \\ g_{k_1, k_2} = g(x_{k_1}, t_{k_2}) \end{cases}$$

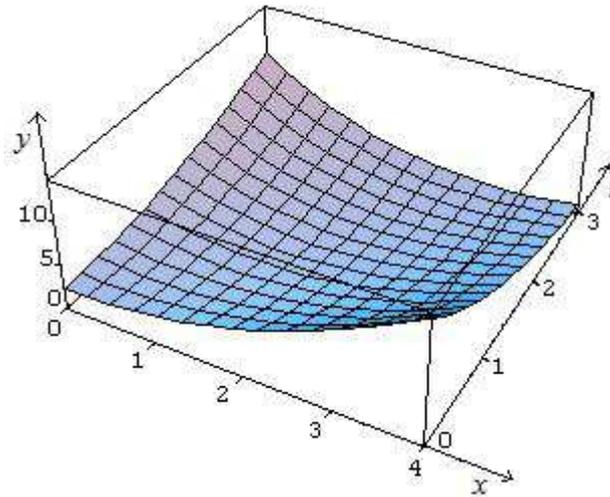
$$\begin{cases} S_1(x_1, t_{k_2}) = 0 \\ S_1(x_{k_1}, t_{k_2}) = S_1(x_{k_1-1}, t_{k_2}) + \frac{g_{k_1, k_2} y_{k_1, k_2} + g_{k_1-1, k_2} y_{k_1-1, k_2}}{2} (x_{k_1} - x_{k_1-1}) \end{cases}$$

Il est inutile de ré-écrire les formules correspondantes aux autres intégrales car on comprend aisément comment établir ces formules, ainsi que d'autres formules correspondant à des intégrales multiples.

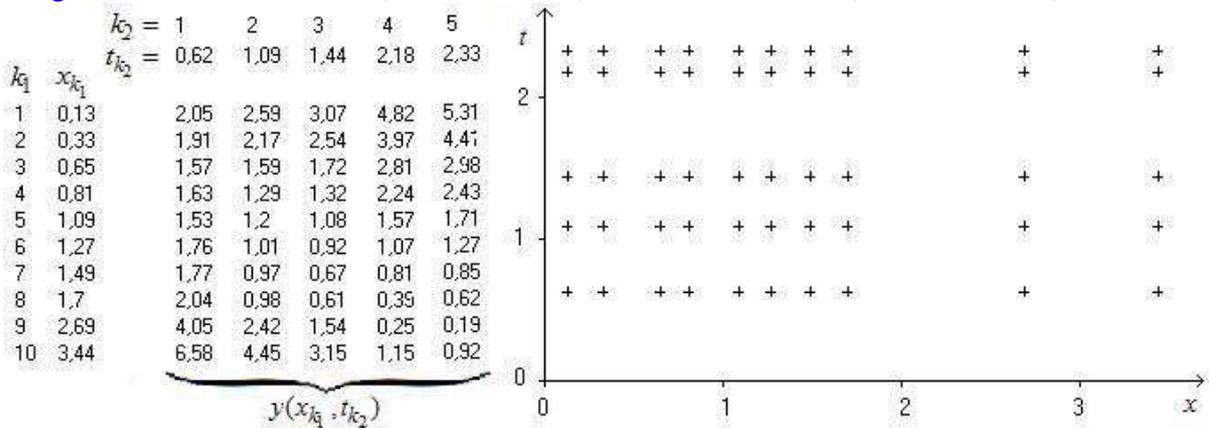
Différentes formes de présentation des données numériques :

Les données peuvent se présenter sous diverses formes. Nous allons considérer, pour une même fonction $y(x,t)$, plusieurs exemples pour lesquels les données numériques se présentent différemment et discuter des conséquences que cela entraîne sur les intégrations numériques respectives.

La figure ci-dessous représente en trois dimensions un exemple de fonction $y(x,t)$.

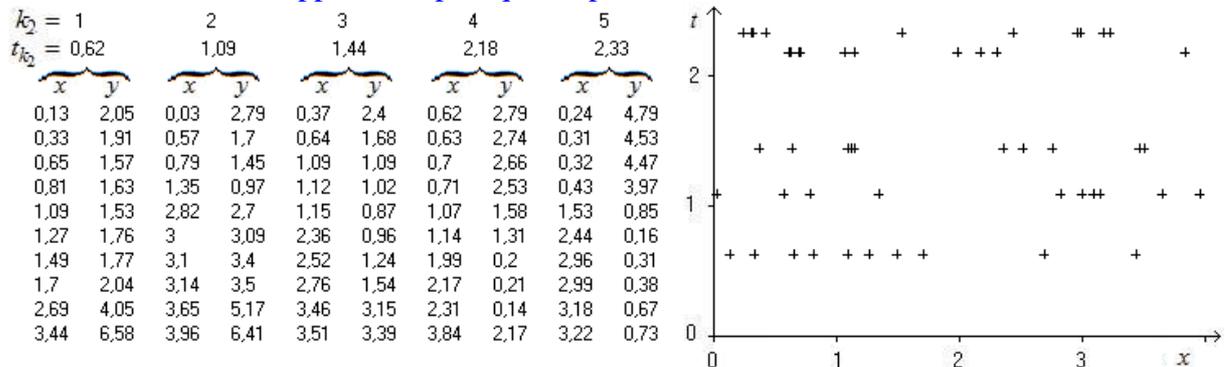


Exemple 1 : Pour cette fonction, le cas de données présentées sous forme de tableau (ci-dessous) correspond à des points (x,t) dont la distribution est représentée sur la figure associée. On a vu, en page précédente, qu'une méthode élémentaire permet de calculer numériquement les valeurs approchées des intégrales dont on a besoin, que ce soit pour les intégrations relativement à x (avec t constant) ou relativement à t (avec x constant).

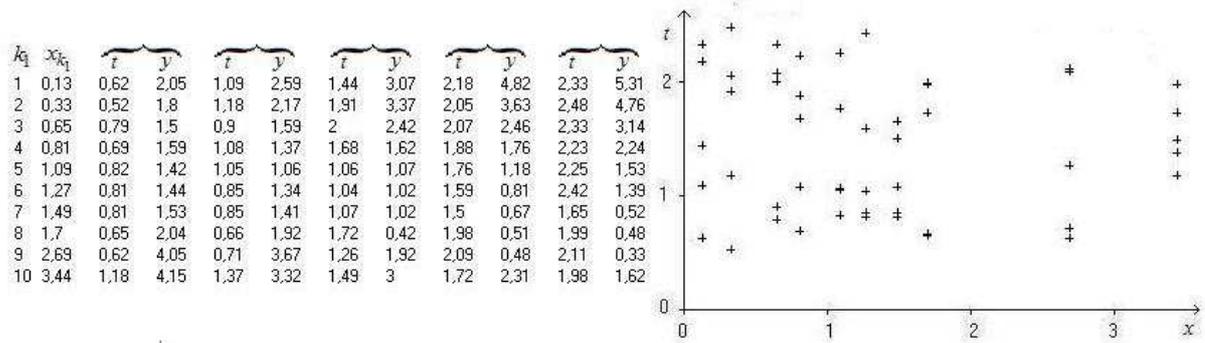


Exemple 2 : Pour la même fonction, dans le cas de données numériques présentées sous la forme suivante, la méthode élémentaire d'intégration ne permet pas, de façon aussi simple, les intégrations relativement à t car il n'y a pas de séries de données à x constant.

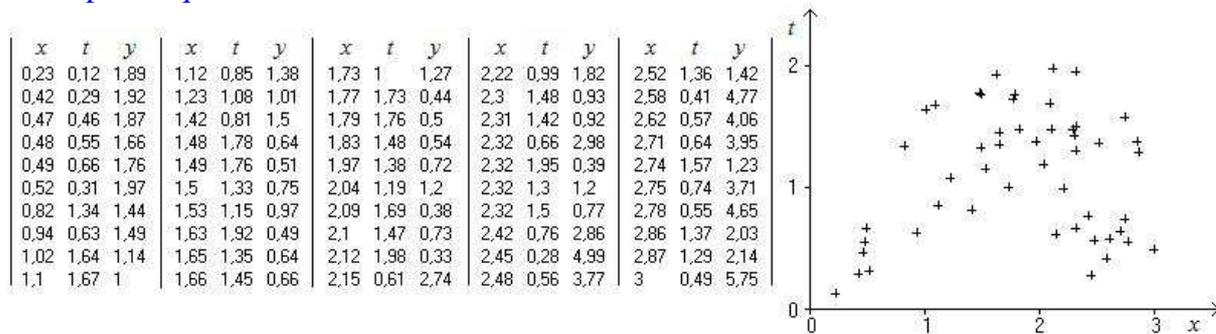
Les intégrations relativement à x (à t constant) sont possibles par la méthode élémentaire. Toutefois, la borne inférieure de l'intégrale n'est pas toujours la même car elle est différente pour chacune des séries de données ayant une valeur commune de t . Ceci complique les notations et alourdi l'application pratique du procédé.



Exemple 3 : Toujours pour la même fonction, dans le cas des données numériques présentées sous la forme suivante, la méthode élémentaire d'intégration ne permet pas, de façon simple, les intégrations relativement à x car il n'y a pas de séries de données à t constant. Les intégrations relativement à t (à x constant) sont possibles par la méthode élémentaire. Toutefois, la borne inférieure de l'intégrale n'est pas toujours la même car elle est différente pour chacune des séries de données ayant une valeur commune de x . Ceci complique les notations et alourdi l'application pratique du procédé.



Exemple 4 : Le cas le plus compliqué se présente lorsque les points (x,t) sont distribués de façon quelconque (exemple ci-dessous, toujours pour la même fonction). La méthode élémentaire d'intégration ne permet, de façon simple, ni les intégrations relativement à x , ni celles relativement à t . Il faudrait faire appel à des méthodes plus élaborées pour le calcul approché d'intégrales doubles sur le domaine sur lequel les points (x,t) sont distribués. Comme on le voit sur la figure correspondante, l'enveloppe de ce domaine d'intégration n'a pas une forme simple. La complication des calculs que cela entraînerait fait perdre l'intérêt de la simplicité qui était recherchée.



On voit que, pour un même problème de régression à plusieurs variables, les calculs numériques d'intégration peuvent être simples ou compliqués selon la forme sous laquelle les données numériques se présentent.

La méthode de linéarisation par équation intégrale est donc surtout intéressante pour simplifier les problèmes de régression non-linéaire à plusieurs variables lorsque les données se présentent sous une forme permettant l'utilisation de méthodes simples d'intégration numérique.

EXEMPLE DÉTAILLÉ DE CALCUL NUMÉRIQUE

$$y(x, t) = b \exp(p x t) + \lambda(x - t)^2$$

Les paramètres à ajuster sont p , b et λ .

L'équation intégrale utilisée est :

$$y(x, t) - y(x_1, t) = p \int_{x_1}^x t y(v, t) dv +$$

$$-\frac{1}{3} \lambda p (t(x-t)^3 - t(x_1-t)^3) + \lambda((x-t)^2 - (x_1-t)^2)$$

Pour simplifier les écritures, les notations sont $k = k_1$ et $h = k_2$

Avec $n_1 = 10$; $n_2 = 5$; $n = n_1 n_2 = 50$, les données sont :

| | | $h = 1$ | 2 | 3 | 4 | 5 | |
|-----|-------|--------------|------|------|------|------|---------------------------|
| k | x_k | $t_h = 0,62$ | 1,09 | 1,44 | 2,18 | 2,33 | |
| 1 | 0,13 | 2,05 | 2,59 | 3,07 | 4,82 | 5,31 | } $y(x_k, t_h) = y_{k,h}$ |
| 2 | 0,33 | 1,91 | 2,17 | 2,54 | 3,97 | 4,41 | |
| 3 | 0,65 | 1,57 | 1,59 | 1,72 | 2,81 | 2,98 | |
| 4 | 0,81 | 1,63 | 1,29 | 1,32 | 2,24 | 2,43 | |
| 5 | 1,09 | 1,53 | 1,2 | 1,08 | 1,57 | 1,71 | |
| 6 | 1,27 | 1,76 | 1,01 | 0,92 | 1,07 | 1,27 | |
| 7 | 1,49 | 1,77 | 0,97 | 0,67 | 0,81 | 0,85 | |
| 8 | 1,7 | 2,04 | 0,98 | 0,61 | 0,39 | 0,62 | |
| 9 | 2,69 | 4,05 | 2,42 | 1,54 | 0,25 | 0,19 | |
| 10 | 3,44 | 6,58 | 4,45 | 3,15 | 1,15 | 0,92 | |

Les t ont été ordonnés par valeurs croissantes. Les x ont été ordonnés par valeurs croissantes.

Calcul de l'intégrale : $S_{k,h} \approx \int_{x_1}^{x_k} t_h y(v, t_h) dv$

$$h = 1 \text{ à } 5 \left\{ \begin{array}{l} S_{1,h} = 0 \\ S_{k,h} = S_{k-1,h} + t_h \left(\frac{y_{k,h} + y_{k-1,h}}{2} \right) (x_k - x_{k-1}) \quad [k = 2 \text{ à } 10] \end{array} \right.$$

| k | $h = 1$ | 2 | 3 | 4 | 5 | |
|-----|-------------|-------------|-------------|-------------|-------------|-------------|
| 1 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | } $S_{k,h}$ |
| 2 | 2,45520E-01 | 5,18840E-01 | 8,07840E-01 | 1,91622E+00 | 2,26476E+00 | |
| 3 | 5,90736E-01 | 1,17458E+00 | 1,78934E+00 | 4,28108E+00 | 5,01975E+00 | |
| 4 | 7,49456E-01 | 1,42572E+00 | 2,13955E+00 | 5,16180E+00 | 6,02818E+00 | |
| 5 | 1,02374E+00 | 1,80569E+00 | 2,62339E+00 | 6,32462E+00 | 7,37864E+00 | |
| 6 | 1,20733E+00 | 2,02250E+00 | 2,88259E+00 | 6,84258E+00 | 8,00355E+00 | |
| 7 | 1,44807E+00 | 2,25990E+00 | 3,13445E+00 | 7,29341E+00 | 8,54691E+00 | |
| 8 | 1,69610E+00 | 2,48307E+00 | 3,32798E+00 | 7,56809E+00 | 8,90654E+00 | |
| 9 | 3,56512E+00 | 4,31754E+00 | 4,86050E+00 | 8,25871E+00 | 9,84076E+00 | |
| 10 | 6,03660E+00 | 7,12566E+00 | 7,39310E+00 | 9,40321E+00 | 1,08106E+01 | |

Pour effectuer la régression linéaire correspondant à la relation :

$$y(x_k, t_h) - y(x_1, t_h) = p S_{k,h} + \lambda_1 f_{1,(k,h)} + \lambda_2 f_{2,(k,h)}$$

d'abord, calculer pour tous les k de 1 à n_1 et pour tous les h de 1 à n_2 :

$$f_{1,(k,h)} = t_h \left((x_k - t_h)^3 - (x_1 - t_h)^3 \right)$$

$$f_{2,(k,h)} = \left((x_k - t_h)^2 - (x_1 - t_h)^2 \right)$$

| k | $h=1$ | 2 | 3 | 4 | 5 | |
|-----|--------------|--------------|--------------|--------------|--------------|-----------------|
| 1 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | } $f_{1,(k,h)}$ |
| 2 | 5,78212E-02 | 4,85878E-01 | 1,26786E+00 | 4,97803E+00 | 6,16984E+00 | |
| 3 | 7,29591E-02 | 8,71512E-01 | 2,52727E+00 | 1,09731E+01 | 1,37618E+01 | |
| 4 | 7,71950E-02 | 9,40435E-01 | 2,87718E+00 | 1,31754E+01 | 1,66273E+01 | |
| 5 | 1,37313E-01 | 9,64362E-01 | 3,17551E+00 | 1,59578E+01 | 2,03674E+01 | |
| 6 | 2,43210E-01 | 9,70719E-01 | 3,23018E+00 | 1,71382E+01 | 2,20348E+01 | |
| 7 | 4,81214E-01 | 1,03412E+00 | 3,23743E+00 | 1,80648E+01 | 2,34288E+01 | |
| 8 | 8,53964E-01 | 1,21177E+00 | 3,26256E+00 | 1,85399E+01 | 2,42272E+01 | |
| 9 | 5,57218E+00 | 5,42900E+00 | 6,04975E+00 | 1,90702E+01 | 2,49185E+01 | |
| 10 | 1,39769E+01 | 1,51102E+01 | 1,47573E+01 | 2,31418E+01 | 2,79964E+01 | |
| 1 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | 0,00000E+00 | } $f_{2,(k,h)}$ |
| 2 | -1,56000E-01 | -3,44000E-01 | -4,84000E-01 | -7,80000E-01 | -8,40000E-01 | |
| 3 | -2,39200E-01 | -7,28000E-01 | -1,09200E+00 | -1,86160E+00 | -2,01760E+00 | |
| 4 | -2,04000E-01 | -8,43200E-01 | -1,31920E+00 | -2,32560E+00 | -2,52960E+00 | |
| 5 | -1,92000E-02 | -9,21600E-01 | -1,59360E+00 | -3,01440E+00 | -3,30240E+00 | |
| 6 | 1,82400E-01 | -8,89200E-01 | -1,68720E+00 | -3,37440E+00 | -3,71640E+00 | |
| 7 | 5,16800E-01 | -7,61600E-01 | -1,71360E+00 | -3,72640E+00 | -4,13440E+00 | |
| 8 | 9,26300E-01 | -5,49500E-01 | -1,64850E+00 | -3,97210E+00 | -4,44310E+00 | |
| 9 | 4,04480E+00 | 1,63840E+00 | -1,53600E-01 | -3,94240E+00 | -4,71040E+00 | |
| 10 | 7,71230E+00 | 4,60090E+00 | 2,28390E+00 | -2,61490E+00 | -3,60790E+00 | |

Ensuite, calculer les coefficients de l'équation matricielle :

$$\Sigma_{0,1} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} (y_{k,h} - y_{1,h}) S_{k,h}$$

$$\Sigma_{0,2} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} (y_{k,h} - y_{1,h}) f_{1,(k,h)}$$

$$\Sigma_{0,3} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} (y_{k,h} - y_{1,h}) f_{2,(k,h)}$$

$$\Sigma_{1,1} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} (S_{k,h})^2 \quad ; \quad \Sigma_{1,2} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} S_{k,h} f_{1,(k,h)}$$

$$\Sigma_{1,3} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} S_{k,h} f_{2,(k,h)} \quad ; \quad \Sigma_{2,2} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} (f_{1,(k,h)})^2$$

$$\Sigma_{2,3} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} f_{1,(k,h)} f_{2,(k,h)} \quad ; \quad \Sigma_{3,3} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} (f_{2,(k,h)})^2$$

Résoudre le système matriciel :

$$\begin{pmatrix} p_1 \\ \lambda_1 \\ \lambda_2 \end{pmatrix} = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} & \Sigma_{1,3} \\ \Sigma_{1,2} & \Sigma_{2,2} & \Sigma_{2,3} \\ \Sigma_{1,3} & \Sigma_{2,3} & \Sigma_{3,3} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{0,1} \\ \Sigma_{0,2} \\ \Sigma_{0,3} \end{pmatrix}$$

$$\begin{bmatrix} 1,22594\text{E}+03 & 2,90228\text{E}+03 & -3,34360\text{E}+02 \\ 2,90228\text{E}+03 & 7,18515\text{E}+03 & -8,75437\text{E}+02 \\ -3,34360\text{E}+02 & -8,75437\text{E}+02 & 3,14234\text{E}+02 \end{bmatrix}^{-1} \begin{bmatrix} -4,84631\text{E}+02 \\ -1,17329\text{E}+03 \\ 2,89415\text{E}+02 \end{bmatrix} = \begin{bmatrix} -4,68563\text{E}-01 \\ 1,17237\text{E}-01 \\ 7,49060\text{E}-01 \end{bmatrix}$$

On a ainsi obtenu une valeur approchée de p : $p_1 \simeq -0,468563$

Les valeurs approchées de b et λ seront ensuite calculées par régression linéaire correspondant à la relation : $y(x, t) = b \exp(p_1 x t) + \lambda (x - t)^2$

On calcule donc les coefficients :

$$\Sigma_{0,1} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} y_{k,h} \exp(p_1 x_k t_h) \quad ; \quad \Sigma_{0,2} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} y_{k,h} (x_k - t_h)^2$$

$$\Sigma_{1,1} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} (\exp(p_1 x_k t_h))^2 \quad ; \quad \Sigma_{2,2} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} ((x_k - t_h)^2)^2$$

$$\Sigma_{1,2} = \sum_{k=1}^{n_1} \sum_{h=1}^{n_2} \exp(p_1 x_k t_h) (x_k - t_h)^2$$

Puis on résout le système matriciel :
$$\begin{pmatrix} b \\ \lambda \end{pmatrix} = \begin{pmatrix} \Sigma_{1,1} & \Sigma_{1,2} \\ \Sigma_{1,2} & \Sigma_{2,2} \end{pmatrix}^{-1} \begin{pmatrix} \Sigma_{0,1} \\ \Sigma_{0,2} \end{pmatrix}$$

$$\begin{bmatrix} 1,56251\text{E}+01 & 3,34931\text{E}+01 \\ 3,34931\text{E}+01 & 2,47632\text{E}+02 \end{bmatrix}^{-1} \begin{bmatrix} 5,55056\text{E}+01 \\ 2,48939\text{E}+02 \end{bmatrix} = \begin{bmatrix} 1,96806\text{E}+00 \\ 7,39092\text{E}-01 \end{bmatrix}$$

On obtient ainsi les valeurs approchées de b et λ reportées dans l'équation $y(x, t)$:

$$y(x, t) \simeq b_1 \exp(p_1 x t) + \lambda_1 (x - t)^2 \quad \begin{cases} p_1 \simeq -0,468563 \\ b_1 \simeq 1,96806 \\ \lambda_1 \simeq 0,739092 \end{cases}$$

Cette équation permet de calculer les valeurs approchées $y_{\text{calculé}}(x_k, t_h)$ et de les comparer aux valeurs initialement données (tableau page suivante).

Ecart quadratique moyen obtenu : $\text{eqm} = 0,0596$

La représentation graphique des courbes tracées à t constant permet de visualiser la proximité de ce réseau de courbes avec les points donnés (figure page suivante).

| k | h= 1 | | 2 | | 3 | | 4 | | 5 | | eqm = 0,0596 |
|----|--------------------|----------------------|--------------------|----------------------|--------------------|----------------------|--------------------|----------------------|--------------------|----------------------|--------------|
| | y _{donné} | y _{calculé} | |
| 1 | 2,05 | 2,072575 | 2,59 | 2,52278 | 3,07 | 3,071141 | 4,82 | 4,829362 | 5,31 | 5,284859 | |
| 2 | 1,91 | 1,850305 | 2,17 | 2,089704 | 2,54 | 2,485842 | 3,97 | 3,934439 | 4,41 | 4,329055 | |
| 3 | 1,57 | 1,630074 | 1,59 | 1,555179 | 1,72 | 1,730577 | 2,81 | 2,743321 | 2,98 | 3,053948 | |
| 4 | 1,63 | 1,582086 | 1,29 | 1,359232 | 1,32 | 1,432762 | 2,24 | 2,247617 | 2,43 | 2,520398 | |
| 5 | 1,53 | 1,597158 | 1,2 | 1,12789 | 1,08 | 1,033804 | 1,57 | 1,524506 | 1,71 | 1,735148 | |
| 6 | 1,76 | 1,673106 | 1,01 | 1,052771 | 0,92 | 0,856747 | 1,07 | 1,149871 | 1,27 | 1,322346 | |
| 7 | 1,77 | 1,836005 | 0,97 | 1,037737 | 0,67 | 0,721994 | 0,81 | 0,781466 | 0,85 | 0,908375 | |
| 8 | 2,04 | 2,063111 | 0,98 | 1,100984 | 0,61 | 0,674969 | 0,39 | 0,516935 | 0,62 | 0,600954 | |
| 9 | 4,05 | 4,067781 | 2,42 | 2,39024 | 1,54 | 1,475298 | 0,25 | 0,318336 | 0,19 | 0,200162 | |
| 10 | 6,58 | 6,602034 | 4,45 | 4,421274 | 3,15 | 3,149569 | 1,15 | 1,231996 | 0,92 | 0,95666 | |

