La droite de Williamson : une technique de régression linéaire injustement oubliée

François Auger, Jean-Christophe Olivier

UNIVERSITÉ DE NANTES

Université de Nantes, IUT de Saint-Nazaire

iUT de Saint-Nazaire
UNIVERSITE DE NANTES

Institut de Recherche en Électrotechnique et Électronique de Nantes-Atlantique

Colloque C2I 2010, 27 janvier 2010, Le Mans

IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

Introduction

- La régression linéaire est un problème fréquent en physique expérimentale, en contrôle-qualité et en métrologie ;
- Paradoxalement, l'outil généralement utilisé (et enseigné) est basé sur des hypothèses qui ne sont pas toujours vérifiées en pratique. Les estimations obtenues présentent alors un fort biais, surtout quand le nombre de points de mesure est faible (2-10) ;
- Des méthodes mieux adaptées existent, mais leur compréhension est plus difficile. L'une d'entre elles, la méthode de Williamson, est basée sur des principes rigoureux. Les équations proposées n'étant pas beaucoup plus compliquées, elle mériterait d'être davantage utilisée.

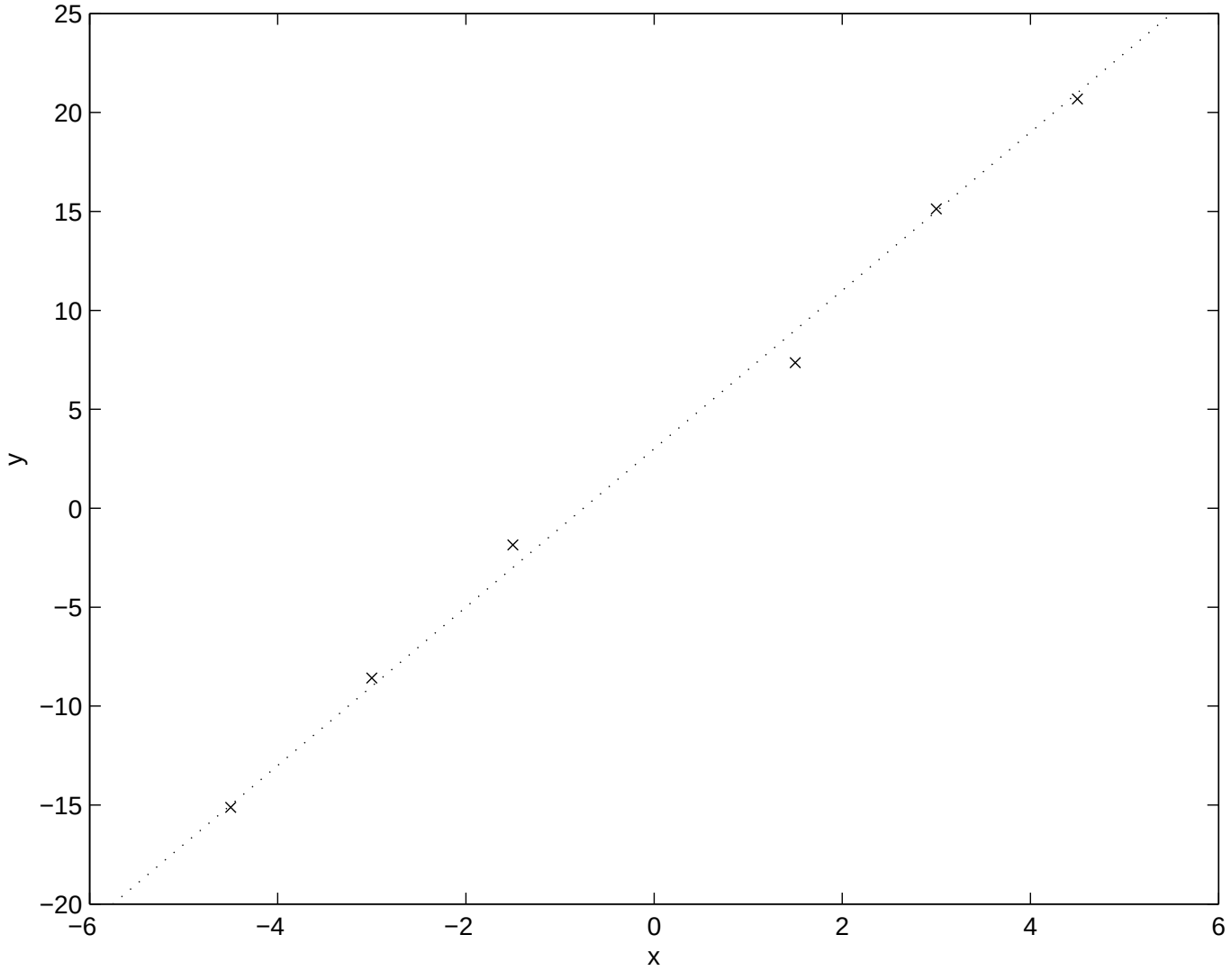
Sommaire

IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La régression linéaire classique : les hypothèses

- Deux grandeurs physiques $\mathcal{X}_i$ et $\mathcal{Y}_i$, liées par une relation linéaire de la forme $\mathcal{Y}_i = a\,\mathcal{X}_i + b$;
- Les deux grandeurs physiques sont supposées être de natures **différentes** ;
- L'abscisse, ou *"variable explicative"*, est supposée **déterministe** et parfaitement connue : $X_i = \mathcal{X}_i$;
- L'ordonnée, ou *"variable expliquée"*, est supposée **aléatoire** et entachée d'un bruit de mesure additif w_i : $Y_i = \mathcal{Y}_i + w_i$.
- les bruits de mesure w_i sont des variables aléatoires gaussiennes de valeurs moyennes nulles et de **même** variance σ_Y^2

La régression linéaire classique : les hypothèses



IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La régression linéaire classique : les résultats

a et *b* sont supposés déterministes (théorie de l'estimation de Fisher)

Estimateur du maximum de vraisemblance

$$\hat{a} = \frac{N S_{XY} - S_X S_Y}{N S_{XX} - S_X^2}, \quad \hat{b} = \frac{S_{XX} S_Y - S_X S_{XY}}{N S_{XX} - S_X^2} = \frac{S_Y - \hat{a} S_X}{N}, \quad \hat{y}_i = \hat{a} X_i + \hat{b}$$

$$\text{avec} \quad S_X = \sum_{i=1}^{N} X_i, \quad S_Y = \sum_{i=1}^{N} Y_i,$$

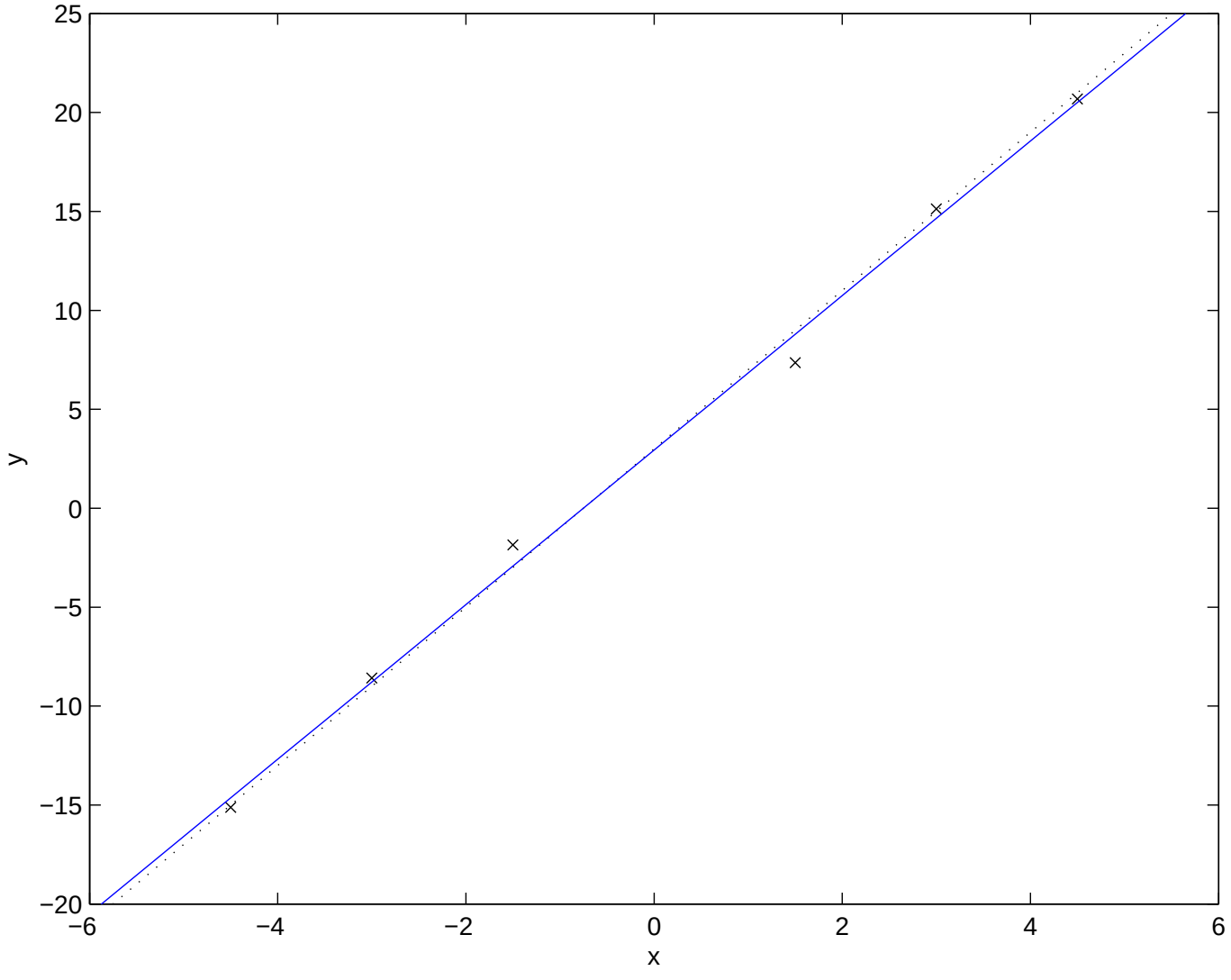
$$S_{XX} = \sum_{i=1}^{N} X_i^2, \quad S_{YY} = \sum_{i=1}^{N} Y_i^2, \quad \text{et} \quad S_{XY} = \sum_{i=1}^{N} X_i Y_i$$

Possibilité de calculer les variances d'erreur d'estimation

$$\sigma_{\hat{a}}^2 = E\left[(\hat{a} - a)^2\right] = \frac{N \sigma_Y^2}{N S_{XX} - S_X^2} \quad \text{et} \quad \sigma_{\hat{b}}^2 = E\left[(\hat{b} - b)^2\right] = \frac{S_{XX} \sigma_Y^2}{N S_{XX} - S_X^2}$$

IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La régression linéaire classique : les résultats



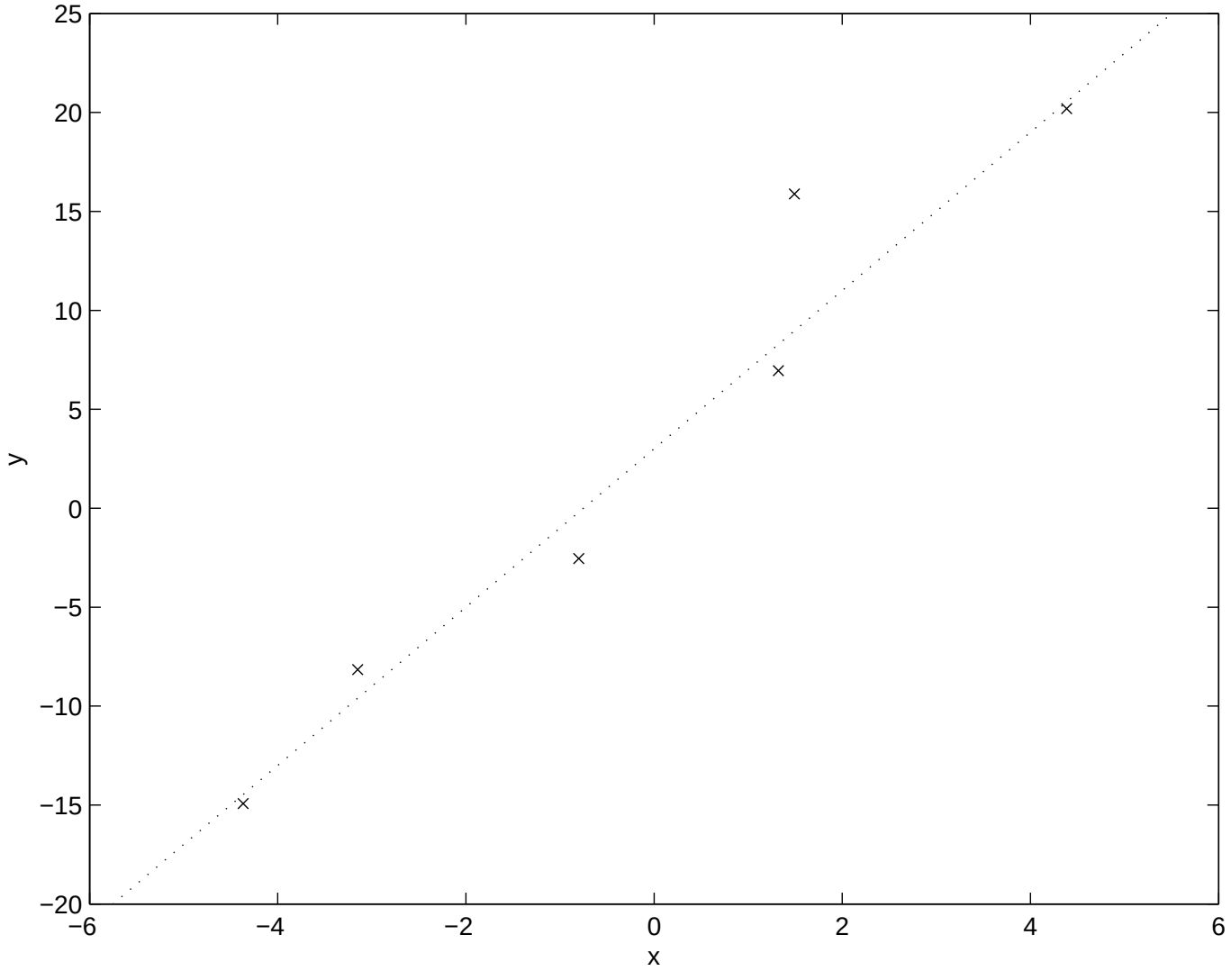
IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La méthode de Williamson : les hypothèses

- Deux grandeurs physiques $\mathcal{X}_i$ et $\mathcal{Y}_i$, liées par une relation linéaire de la forme $\mathcal{Y}_i = a\,\mathcal{X}_i + b$;
- Les deux grandeurs physiques sont supposées être de **même** nature ;
- L'abscisse, ou *"variable explicative"*, est supposée **aléatoire** et entachée d'un bruit de mesure additif v_i : $X_i = \mathcal{X}_i + v_i$;
- L'ordonnée, ou *"variable expliquée"*, est supposée **aléatoire** et entachée d'un bruit de mesure additif w_i : $Y_i = \mathcal{Y}_i + w_i$.
- les bruits de mesure v_i et w_i sont des variables aléatoires gaussiennes **indépendantes** de valeurs moyennes nulles et de variances respectives $\sigma^2_{X_i}$ et $\sigma^2_{Y_i}$

IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La méthode de Williamson : les hypothèses



IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La méthode de Williamson : la démarche

densité de probabilité conjointe de v_i et w_i

$$\rho(v_i, w_i) = \rho(v_i)\,\rho(w_i) = \frac{1}{2\pi\sigma_{X_i}\sigma_{Y_i}}\, e^{-\frac{1}{2}\left(\frac{v_i^2}{\sigma_{X_i}^2}+\frac{w_i^2}{\sigma_{Y_i}^2}\right)}$$

densité de probabilité conjointe des vecteurs des bruits de mesure

$$\rho(V, W) = \Pi_{i=1}^{N}\rho(v_i, w_i) = \frac{1}{(2\pi)^N\Pi_{i=1}^{N}\sigma_{X_i}\sigma_{Y_i}}\, e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{v_i^2}{\sigma_{X_i}^2}+\frac{w_i^2}{\sigma_{Y_i}^2}\right)}$$

densité de probabilité conjointe des vecteurs de mesure

$v_i = X_i - \mathcal{X}_i$ et $w_i = Y_i - \mathcal{Y}_i = Y_i - a\,\mathcal{X}_i - b$, donc

$$\rho(X, Y; a, b, \mathcal{X}) = \frac{1}{(2\pi)^N\Pi_{i=1}^{N}\sigma_{X_i}\sigma_{Y_i}}\, e^{-\frac{1}{2}\sum_{i=1}^{N}\left(\frac{(X_i-\mathcal{X}_i)^2}{\sigma_{X_i}^2}+\frac{(Y_i-a\,\mathcal{X}_i-b)^2}{\sigma_{Y_i}^2}\right)}$$

La méthode de Williamson : les résultats (1/2)

a, b et tous les $\mathcal{X}_i$ sont supposés déterministes (théorie de l'estimation de Fisher)

estimateurs du maximum de vraisemblance des $\mathcal{X}_i$

$$\frac{\partial}{\partial \mathcal{X}_i}\left[\ln(\rho(X, Y; a, b, \mathcal{X}))\right] = 0 \Longleftrightarrow \begin{cases} \hat{\mathcal{X}}_i = X_i + \frac{\hat{a}\,\lambda_i^2}{\hat{a}^2\,\lambda_i^2+1}\left(Y_i - \hat{a}\,X_i - \hat{b}\right) \\ \hat{\mathcal{Y}}_i = Y_i - \frac{1}{\hat{a}^2\,\lambda_i^2+1}\left(Y_i - \hat{a}\,X_i - \hat{b}\right) \end{cases},$$

avec $\lambda_i = \sigma_{X_i}/\sigma_{Y_i}$

estimateur du maximum de vraisemblance de $\hat{b}$

$$\frac{\partial}{\partial b}\left[\ln(\rho(X, Y; a, b, \mathcal{X}))\right] = 0 \Longleftrightarrow \sum_{i=1}^{N} \frac{Y_i - \hat{\mathcal{Y}}_i}{\sigma_{Y_i}^2} = \sum_{i=1}^{N} \frac{Y_i - \hat{a}\,X_i - \hat{b}}{\hat{a}^2\,\lambda_i^2 + 1} = 0$$

Si les λ_i sont constants, cette équation se ramène simplement à $\hat{b} = (S_Y - \hat{a}\,S_X)/N$

IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La méthode de Williamson : les résultats (2/2)

estimateur du maximum de vraisemblance de $\hat{a}$

$$\frac{\partial}{\partial a}\left[\ln(\rho(X, Y; a, b, \mathcal{X}))\right] = 0 \Longleftrightarrow \sum_{i=1}^{N} \frac{\hat{\mathcal{X}}_i \left(Y_i - \hat{\mathcal{Y}}_i\right)}{\sigma_{Y_i}^2} = 0$$

Dans le cas général, la solution de cette équation n'a pas d'expression simple, et doit être approchée par des algorithmes numériques.
Dans le cas où $\sigma_{X_i} = \sigma_X$ et $\sigma_{Y_i} = \sigma_Y$, elle se réduit par contre à une équation du second degré

$$\lambda^2 S_{xy}\, \hat{a}^2 + \left(S_{xx} - \lambda^2 S_{yy}\right) \hat{a} - S_{xy} = 0$$

avec $S_{xx} = S_{XX} - S_X^2/N$, $S_{xy} = S_{XY} - S_X S_Y/N$ et $S_{yy} = S_{YY} - S_Y^2/N$.
Puisque a doit être de même signe que S_{xy}, on retient

$$\hat{a} = \frac{-\left(S_{xx} - \lambda^2 S_{yy}\right) + \sqrt{\left(S_{xx} - \lambda^2 S_{yy}\right)^2 + 4\lambda^2 S_{xy}^2}}{2\lambda^2 S_{xy}}$$

IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La méthode de Williamson : en résumé

estimation des moments statistiques

$$S_X = \sum_{i=1}^{N} X_i, \quad S_Y = \sum_{i=1}^{N} Y_i, \quad S_{XX} = \sum_{i=1}^{N} X_i^2, \quad S_{YY} = \sum_{i=1}^{N} Y_i^2, \quad \text{et} \quad S_{XY} = \sum_{i=1}^{N} X_i\, Y_i$$

$$S_{xx} = S_{XX} - S_X^2/N, \quad S_{xy} = S_{XY} - S_X S_Y/N \quad \text{et} \quad S_{yy} = S_{YY} - S_Y^2/N$$

au lieu de

$$\hat{a} = \frac{S_{xy}}{S_{xx}}$$

$$\hat{b} = (S_Y - \hat{a}\, S_X)/N$$

$$\hat{y}_i = \hat{a}\, X_i + \hat{b}$$

utiliser (si $\sigma_{X_i} = \sigma_X$ et $\sigma_{Y_i} = \sigma_Y$)

$$\hat{a} = \frac{-(S_{xx} - \lambda^2 S_{yy}) + \sqrt{(S_{xx} - \lambda^2 S_{yy})^2 + 4\lambda^2 S_{xy}^2}}{2\lambda^2 S_{xy}}$$

$$\hat{b} = (S_Y - \hat{a}\, S_X)/N$$

$$\hat{y}_i = Y_i - \frac{1}{\hat{a}^2\, \lambda^2 + 1} \left(Y_i - \hat{a}\, X_i - \hat{b} \right)$$

$$\hat{x}_i = X_i + \frac{\hat{a}\, \lambda_i^2}{\hat{a}^2\, \lambda_i^2 + 1} \left(Y_i - \hat{a}\, X_i - \hat{b} \right)$$

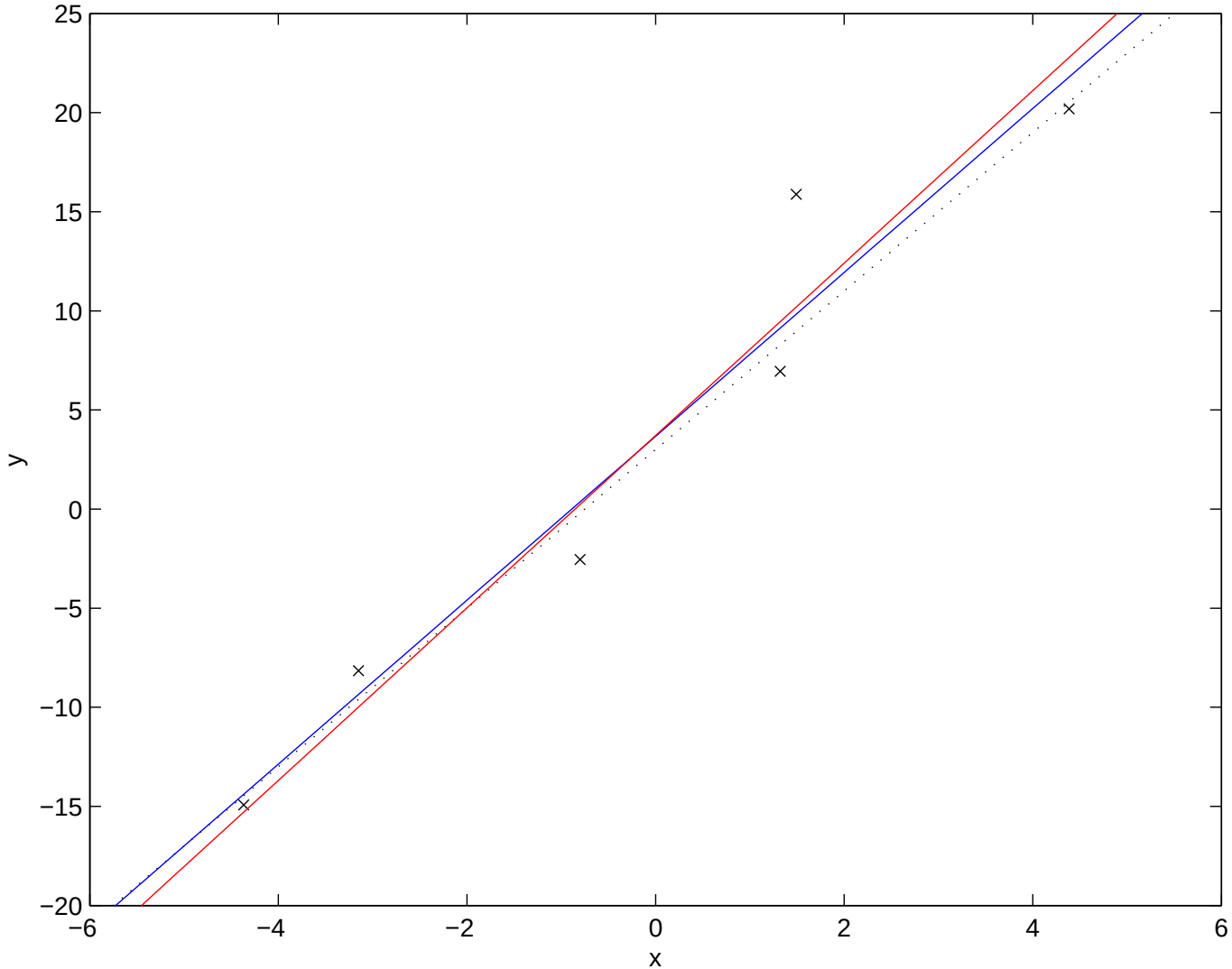
avec $\lambda = \sigma_X / \sigma_Y$

La méthode de Williamson : en résumé

```
function [ahat,bhat]=WilliamsonStraightLineFit(X,Y,lambda);
% [a,b]=WilliamsonStraightLineFit(X,Y,lambda)
% find the coefficients of the straightline which best fits data
% using the Williamson approach. X and Y are the data, lambda is
% the ratio of the standard deviations sigmaX/sigmaY
%
% For lambda=0, the classical straight line fit is delivered

% F. Auger, IREENA, francois.auger@univ-nantes.fr, jan 2010

if nargin==0,
 help WilliamsonStraightLineFit; return;
elseif (nargin==1),
 error('at least 2 parameters required');
elseif (nargin==2),
 lambda=0;
end;

NbPoints=length(X); NbPointsY=length(Y);
if (NbPoints~=NbPointsY)
 error('X and Y must have the same length');
end;

Xbar=sum(X)/NbPoints; x=X-Xbar; Ybar=sum(Y)/NbPoints; y=Y-Ybar;
Sxx=sum(x.*x); Sxy=sum(y.*x); Syy=sum(y.*y);

if (lambda==0)
 ahat=Sxy/Sxx; bhat=Ybar-ahat*Xbar;
else
 Delta=(Sxx-lambda^2*Syy)^2+4*Sxy^2*lambda^2;
 ahat=(-(Sxx-lambda^2*Syy)+sqrt(Delta))/(2*lambda^2*Sxy);
 bhat=Ybar-ahat*Xbar;
end;
```

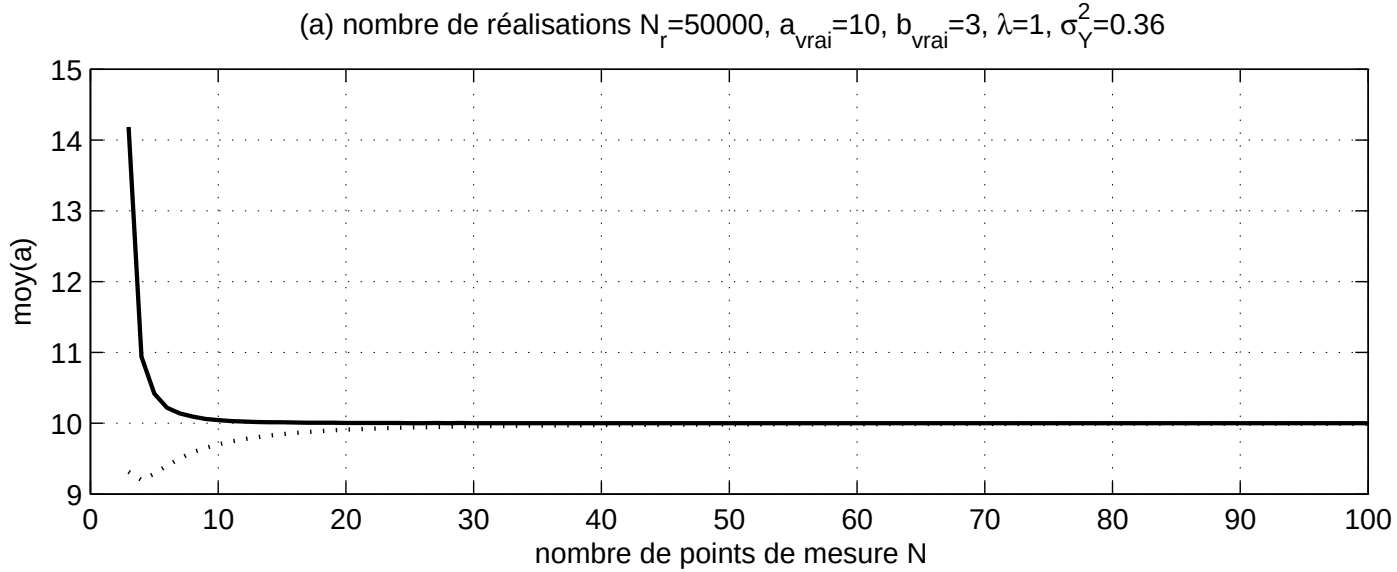
IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

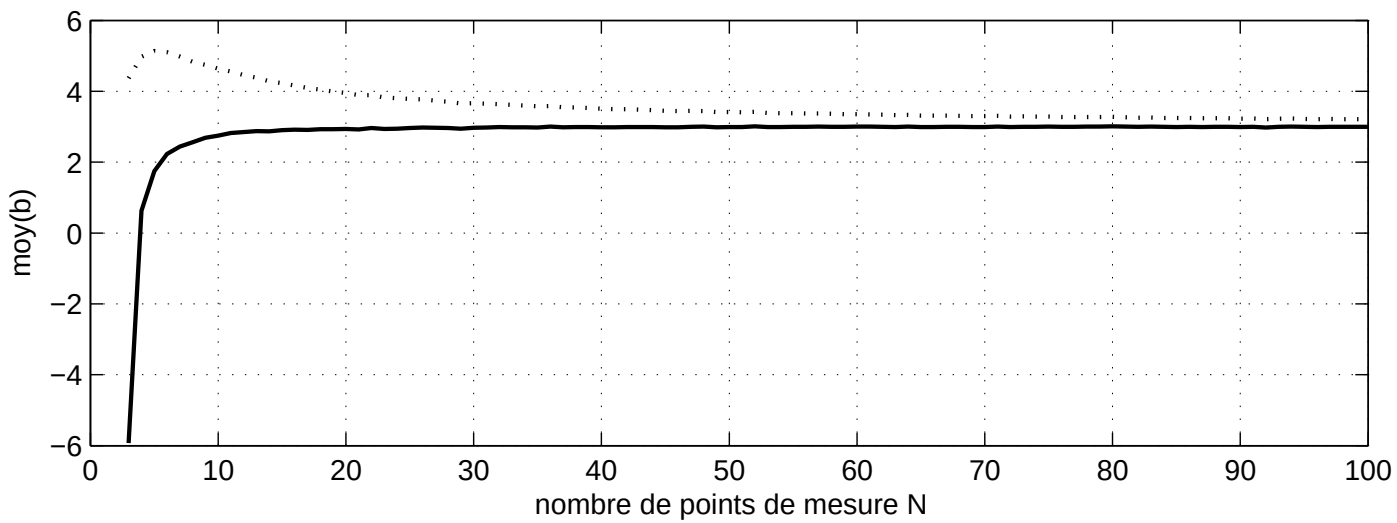
La méthode de Williamson : en résumé



IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

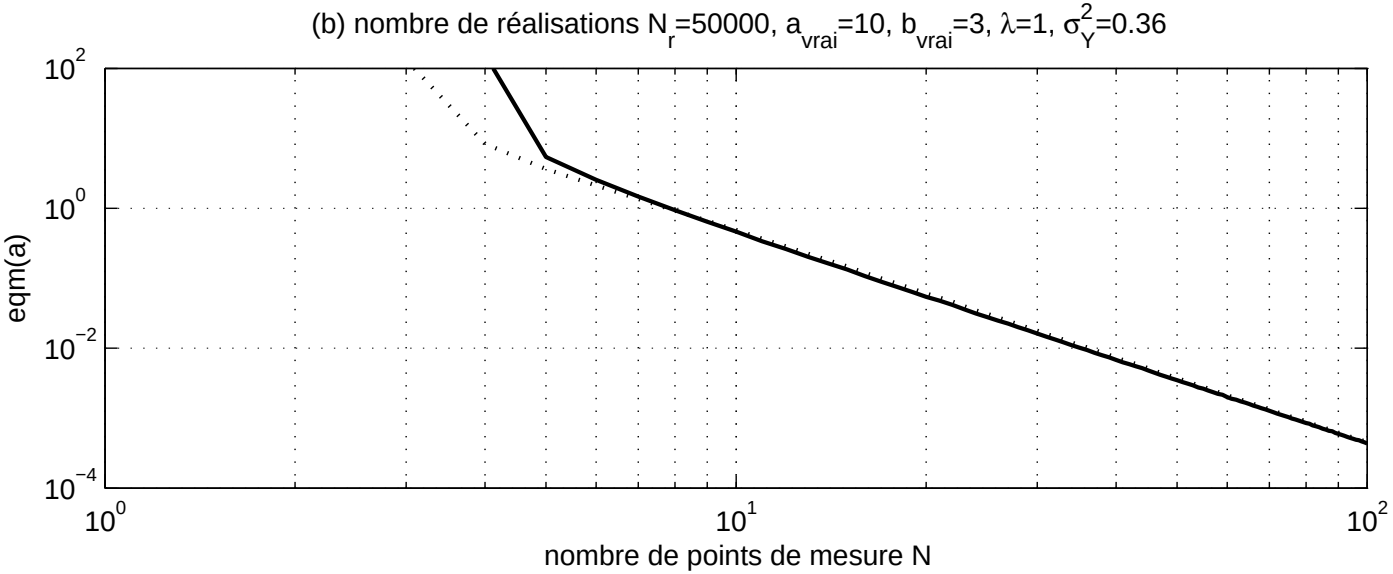
La méthode de Williamson : performances (1/5)

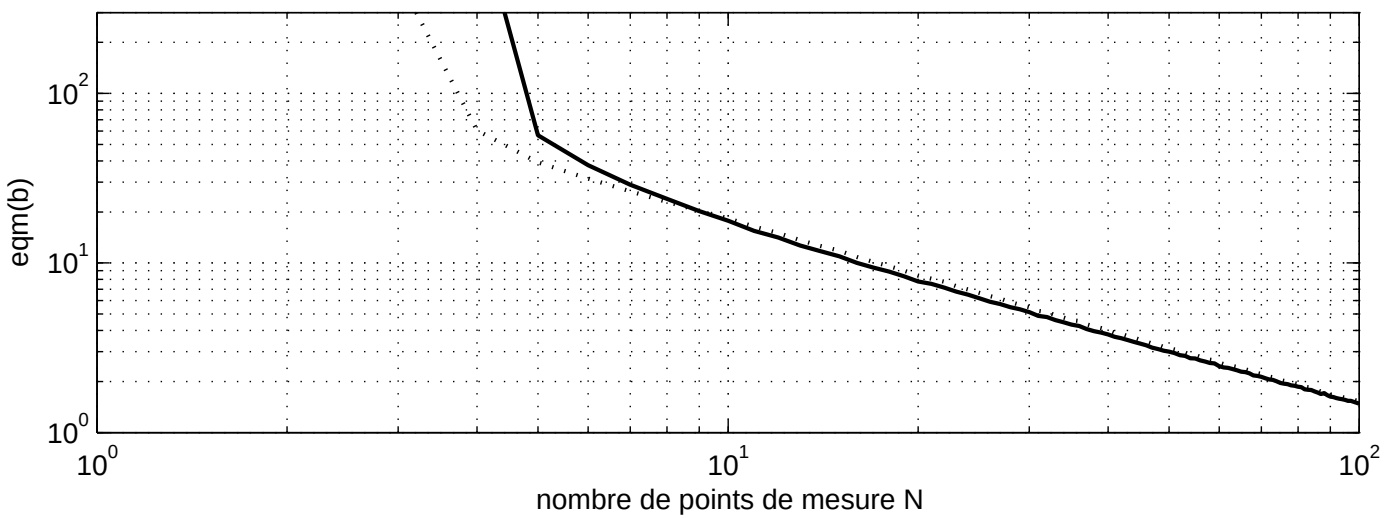




IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

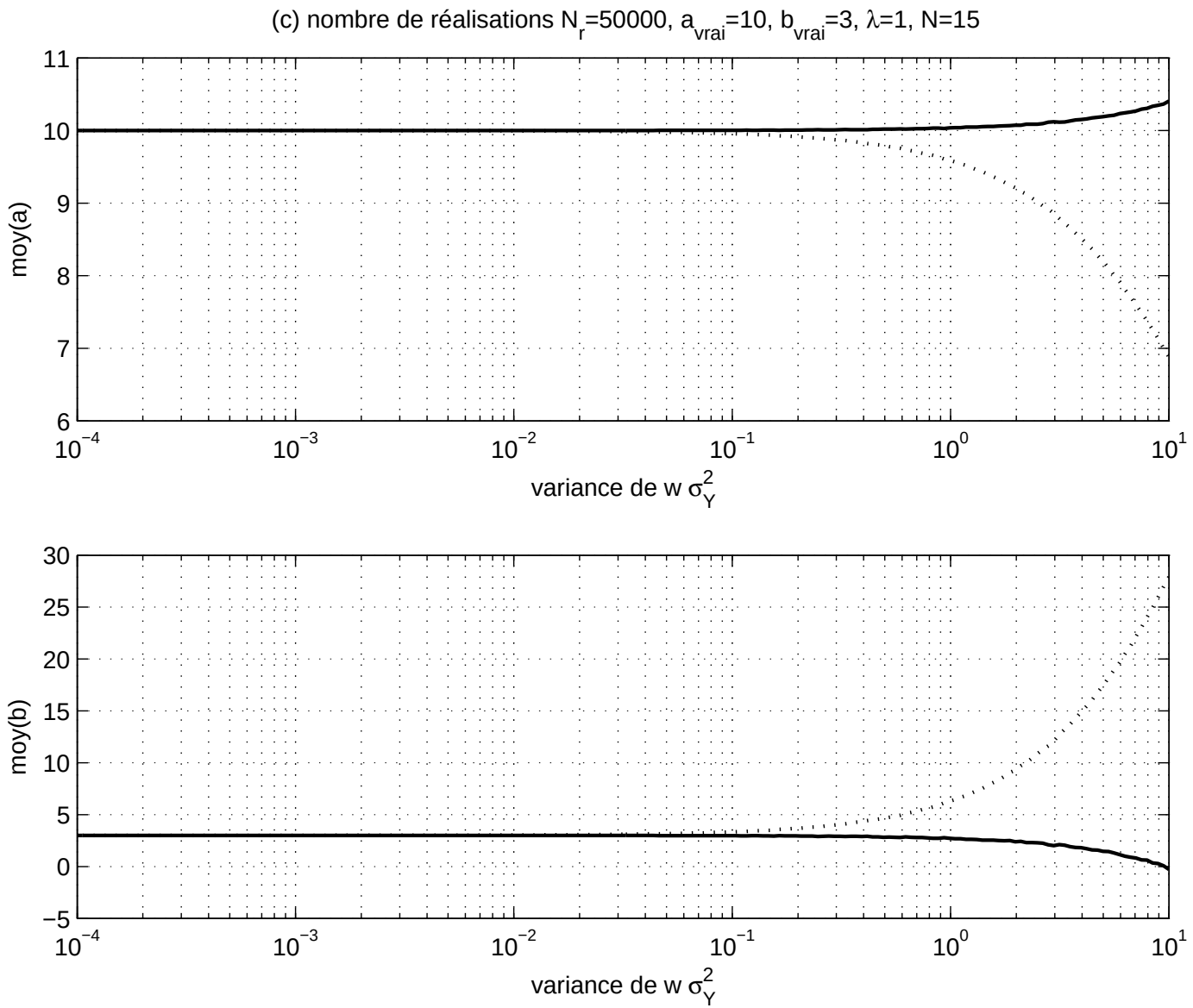
La méthode de Williamson : performances (2/5)





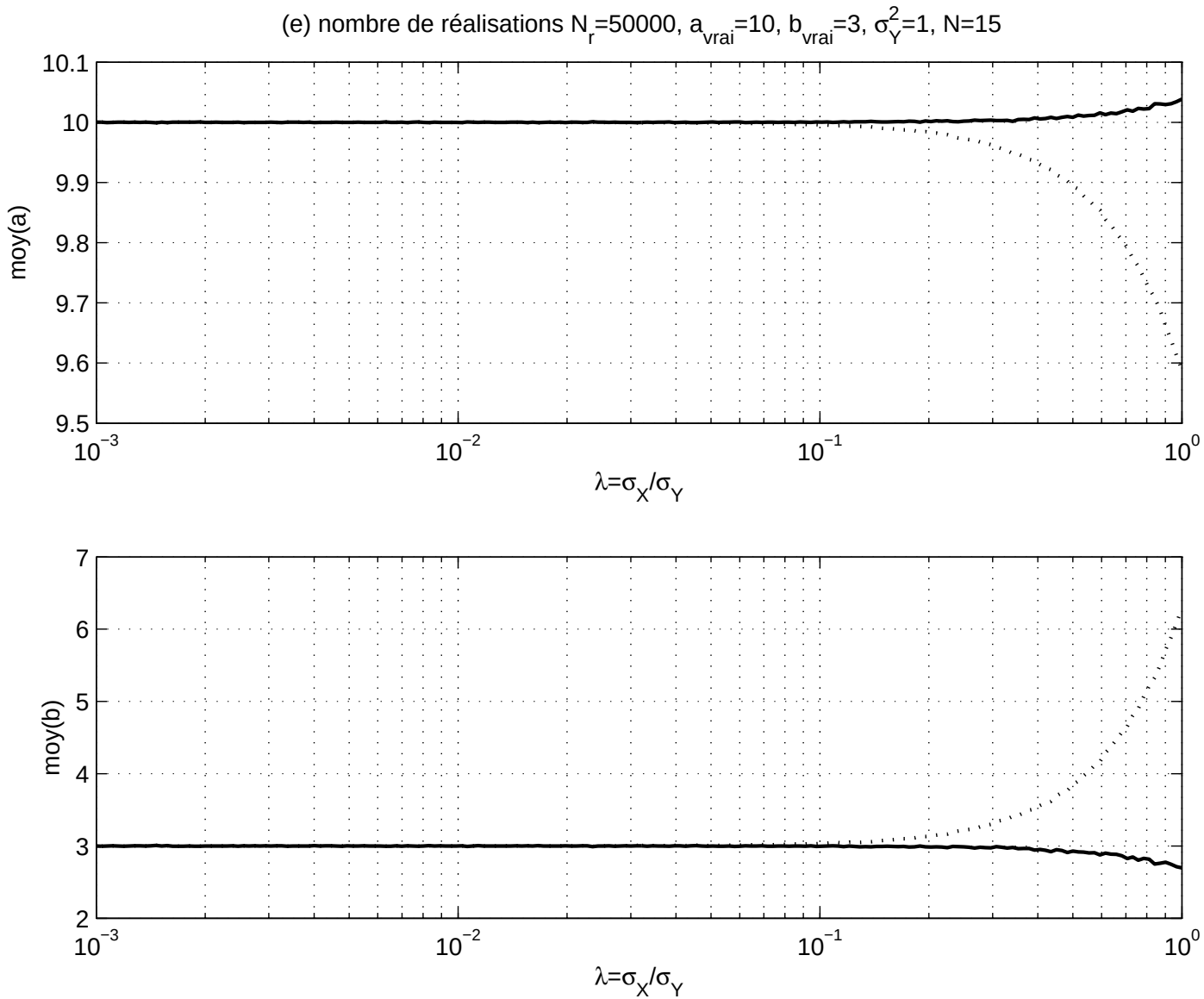
IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La méthode de Williamson : performances (3/5)



IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

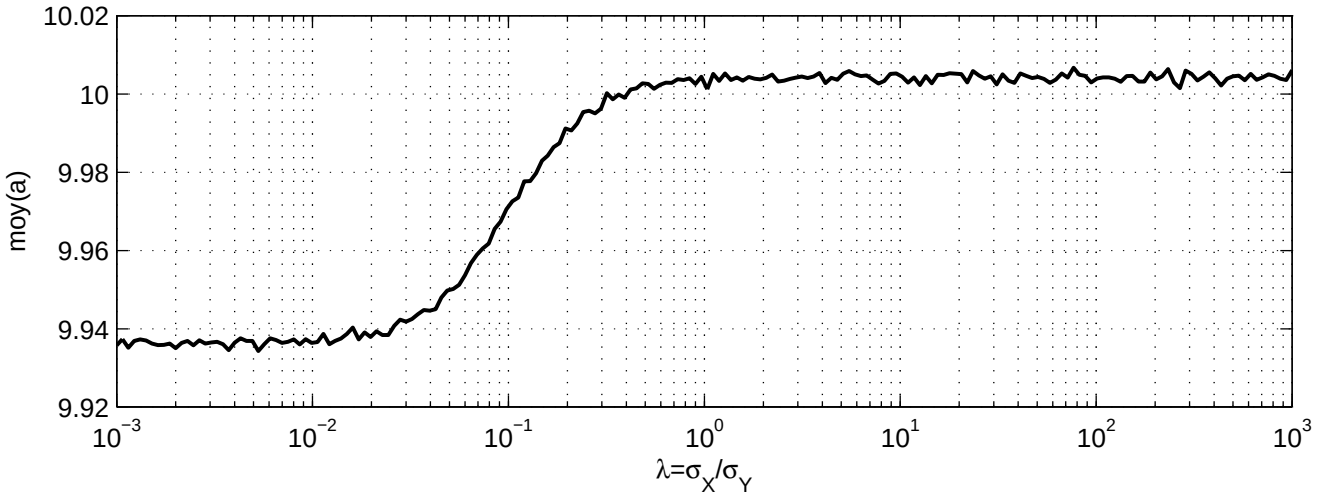
La méthode de Williamson : performances (4/5)



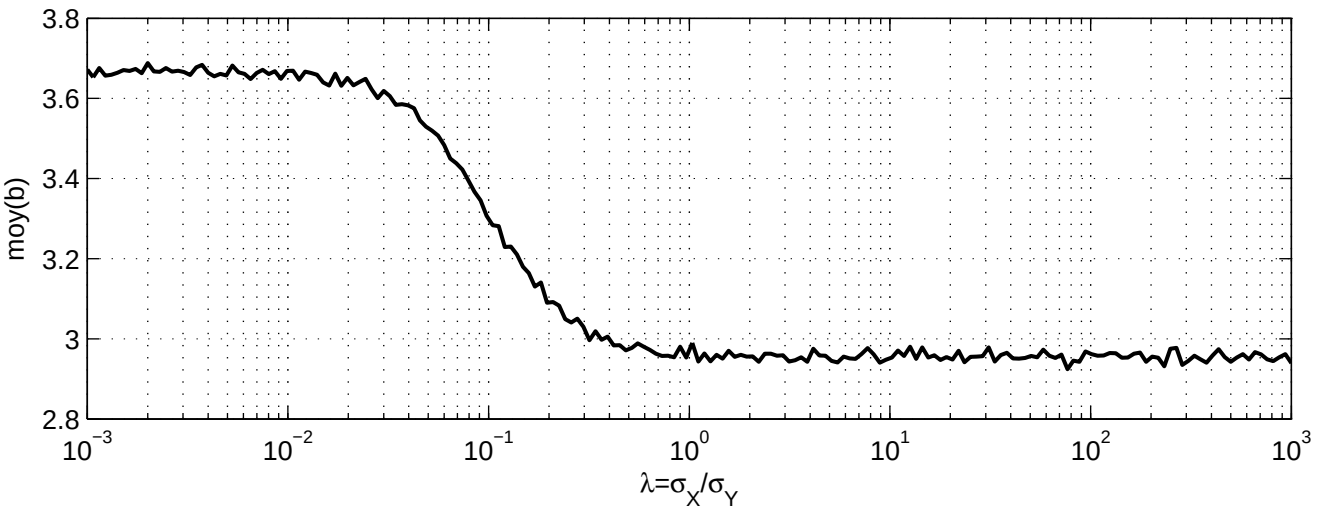
IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

La méthode de Williamson : performances (5/5)

(g) nombre de réalisations N_r=50000, a_{vrai}=10, b_{vrai}=3, σ_Y^2=0.25, λ_{vrai}=1, N=20



(h) nombre de réalisations N_r=50000, a_{vrai}=10, b_{vrai}=3, σ_Y^2=0.25, λ_{vrai}=1, N=20



IREENA
Institut de Recherche en Electrotechnique et Electronique de Nantes-Atlantique

Linéarité d'un capteur de courant

Vérification d'une carte de mesure de courants triphasés

X_i mesure obtenue avec une pince `Tektronix A 6303`
Y_i mesure obtenue avec un capetur `LEM LAH 100-P`

X_i (A)	-79.900	-60.440	-42.600	-23.100	-0.118
Y_i (A)	-79.940	-60.710	-43.350	-23.410	-0.281
X_i (A)	0.286	20.580	37.280	60.580	79.460
Y_i (A)	0.013	20.460	37.580	60.750	80.290

méthode	$\hat{a}$	$\hat{b}$
régression classique	1.00591624	-0.05788357
Williamson ($\lambda = 3/16$)	1.00591733	-0.05788270

Dans une procédure de vérification, $\lambda \ll 1$, donc la méthode de Williamson n'apporte qu'une faible amélioration, et les valeurs "vraies" de a et b sont inconnues. Cet exemple montre principalement la possibilité d'utiliser la méthode de Williamson dans des cas pratiques.

Construction d'estimateurs récursifs

Dans certains cas, il peut être intéressant de disposer des estimations de a et b au fur et à mesure que de nouveaux points de mesure sont apportés.

Estimateurs récursifs des moments statistiques

- fenêtre de taille croissante : $S_X[n+1] = S_X[n] + X_{n+1}$.
- fenêtre glissante de taille constante N : $S_X[n+1] = S_X[n] + X_{n+1} - X_{n-N+1}$.
- facteur d'oubli : $S_X[n+1] = \alpha\, S_X[n] + (1-\alpha)\, X_{n+1}$, avec $0 \leq \alpha \leq 1$.

Estimateurs récursifs bayésiens déduits d'un modèle d'état

$$x_1[n] = a, \quad x_2[n] = b/a \quad \text{et} \quad x_3[n] = \mathcal{X}_n + b/a = X_n + b/a - v_n$$

$$\begin{aligned} X_{[n+1]} &= A\, X_{[n]} + B\, X_{n+1} + V_{[n+1]}, \\ Y_{[n+1]} &= Y_{n+1} = \mathcal{H}(X_{[n+1]}) + w_{[n+1]}, \end{aligned}$$

$$A = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ 0 & 1 & 0 \end{pmatrix}, \quad X = \begin{pmatrix} x_1 \\ x_2 \\ x_3 \end{pmatrix}, \quad B = \begin{pmatrix} 0 \\ 0 \\ 1 \end{pmatrix} \quad \text{et} \quad \mathcal{H}(X) = x_1\, x_3$$

Possibilité de construire des estimateurs de Kalman à partir de ce modèle

Conclusion

- La méthode de Williamson fournit des estimations plus pertinentes que l'estimateur basique lorsque le nombre de points de mesure est faible et lorsque les données en X sont significativement bruitées ;
- Le surcoût de calcul est faible ;
- Il faudrait établir des valeurs approchées des erreurs d'estimation et déterminer les points qui contribuent le plus à la détermination des coefficients de la droite de régression ;
- Un outil logiciel pourrait être développé à partir de cette contribution.
- `http://www.univ-nantes.fr/auger-f`

retour