Université Nice SOPHIA ANTIPOLIS

I3S

CENTRE NATIONAL
DE LA RECHERCHE
SCIENTIFIQUE

ANALYSE CRITIQUE

Luc Pronzato, 2004

1) Inspection des résultats

- ☞ Paramètres estimés $\hat{\theta}$ «réalistes» ? (contraintes respectées ?)
- ☞ Recommencer avec d'autres conditions initiales, perturbations des quantités supposées connues $\rightarrow$ robustesse ? (stabilité des résultats ?)
- ☞ Prédire les valeurs d'observations correspondant à des données non utilisées pour l'identification (données de validation)
- ☞ Analyse graphique des résidus $e(\hat{\theta},k)$, $k = 1,\ldots,N$ $\rightarrow$ doivent ressembler à des erreurs (que l'on a souvent supposées de moyenne nulle, variance constante, etc.)

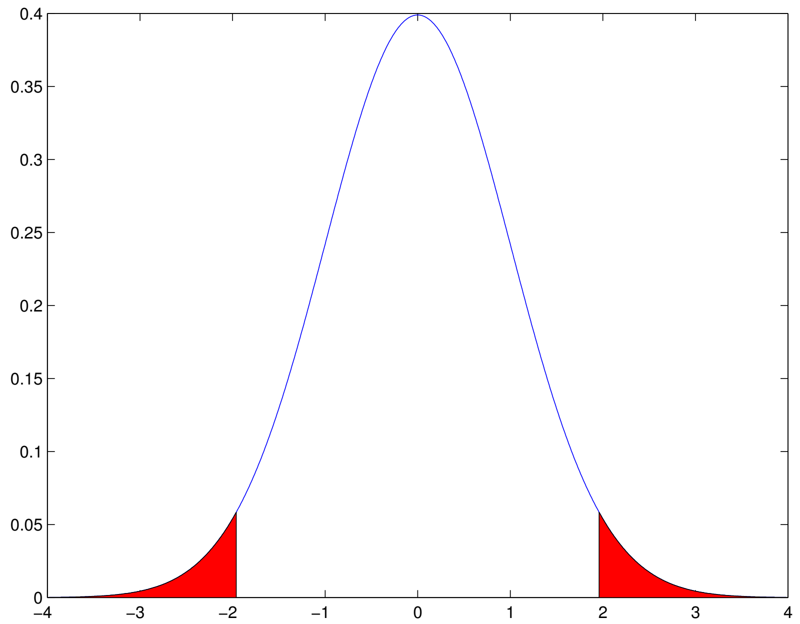
2) Analyse statistique des résidus

Ex1: On a supposé que $(e(\bar{\theta},k))_k = (\epsilon_k)_k$ est i.i.d. $\mathcal{N}(0,\sigma^2)$
On dispose de résidus $e(\hat{\theta},k)$, $k = 1,\ldots,N$, et on veut tester l'hypothèse H_0 que leur moyenne est nulle

Moyenne estimée : $\hat{m} = \frac{1}{N}\sum_{k=1}^{N} e(\hat{\theta},k)$
Variance estimée : $\hat{v} = \frac{1}{N-1}\sum_{k=1}^{N}[e(\hat{\theta},k) - \hat{m}]^2$

Sous l'hypothèse que les erreurs $e(\hat{\theta},k)$, $k = 1,\ldots,N$ forment un échantillon de taille N de $(\epsilon_k)_k$, la variable $t = \sqrt{N}\frac{\hat{m}}{\sqrt{\hat{v}}}$
est distribuée suivant une loi de Student à $N-1$ degrés de liberté
(exact si les erreurs sont normales, $\approx$ si N assez grand et erreurs indépendantes)
On rejette l'hypothèse de moyenne nulle si $t > t_{N-1}(\alpha)$, par exemple $\alpha = 95\%$ ($\rightarrow$ valeur tabulée)

Pour N grand, $t \sim \mathcal{N}(0,1)$



On rejette l'hypothèse de moyenne nulle si $t > 2$
c'est-à-dire si $\hat{m} > 2\frac{\sqrt{\hat{v}}}{\sqrt{N}}$ (sous H_0, probabilité $< 5\%$)

Quand on décide de rejeter H_0 : introduire un terme constant θ_0 dans la réponse, estimer θ_0, tester de nouveau

Hypothèse d'indépendance essentielle : voir + loin comment la tester ...

3) Test de normalité

A partir de $\hat{m}$ et $\hat{v}$ $\rightarrow$ résidus normalisés

$$e_n(\hat{\theta},k) = \frac{e(\hat{\theta},k) - \hat{m}}{\sqrt{\hat{v}}}, \ \ k = 1, \dots ,N$$

On les classe par valeurs croissantes : $e_n(\hat{\theta},k_1) \leq e_n(\hat{\theta},k_2) \leq \cdots \leq e_n(\hat{\theta},k_N)$, et on calcule la fonction de répartition empirique $F_e(\cdot)$

$$F_e(x) = \begin{cases} 0 & \text{si } x < e_n(\hat{\theta},k_1) \\ i/N & \text{si } e_n(\hat{\theta},k_i) \leq x < e_n(\hat{\theta},k_{i+1}) \\ 1 & \text{si } e_n(\hat{\theta},k_N) \leq x \end{cases}$$

que l'on peut comparer à sa valeur théorique [pour $\mathcal{N}(0,1)$]
$F(x) = \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{x} \exp\left(-\frac{u^2}{2}\right) du$

On peut également tracer un *pp-plot* : $F([e_n(\hat{\theta},k_i)]$ en fonction de i/N

$\rightarrow$ diagonale principale si distribution normale

Ex2 : 4 jeux de données

- (i) 100 variables i.i.d. $\mathcal{N}(0,1)$
- (ii) 100 variables i.i.d. uniformes dans $[-1,1]$
- (iii) $\mathbf{y} = \mathbf{R}\bar{\theta} + \epsilon$, avec $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{I}_{100})$, $\mathbf{R} = \begin{pmatrix} \mathbf{R}_1 \\ \mathbf{R}_2 \end{pmatrix}$,

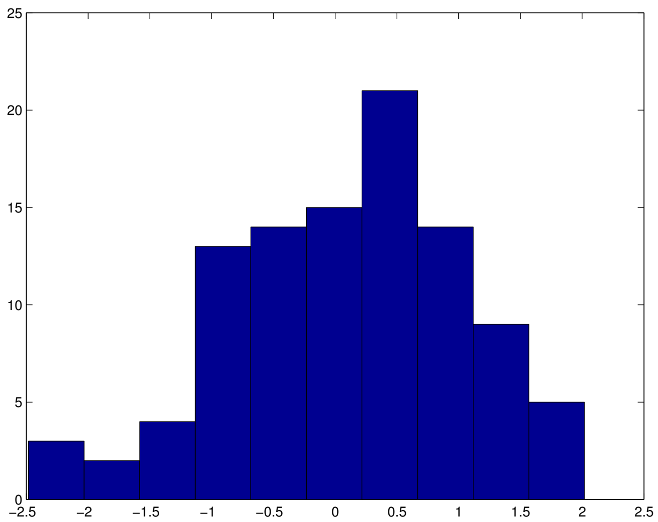
  $\mathbf{R}_1 = \mathrm{diag}(1,2,\ldots,50)$, $\mathbf{R}_2 = \mathrm{diag}(51,52,\ldots,100)$, $\bar{\theta}_i = 1$, $i = 1,\ldots,50$

  50 paramètres pour 100 observations…

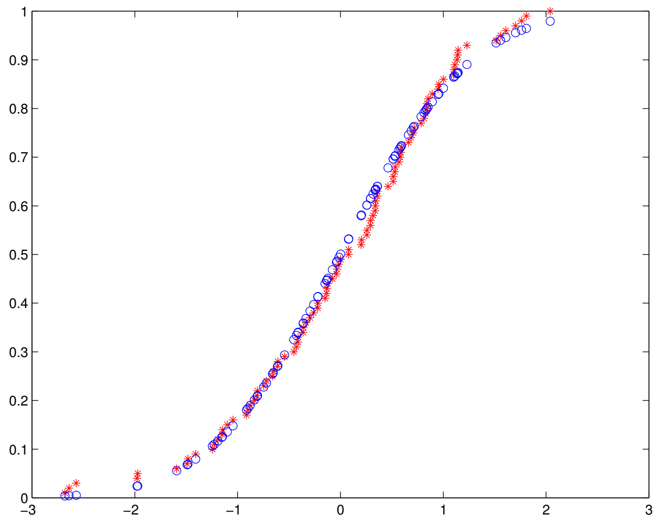
  100 résidus d'estimation par MC : $\hat{\theta} = (\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top\mathbf{y}$
- (iv) Bruit blanc filtré par un AR : $x(k+1) = -x(k) + \epsilon_{k+1}$, $(\epsilon_k)_k$ i.i.d. $\mathcal{N}(0,1)$

$\rightarrow$ histogrammes, fonctions de répartition, *pp-plots*

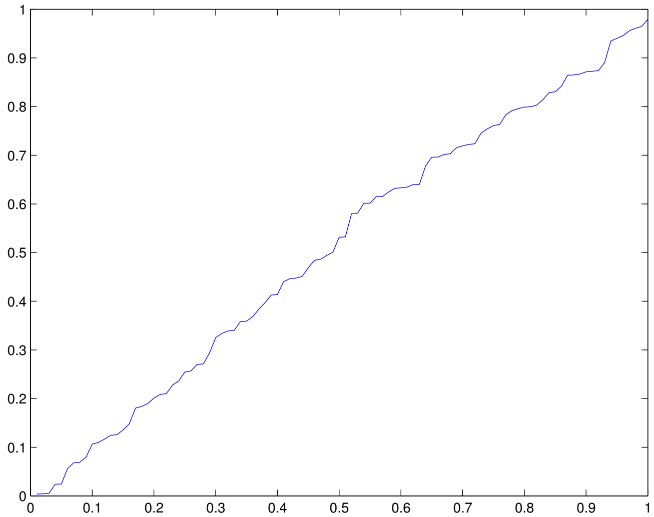
3.1) tracés :



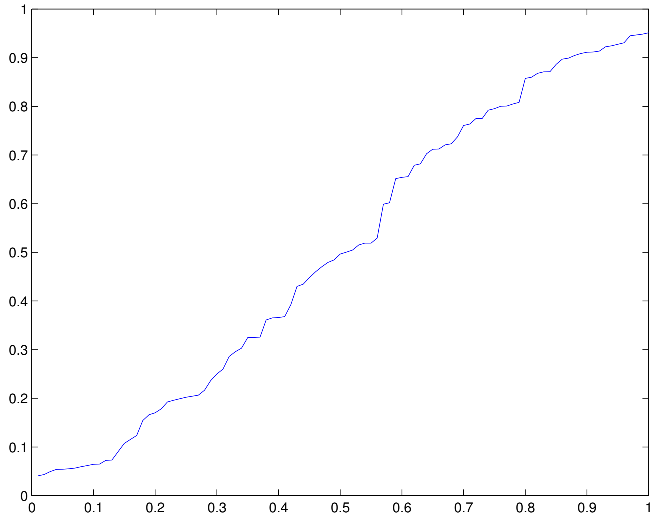
Histogramme, jeu (i)



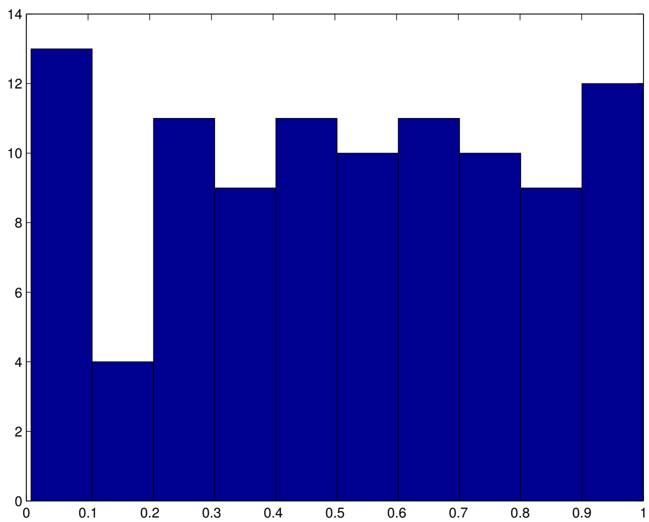
fonction de répartition
empirique et théorique, jeu (i)



pp-plot, jeu (i)



pp-plot, jeu (ii)



Histogramme, jeu (ii)

Variable uniformes :

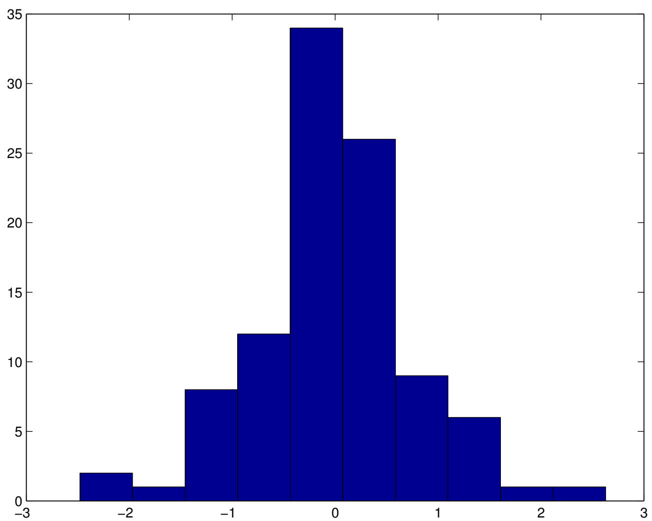
très $\neq$ de variables normales

Mais histogramme souvent

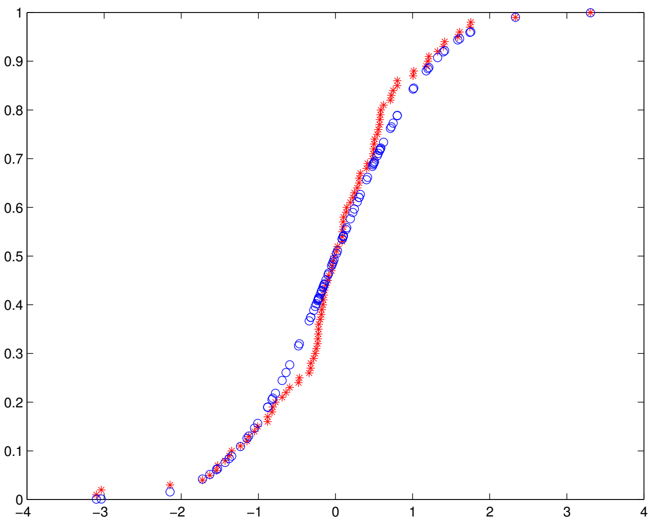
assez peu révélateur si

N pas très grand et

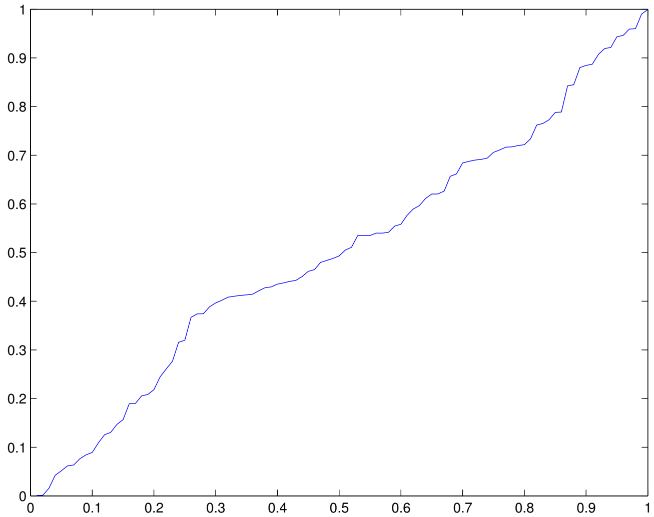
distribution presque normale



Histogramme, jeu (iii)



fonction de répartition
empirique et théorique, jeu (iii)



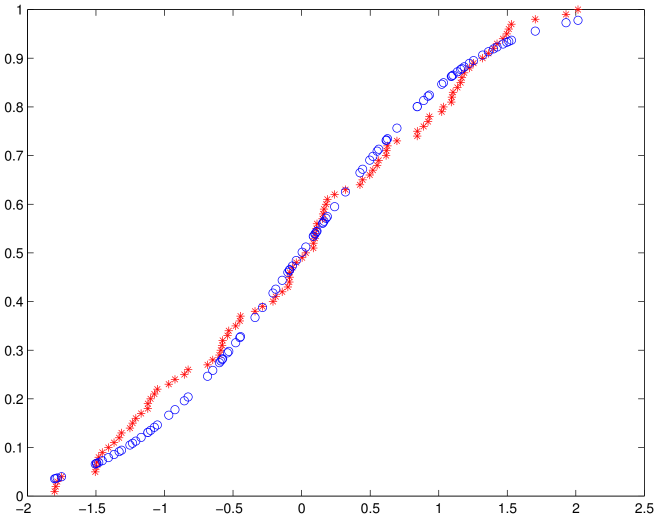
pp-plot, jeu (iii)

50 paramètres pour 100 observations

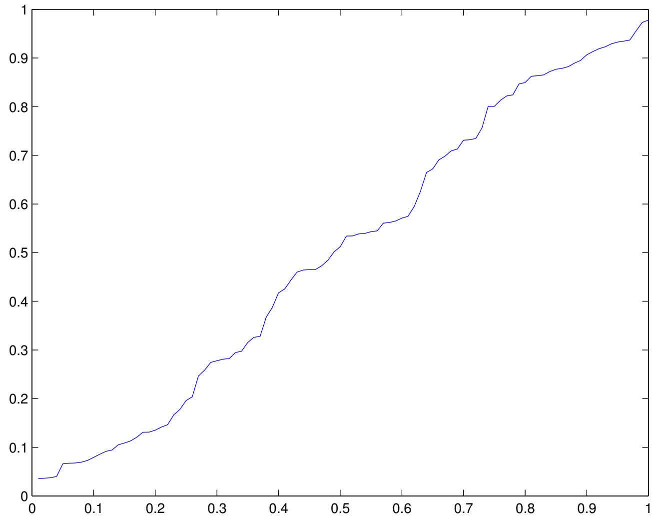
les résidus sont très corréles

→ fausse l'analyse !

(. . . voir + loin comment faire)



fonction de répartition
empirique et théorique, jeu (iv)



pp-plot, jeu (iv)

Résultats voisins de ceux du jeu (iii) : la corrélation fausse l'analyse

Analyse du jeu (iii) ($N = 100$, $p = 50$) :

Résidus d'estimation par MC : $\mathbf{e}(\hat{\theta}) = \mathbf{y} - \mathbf{R}\hat{\theta} = [\mathbf{I}_N - \mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}^\top]\mathbf{y}$

Matrice de covariance $\mathbf{V} = \sigma^2[\mathbf{I}_N - \mathbf{R}(\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}]$

$\rightarrow$ résidus corréles et non stationnaires

Mais on peut en extraire $N - p$ variables de covariance $\sigma^2\mathbf{I}_{N-p}$:

- ☞ Partitionner $\mathbf{R}$ et $\mathbf{y}$ en $\mathbf{R} = \begin{pmatrix} \mathbf{R}_0 \\ \mathbf{R}_1 \end{pmatrix}$, $\mathbf{y} = \begin{pmatrix} \mathbf{y}_0 \\ \mathbf{y}_1 \end{pmatrix}$

  avec $\mathbf{R}_0 \in \mathbb{R}^{p\times p}$ non singulière
- ☞ Calculer les résidus associés au partitionnement précédent :

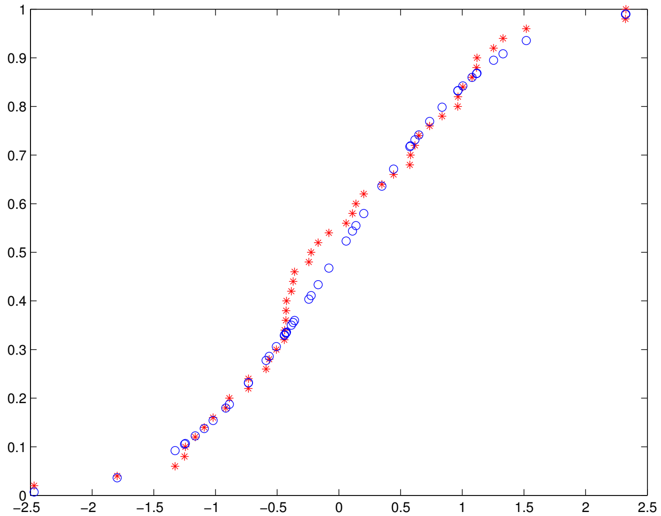
$$\mathbf{e}_0(\hat{\theta}) = \mathbf{y}_0 - \mathbf{R}_0\hat{\theta},\ \ \mathbf{e}_1(\hat{\theta}) = \mathbf{y}_1 - \mathbf{R}_1\hat{\theta}$$

- ☞ Calculer $\mathbf{M} = \mathbf{R}_0(\mathbf{R}^\top\mathbf{R})^{-1}\mathbf{R}_0^\top$

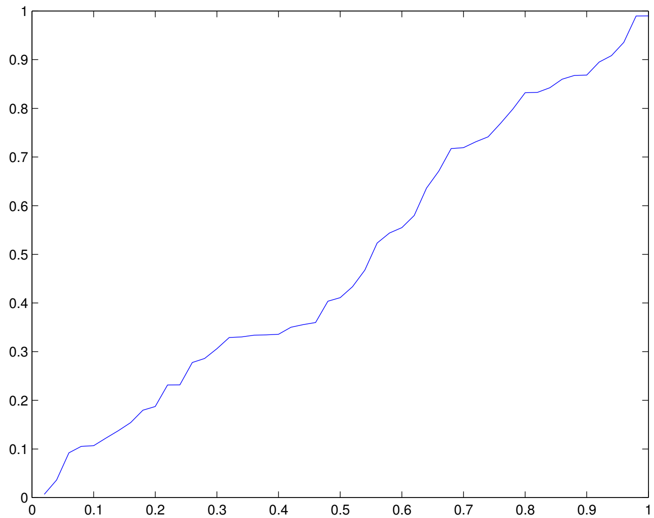
☞ Soient $\lambda_1, \ldots, \lambda_n$, $n \leq p$, les n valeurs propres de $\mathbf{M} < 1$, et $\mathbf{v}_i$, $i = 1, \ldots, n$ les vecteurs propres associés. Les variables $\mathbf{e} \in \mathbb{R}^{N-p}$ à considérer sont

$$\mathbf{e} = \mathbf{e}_1(\hat{\theta}) - \mathbf{R}_1 \mathbf{R}_0^{-1} \left(\sum_{i=1}^{n} \frac{\sqrt{\lambda_i}}{1 + \sqrt{\lambda_i}} \mathbf{v}_i \mathbf{v}_i^\top \right) \mathbf{e}_0(\hat{\theta})$$

Ex2 (suite) : Appliqué au jeu de donné (iii), cela donne un échantillon de $N - p = 50$ variables $\rightarrow$ jeu de données (iii')



fonction de répartition
empirique et théorique, jeu (iii′)



pp-plot, jeu (iii′)

Le traitement appliqué au jeu (iii) conduit à une meilleure analyse de la normalité

3.2) Utilisation des moments empiriques :

$(e_k)_k$ une suite de variables aléatoires i.i.d., de variance σ^2

Coefficient de symétrie (*skewness*) de leur loi :

$$\gamma_1 = \frac{\mathtt{E}\{(e_1 - \mathtt{E}\{e_1\})^3\}}{\sigma^3}$$

$\gamma_1 = 0$ pour une loi symétrique, $\gamma_1 > 0$ si plus de poids à droite qu'à gauche de la moyenne

Coefficient d'aplatissement (*kurtosis*) :

$$\gamma_2 = \frac{\mathtt{E}\{(e_1 - \mathtt{E}\{e_1\})^4\}}{\sigma^4} - 3$$

$\gamma_2 = 0$ pour la loi normale, $\gamma_2 > 0$ si décroissance plus lente que pour la loi normale quand $|x| \to \infty$

→ on va utiliser les valeurs empiriques de γ_1 et γ_2

$$\begin{aligned}\hat{\gamma}_1 &= \frac{\frac{1}{N}\sum_{k=1}^{N}[e(\hat{\theta},k)-\hat{m}]^3}{\hat{v}^{3/2}} \\ \hat{\gamma}_2 &= \frac{\frac{1}{N}\sum_{k=1}^{N}[e(\hat{\theta},k)-\hat{m}]^4}{\hat{v}^{2}} - 3\end{aligned}$$

Sous l'hypothèse de normalité, $\hat{\gamma}_1$ et $\hat{\gamma}_2$ ont pour moyenne

$$m_1 = 0 \text{ et } m_2 = -\frac{6}{N+1}$$

et pour variance

$$v_1 = \frac{6(N-2)}{(N+1)(N+3)} \text{ et } v_2 = \frac{24N(N-2)(N-3)}{(N+1)^2(N+3)(N+5)}$$

et asymptotiquement ($N \to \infty$) $\hat{\gamma}_1$ et $\hat{\gamma}_2$ sont gaussiens

→ on rejette l'hypothèse de normalité si

$$T_1 = \frac{|\hat{\gamma}_1|}{2\sqrt{v_1}} > 1 \ \text{ ou } T_2 = \frac{|\hat{\gamma}_2 - m_2|}{2\sqrt{v_2}} > 1$$

jeu	T_1	T_2	conclusion
(i)	0.9437	0.0834	normal
(ii)	0.1725	1.1972	pas normal
(iii)	0.3471	1.8250	pas normal
(iv)	0.0385	1.05	??
(iii')	0.2346	0.0325	normal

4) Test de stationnarité

Stationnarité du second ordre (de σ^2)

Par exemple, estimation par MC ordinaires $\rightarrow$ hypothèse de stationnarité rejetée $\rightarrow$ MC pondérés

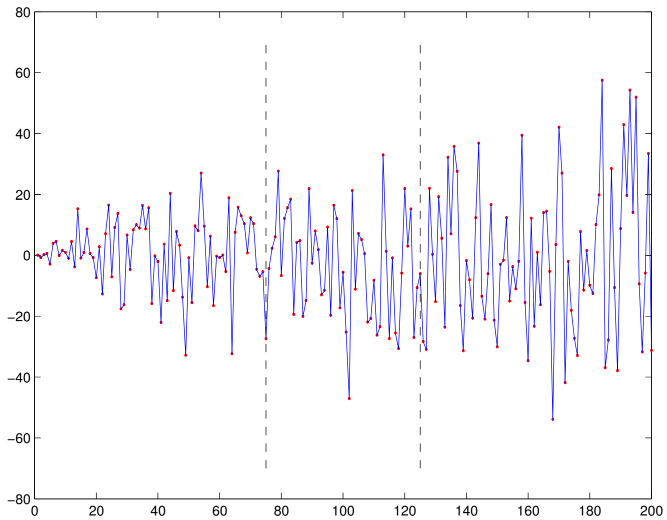
4.1) Goldfeld-Quandt :

La variable «explicative» pour la non-stationnarité éventuelle est le temps (indice du résidu)

On divise les résidus en 3 parties : les k_1 premiers et k_3 derniers seront utilisés, les k_2 du milieu sont écartés (typiquement, $k_2 = N/4$ et $k_1 = k_3$)

On calcule ensuite $\hat{m}_1$ et $\hat{m}_3$ les moyennes pour les k_1 premiers et k_3 derniers résidus, puis

$$r_3 = \frac{k_3 - p}{k_1 - p} \, \frac{\sum_{k=1}^{k_1} [e(\hat{\theta},k) - \hat{m}_1]^2}{\sum_{k=k_1+k_2+1}^{N} [e(\hat{\theta},k) - \hat{m}_3]^2}$$



processus non stationnaire,
méthode de Goldfeld-Quandt

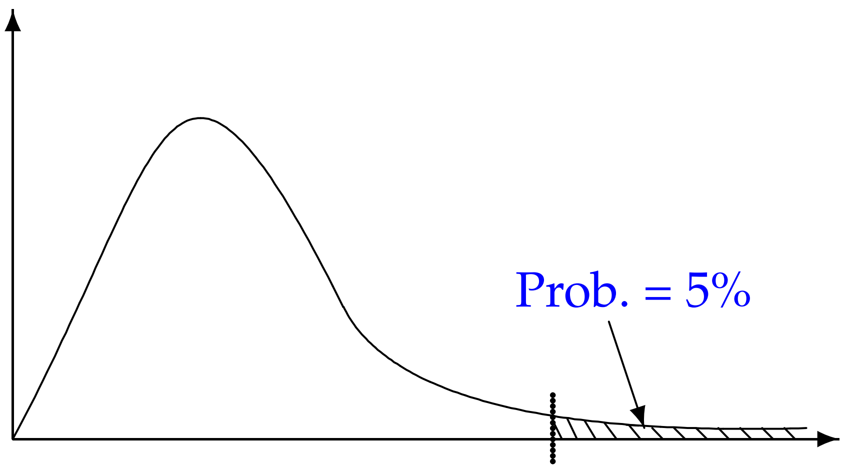
$\longrightarrow$ revient à calculer le rapport de variances estimées sur les k_1 premiers et k_3 derniers résidus

On doit ensuite tester si r_3 est égal à 1 (hypothèse H_0 : stationnarité), ou «significativement différent de 1»

cela dépend de l'hypothèse alternative H_1 ...

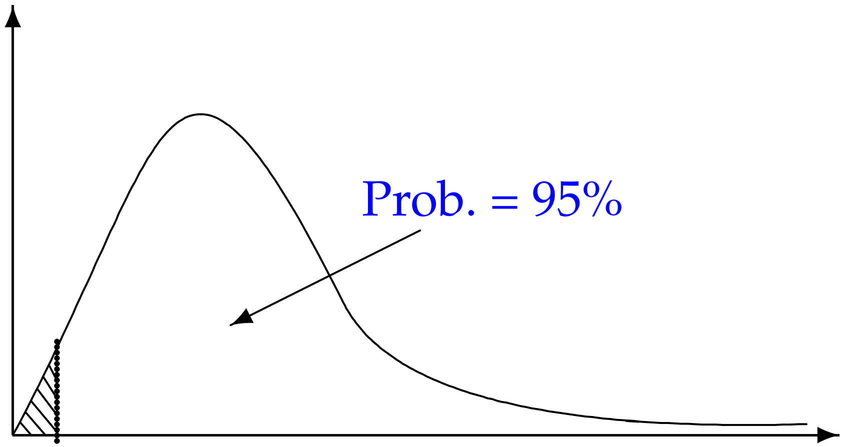
Sous H_0, avec une structure LP et un bruit i.i.d. normal, les résidus d'estimation par MC ordinaires (non pondérés) sont tels que $r_3 \sim$ Fisher-Snedecor à $k_1 - p$ et $k_3 - p$ degrés de liberté

Si H_1 = variance décroissante avec k, on teste si r_3 est significativement grand



$T_3 = \frac{r_3}{F_5(k_1-p,k_3-p)}$
Stationnarité si $T_3 < 1$,
variance décroissante sinon

Si H_1 = variance croissante avec k, on teste si r_3 est significativement petit



$F_{95}(k_1 - p, k_3 - p)$ (tabulé)

$T'_3 = \frac{r_3}{F_{95}(k_1 - p, k_3 - p)}$

Stationnarité si $T'_3 > 1$,

variance croissante sinon

4.2) Régression du carré des résidus :

$z_i(k)$, $i = 1,\ldots,n$ des variables explicatives, Par exemple, $z(k) = k$, ou $\mathbf{z}(k) = \mathbf{r}(k)$ ($n = p$)

A partir de $\hat{m}$ et $\hat{v} \longrightarrow$ résidus normalisés

$$e_n(\hat{\theta},k) = \frac{e(\hat{\theta},k) - \hat{m}}{\sqrt{\hat{v}}}, \ \ k = 1,\ldots,N$$

puis nouvelles «observations» $v_n(\hat{\theta},k) = e_n^2(\hat{\theta},k)$ que l'on va tenter de modéliser avec les $z_i(k)$

$$v_n(\hat{\theta},k) = \alpha_0 + \sum_{i=1}^{n} \alpha_i z_i(k) + \epsilon'_k$$

Stationnarité $\longrightarrow \alpha_1 = \alpha_2 = \cdots = \alpha_n = 0$ et $\alpha_0 = \sigma^2$

$\longrightarrow$ estimer $\alpha = (\alpha_0,\alpha_1,\ldots,\alpha_n)^\top$ par MC

$$\hat{\alpha} = (\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{v}_n(\hat{\theta})$$

avec $\mathbf{v}_n(\hat{\theta}) = [v_n(\hat{\theta},1), \ldots, v_n(\hat{\theta},N)]^\top$ et $\mathbf{Z}$ la matrice des régresseurs $(N \times n)$

$$\mathbf{Z} = \begin{pmatrix} 1 & z_1(1) & \cdots & z_n(1) \\ 1 & z_1(2) & \cdots & z_n(2) \\ \vdots & \vdots & \vdots & \vdots \\ 1 & z_1(N) & \cdots & z_n(N) \end{pmatrix}$$

On va calculer la somme des carrés des résidus expliquée par la régression sur les $z_i(k)$

Somme des carrés des résidus avec tous les $z_i(k)$, $i = 1, \ldots, n$:

$$\|\mathbf{v}_n(\hat{\theta}) - \mathbf{Z}\hat{\alpha}\|^2 = \|\mathbf{v}_n(\hat{\theta})\|^2 - \mathbf{v}_n^\top(\hat{\theta})\mathbf{Z}(\mathbf{Z}^\top \mathbf{Z})^{-1} \mathbf{Z}^\top \mathbf{v}_n(\hat{\theta})$$

avec α_0 seul : $\|\mathbf{v}_n(\hat{\theta})\|^2 - \frac{1}{N}\left[\sum_{k=1}^N v_n(\hat{\theta},k)\right]^2$

$\rightarrow$ partie expliquée par les $z_i(k)$:

$$s_4 = \mathbf{v}_n^\top(\hat{\theta})\mathbf{Z}(\mathbf{Z}^\top\mathbf{Z})^{-1}\mathbf{Z}^\top\mathbf{v}_n(\hat{\theta}) - \frac{1}{N}\left[\sum_{k=1}^{N} v_n(\hat{\theta},k)\right]^2$$

Sous l'hypothèse de stationnarité, on doit avoir $r_4 = s_4/\|\mathbf{v}_n(\hat{\theta})\|^2 << 1$

On montre que $s_4/2$ suit une loi du $\mathcal{X}^2$ à n degrés de liberté $\rightarrow$

$$T_4 = \frac{s_4}{2\mathcal{X}_5^2(n)}$$

$\mathcal{X}_5^2(n)$ ayant une probabilité de 5% d'être dépassé par une variable distribuée $\mathcal{X}^2(n)$ (valeur tabulée)

4.3) Corrélation avec variables explicatives :

On calcule la corrélation entre les $v_n(\hat{\theta},k)$ et les $z_i(k)$, pour chaque i

$$c_5(i) = \frac{\sum_{k=1}^{N}[v_n(\hat{\theta},k) - m_v][z_i(k) - m_{z_i}]}{\left\{\sum_{k=1}^{N}[v_n(\hat{\theta},k) - m_v]^2 \sum_{k=1}^{N}[z_i(k) - m_{z_i}]^2\right\}^{1/2}}$$

avec m_v la moyenne des $v_n(\hat{\theta},k)$ et m_{z_i} celle des $z_i(k)$

Stationnarité $\rightarrow c_5(i)$ doit être proche de 0

On montre que $\sqrt{N-p}\,c_5(i) \sim \mathcal{N}(0,1)$, d'où

$$T_5(i) = \frac{|c_5(i)|\sqrt{N-p}}{2}$$

et on rejette la stationnarité si $T_5(i) > 1$

Ex2 (suite) : 3 jeux de données de plus

Régression linéaire, $\mathbf{y} = \mathbf{R}\bar{\theta} + \epsilon$, avec $\mathbf{R} = (\mathbf{w}_1,\mathbf{w}_2)$,

$$w_{1i} = \cos\left(\frac{i-1}{99}\frac{\pi}{2}\right), w_{2i} = \sin\left(\frac{i-1}{99}\frac{\pi}{2}\right), i = 1,\ldots,100$$

$\bar{\theta} = (1,1)^\top$, résidus d'estimation par MC non pondérés

- (v) $(\epsilon_k)_k$ i.i.d. $\mathcal{N}(0,1)$
- (vi) $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{V}_1)$, $\mathbf{V}_1 = \mathrm{diag}(1,2^2,\ldots,100^2)$ (non stationnaire)
- (vii) $\epsilon \sim \mathcal{N}(\mathbf{0},\mathbf{V}_2)$, $\mathbf{V}_2 = \mathbf{M}\mathbf{V}_1\mathbf{M}^\top$ et

$$\mathbf{M} = \begin{pmatrix} 1 & 1 & 0 & 0 & \cdots & \cdots & \cdots & 0 \\ 0 & 1 & 1 & 0 & \cdots & \cdots & \cdots & \cdots \\ 0 & 0 & 1 & 1 & \cdots & \cdots & \cdots & \cdots \\ \vdots & \vdots & \vdots & \vdots & \cdots & 1 & 1 & 0 \\ 0 & \cdots & \cdots & \cdots & \cdots & 0 & 1 & 1 \end{pmatrix}$$

(non stationnaire, corrélation)

Variables explicatives :

$z(k) = k$ pour (i) à (iv), $\mathbf{z}(k) = \mathbf{r}(k)$ pour (v), (vi) et (vii)

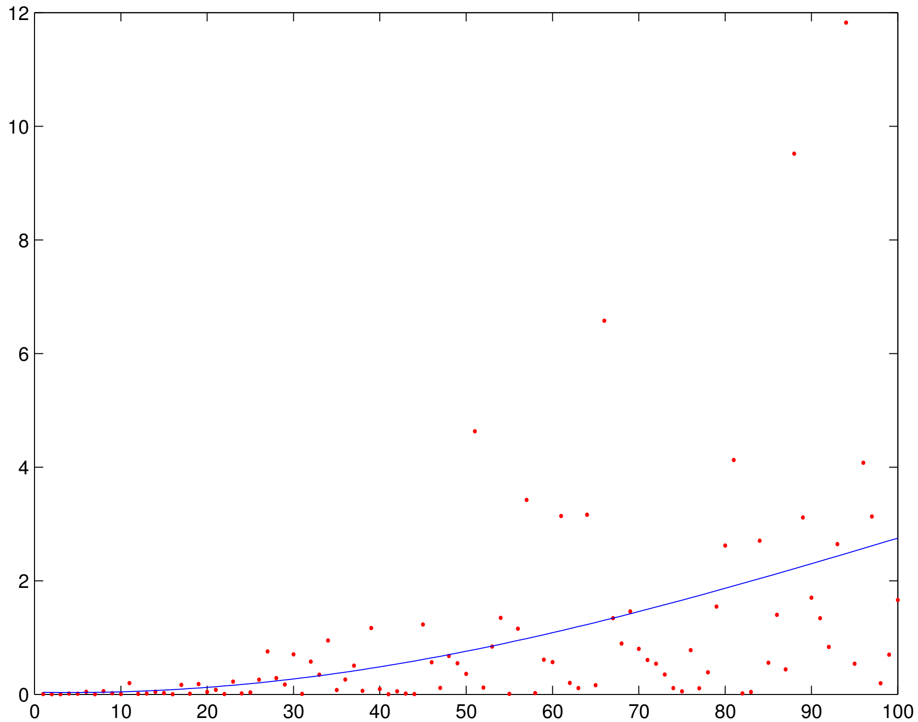
(iii) p trop grand pour T_3 et T_3'

jeu	T_3	T_3'	r_4	T_4	$c_5(1)$	$T_5(1)$	$c_5(2)$	$T_5(2)$
(i)	0.509	1.5	0.002	0.076	0.055	0.273	–	–
(ii)	0.619	1.822	0.0023	0.003	-0.070	0.352	–	–
(iii)	–	–	0.071	4.165	-0.301	1.064	–	–
(iii′)	0.276	1.296	0.014	0.246	0.143	0.506	–	–
(iv)	1.732	5.103	0.030	0.752	-0.244	1.222	–	–
(v)	0.487	1.478	0.033	0.752	0.050	0.248	0.043	0.212
(vi)	0.048	0.147	0.166	5.990	-0.462	2.289	0.420	2.080
(vii)	0.005	0.016	0.228	12.815	-0.497	2.459	0.398	1.971

Non stationnaire : correct, incorrect

Régression du carré des résidus et «observations $v_n(\hat{\theta},k)$», jeu (vi)

variables explicatives : $\mathbf{z}(k) = \mathbf{r}(k)$



5) Test d'indépendance

5.1) Corrélation

Rappel : indépendance $\Rightarrow$ pas de corrélation
(et donc corrélation $\Rightarrow$ dépendance)
et indépendance $\Leftrightarrow$ pas de corrélation si variables normales

Fonction d'autocorrélation des erreurs :

$$\hat{c}_e(k) = \frac{1}{N-k} \sum_{i=1}^{N-k} [e(\hat{\theta},i) - \hat{m}][e(\hat{\theta},i+k) - \hat{m}]$$

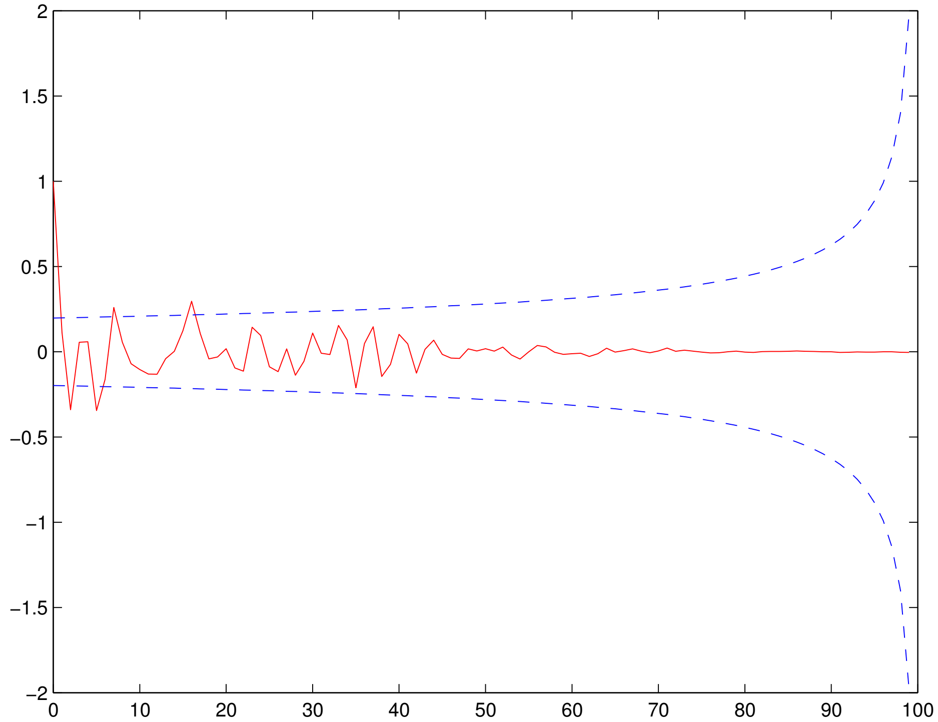
$\longrightarrow$ fonction normalisée $\hat{c}_n(k) = \frac{\hat{c}_e(k)}{\hat{c}_e(0)}$ $(\longrightarrow \hat{c}_n(0) = 1)$

Indépendance $\Rightarrow$ pour $N >> k$, $\hat{c}_n(k) \sim \mathcal{N}(0,1/N)$
plus précisément (le nb. de termes de la somme dépend de $k \ldots$),

$$\hat{c}_n(k) \sim \mathcal{N}\left(0, \frac{N}{(N+2)(N-k)}\right)$$

$\Rightarrow \hat{c}_n(k)$ doit être dans l'intervalle $\mathcal{I}_k = \left[-\frac{2\sqrt{N}}{\sqrt{(N+2)(N-k)}}, \frac{2\sqrt{N}}{\sqrt{(N+2)(N-k)}} \right]$ avec une prob. 95%

Ex2 (suite) : $\hat{c}_n(k)$ et extrémités de l'intervalle $\mathcal{I}_k$, jeu (vii)



$\longrightarrow$ dépendance

Pour N grand,

$$v_6 = \frac{N+2}{N} \sum_{k=1}^{N} (N-k)\hat{c}_n^2(k)$$

est approximativement distribuée suivant $\mathcal{X}^2(N-p)$

$$T_6 = \frac{v_6}{\mathcal{X}_5^2(N-p)}$$

avec $\mathcal{X}_5^2(N-p)$ ayant une probabilité de 5% d'être dépassé par une variable distribuée suivant une loi $\mathcal{X}^2(N-p)$
$\rightarrow$ on rejette l'hypothèse d'indépendance si $T_6 > 1$

Rq : On peut aussi considérer la corrélation entre les résidus et les entrées (pas de corrélation si le modèle a «extrait toute l'information»)

5.2) Méthodes non paramétriques

5.2.1) Comparaison à la médiane

Si la suite de variables (erreurs normalisées) est indépendante, chacune à une probabilité $1/2$ d'être au dessus comme au dessous de la médiane m

$$x(k) = \begin{cases} 1 & \text{si } e_n(\hat{\theta},k) > m \\ 0 & \text{si } e_n(\hat{\theta},k) < m \end{cases} \quad \text{et 0 ou 1 avec prob. } 1/2 \text{ si } e_n(\hat{\theta},k) = m$$

Déf.: *run* = suite continue d'éléments de $x(k)$ ayant la même valeur

Ex : $(x(k)) = \{011\,\underbrace{000}\,11100111110010\} \rightarrow 9$ *runs*

r le nb. de *runs*, n le nb. de 1
Pour N grand, indépendance $\rightarrow r \sim \mathcal{N}(m_7,v_7)$, avec

$$m_7 = \frac{2n(N-n)}{N} + 1, \;\; v_7 = \frac{2n(N-n)[2n(N-n)-N]}{N^2(N-1)}$$

Le test est $T_7 = \frac{|r-m_7|}{2\sqrt{v_7}}$, et on rejette l'indépendance si $T_7 > 1$

5.2.2) Comparaison des valeurs successives

Si la suite est indépendante, chaque terme à la même probabilité 1/2 d'être plus petit ou plus grand que le précédent

$$x(k) = \begin{cases} 1 & \text{si } e_n(\hat{\theta},k+1) > e_n(\hat{\theta},k) \\ 0 & \text{si } e_n(\hat{\theta},k+1) > e_n(\hat{\theta},k) \end{cases}$$

et $x(k) = 0$ ou 1 avec prob. 1/2 si $e_n(\hat{\theta},k+1) = e_n(\hat{\theta},k)$

Pour N grand, indépendance $\rightarrow$ $r \sim \mathcal{N}(m_8,v_8)$, avec

$$m_8 = \frac{2N-1}{3}, \; v_8 = \frac{16N-29}{90}$$

Le test est $T_8 = \frac{|r-m_8|}{2\sqrt{v_8}}$, et on rejette l'indépendance si $T_8 > 1$

5.2.3) Rapport de Von-Neumann

On classe les erreurs par valeurs croissantes
$e_n(\hat{\theta},k_1) \leq e_n(\hat{\theta},k_2) \leq \cdots \leq e_n(\hat{\theta},k_N) \rightarrow r(k)$ le rang de l'erreur $e_n(\hat{\theta},k)$

Rapport de Von-Neumann :

$$r_9 = 12\frac{\sum_{k=2}^{N}[r(k)-r(k-1)]^2}{N(N^2-1)}$$

Pour N grand, indépendance $\rightarrow r_9 \sim \mathcal{N}\left(2,\frac{20}{5N+7}\right)$

Le test est

$$T_9 = \frac{|r_9-2|}{2\sqrt{\frac{20}{5N+7}}}$$

et on rejette l'indépendance si $T_9 > 1$

Ex2 (suite) :

jeu	T_6	T_7	T_8	T_9
(i)	0.5778	0	0.4388	0.3274
(ii)	0.4256	0.4020	0.6383	0.2808
(iii)	0.5834	0.4020	0.7579	0.2378
(iii′)	0.3336	0	0.6833	0.1374
(iv)	27.6005	4.0204	3.1913	4.9478
(v)	0.3645	0.1005	0.4388	0.1690
(vi)	0.9841	0.5025	0.6383	0.0828
(vii)	1.6952	1.5076	1.3563	0.8467

Dépendance : correct, pas détectée

6) Simulation

Si résultats de tests ambigus ?

☞ simuler le modèle avec la valeur estimée $\hat{\theta}$ des paramètres,
les mêmes conditions expérimentales que celles utilisées pour obtenir $\hat{\theta}$,
des perturbations respectant les hypothèses faites,

☞ recueillir les données comme pour un vrai système,

☞ recommencer l'identification à partir de ces données

Si le modèle est correct, elles devraient être semblables au données initiales
… et conduire aux mêmes ambiguïtés !

S'il le faut, recueillir d'autres observations sur le système
(penser «planification d'expériences» …)