

Econométrie des Séries Temporelles Univariées

Chapitre 2 : Estimation et Sélection de Modèles

Gilles de Truchis

Master 1 ESA

2022-2023

Le plan du Chapitre

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE

- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

Plan

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE

- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

Estimation des modèles linéaires

Objectif Identifier un modèle linéaire approprié pour une série $\{\tilde{X}_t\}_{t=1}^n$

- Cela implique de résoudre plusieurs problèmes itérativement
 - estimer μ_X (afin de travailler avec $X_t = \tilde{X}_t - \mu_X$)
 - estimer les coefficients du modèle (dans la classe des $\text{ARMA}(p, q)$)
 - sélectionner l'ordre des retards optimaux p et q
 - estimer la variance du bruit blanc σ_ε^2
- Dans un second temps, le modèle sélectionné devra être
 - soumis à des tests de diagnostic (e.g. sphéricité des erreurs)
 - être utilisé pour de la prédiction

Estimation de μ_X

- L'estimateur de la moyenne μ_X d'un processus stationnaire X_t est

$$\bar{X}_n = n^{-1} \sum_{j=1}^n X_j$$

- \bar{X}_n , la moyenne empirique, est un estimateur non-biaisé car

$$\mathbb{E}(\bar{X}_n) = n^{-1}(\mathbb{E}X_1 + \dots + \mathbb{E}X_n) = \mu_X$$

- La MSE de \bar{X}_n est donnée par

$$\begin{aligned}\mathbb{E}(\bar{X}_n - \mu_X)^2 &= \mathbb{V}(\bar{X}_n) = n^{-2} \sum_{i=1}^n \sum_{j=1}^n \text{Cov}(X_i, X_j) \\ &= n^{-2} \sum_{i-j=-n}^n (n - |i - j|) \gamma_X(i - j) \\ &= n^{-1} \sum_{h=-n}^n (1 - n^{-1}|h|) \gamma_X(h)\end{aligned}$$

- Si $X_t \sim ARMA(p, q)$, $\gamma_X(h)$ est absolument sommable et

$$\lim_{n \rightarrow \infty} \mathbb{V}(\bar{X}_n) = n^{-1} \sum_{h=-\infty}^{\infty} |\gamma_X(h)|$$

Inférence autour de μ_X

- Sous certaines conditions on peut montrer que

$$n^{1/2}(\bar{X}_n - \mu_X) \sim \mathcal{N}\left(0, \sum_{|h| < n} (1 - n^{-1}|h|)\gamma(h)\right)$$

e.g. Si X_t est Gaussien ou linéaire ce résultat est valide

- Une approximation de l'intervalle de confiance (IC) à 95% est alors

$$\left(\bar{X}_n - 1.96 \frac{\nu^{1/2}}{n^{1/2}}, \bar{X}_n + 1.96 \frac{\nu^{1/2}}{n^{1/2}}\right)$$

où $\nu = \sum_{h=-\infty}^{\infty} \gamma_X(h)$, généralement inconnu, devra être estimé aussi

- Considérons un exemple : soit un AR(1) avec $|\phi_1| < 1$,

$$X_t - \mu_X = \phi_1(X_{t-1} - \mu_X) + \varepsilon_t, \quad \gamma_X(h) = \phi_1^{|h|} \sigma_\varepsilon^2 (1 - \phi_1^2)^{-1}$$

$$\text{et donc } \nu = (1 + 2 \sum_{h=1}^{\infty} \phi_1^h) \sigma_\varepsilon^2 (1 - \phi_1^2)^{-1} = \sigma_\varepsilon^2 (1 - \phi_1)^{-2}$$

- On peut alors construire l'IC à 95% pour μ_X :

$$\bar{x}_n \pm 1.96 \sigma n^{-1/2} (1 - \phi_1)^{-1}$$

Estimateurs de $\gamma_X(\cdot)$ et $\rho_X(\cdot)$

- Les estimateurs de l'ACovF et l'ACF sont données par

$$\hat{\gamma}(h) = n^{-1} \sum_{t=1}^{n-|h|} (X_{t+|h|} - \bar{X}_n)$$

et $\hat{\rho}(h) = \hat{\gamma}(h)/\hat{\gamma}(0)$

- Ces estimateurs sont biaisés mais si on remplace n^{-1} par

$$(n - h)^{-1}$$

ils sont presque pas biaisés pour n grand

- On peut montrer que pour $\boldsymbol{\rho} = (\rho(1), \dots, \rho(k))'$ les ACF empiriques sont approximativement Normales

$$\hat{\boldsymbol{\rho}} \xrightarrow{a.a.d} \mathcal{N}(\boldsymbol{\rho}, n^{-1}W)$$

avec W une matrice de covariance dont les coefficients sont donnés par la formule de Bartlett

$$w_{ij} = \sum_{k=1}^{\infty} (\rho(k+i) + \rho(k-i) - 2\rho(i)\rho(k)) \\ \times (\rho(k+j) + \rho(k-j) - 2\rho(j)\rho(k))$$

Inférence autour de $\gamma_X(\cdot)$ et $\rho_X(\cdot)$

- Prenons l'exemple d'un AR(1) et comparons-le avec l'ACF empirique

$$X_t = \phi_1 X_{t-1} + \varepsilon_t$$

avec $|\phi_1| < 1$ de sorte que nous savons que $\rho_X(h) = \phi_1^{|h|}$ et donc

$$\begin{aligned} w_{ii} &= \sum_{k=1}^i \phi_1^{2i} (\phi_1^{-k} - \phi_1^k)^2 + \sum_{k=i+1}^{\infty} \phi_1^{2k} (\phi_1^{-i} - \phi_1^i)^2 \\ &= (1 - \phi_1^{2i})(1 + \phi_1^2)(1 - \phi_1^2)^{-1} - 2i\phi_1^{2i} \end{aligned}$$

- Considérons le niveau annuel en “pieds” du Lac Huron (1875-1972)
- Supposons que l'estimation du modèle AR(1) donne

$$x_t - \bar{x} = 0.791(x_{t-1} - \bar{x}) + \varepsilon_t$$

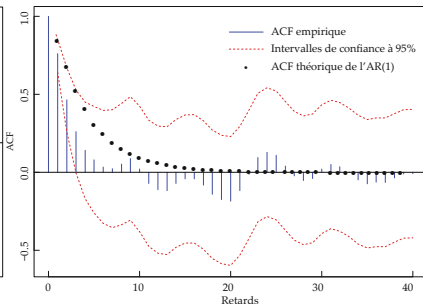
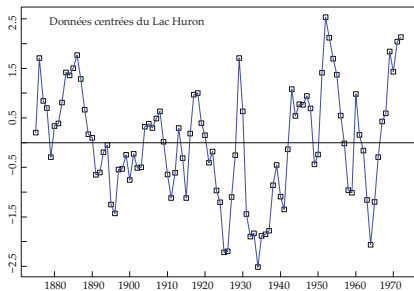
- On peut alors calculer $\hat{\rho}(h)$ et le comparer avec

$$\rho_X(h) = \phi_1^{|h|}$$

les intervalles de confiance à 95% étant donnés par

$$\hat{\rho}(h) \pm 1.96n^{-1/2}w_{ii}^{1/2}, \quad i = 1, \dots, h$$

Analyse graphique de $\rho_X(\cdot)$ et $\hat{\rho}(h)$



■ L'allure des données :

⇒ stationnarité et faible de dépendance donc l'AR(1) est un bon candidat

■ Analyse de l'ACF empirique et ses intervalles de confiance :

⇒ l'ACF théorique touche les intervalles aux retards 2 à 4

⇒ cela suggère une certaine incompatibilité entre les données et le modèle

Analyse graphique de $\rho_X(\cdot)$ et $\hat{\rho}(h)$

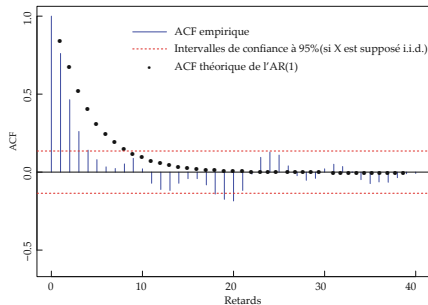
- Dans l'hypothèse où X_t serait une séquence i.i.d. ($0, \sigma_X^2 < \infty$)

$$\hat{\rho} \xrightarrow{a.a.d} \mathcal{N}(\rho, n^{-1}), \quad \rho = 0$$

ce qui implique que 95% des fois, $\hat{\rho}$ devrait tomber dans l'intervalle

$$\pm 1.96n^{-1/2}$$

ce qui ici revient à ± 0.1990 car $n = 97$



Les estimateurs de Yule-Walker d'un AR

- Soit un processus $AR(p)$ stationnaire : $X_t = \sum_{j=1}^p \phi_j X_{t-j} + \varepsilon_t$
 - On a vu au Ch. 1 qu'en multipliant par X_{t-j} , de chaque côté ...
- ... et en prenant l'espérance, on obtenait les équations de Yule-Walker
- ⇒ Exprimées en terme d'ACovF sous forme matricielle on a

$$\Gamma_p \phi_p = \gamma_p$$

avec $\Gamma_p = [\gamma(i-j)]_{i,j=1}^p$ et $\gamma_p = (\gamma(1), \dots, \gamma(p))'$

- Dans la pratique on voudra remplacer $\gamma(j)$ par $\hat{\gamma}(j)$ et construire

$$\hat{\phi}_p = \hat{R}_p^{-1} \hat{\rho}_p$$

l'estimateur de Yule-Walker des p coefficients AR ainsi que

$$\hat{\sigma}_\varepsilon^2 = \hat{\gamma}(0) (1 - \hat{\rho}_p' \hat{R}_p^{-1} \hat{\rho}_p)$$

où $\hat{\rho}_p = (\rho(1), \dots, \rho(p))' = \gamma_p / \gamma(0)$

Inférence et équations de Yule-Walker

- On peut montrer que pour $n \rightarrow \infty$ sous certaines conditions

$$\hat{\phi}_p \xrightarrow{a.a.d} \mathcal{N}(\phi, n^{-1} \sigma_\varepsilon^2 \Gamma_p^{-1})$$

- En remplaçant σ_ε^2 par $\hat{\sigma}_\varepsilon^2$, on peut écrire que l'intervalle

$$\hat{\phi}_j \pm \Phi_{1-\alpha/2} n^{-1/2} \hat{\sigma}_\varepsilon$$

contient ϕ_j avec une probabilité de $(1 - \alpha)$ avec

$$\Phi_{1-\alpha/2}$$

le quantile à $(1 - \alpha)$ d'une distribution normale centrée réduite

Les estimateurs de Yule-Walker en présence de composantes MA

- La limite de Yule-Walker se rencontre pour les ARMA($p, q > 0$)

$$X_t = \sum_{j=1}^p \phi_j X_{t-j} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

- En effet, les équations à solutionner pour $p > 0$ et $q > 0$ sont

$$\hat{\gamma}(h) - \phi_1 \hat{\gamma}(h-1) - \dots - \phi_p \hat{\gamma}(h-p) = \sigma_\varepsilon^2 \sum_{j=h}^q \theta_j \psi_{j-h}$$

où les $\Psi_z = \Theta(z)\Phi(z)^{-1}$ sont les coefficients de la forme MA(∞)

- On voit que malgré son écriture simple, ce problème est non-linéaire
- \Rightarrow On ne peut garantir l'existence et l'unicité de la solution

Note Pour un processus MA pur, il est possible d'utiliser Yule-Walker

Les estimateurs de Yule-Walker d'un pure MA

■ Soit un MA(1) avec $|\theta_1| < 1$: $X_t = \varepsilon_t + \theta_1 \varepsilon_{t-1}$

■ On sait que (cf. Ch1) $\hat{\gamma}(0) = \hat{\sigma}_\varepsilon^2(1 + \hat{\theta}_1^2)$ et

$$\hat{\rho}(1) = \frac{\hat{\theta}_1}{1 + \hat{\theta}_1^2}$$

■ On a vu qu'on peut en déduire $\hat{\theta}_1 \in \mathbb{R}$ si $|\hat{\rho}(1)| \leq 1/2$ ($\hat{\theta}_1 \in \mathbb{C}$ sinon)

$$\hat{\theta}(1) = (2\hat{\rho}(1))^{-1}(1 - (1 - 4\hat{\rho}^2(1))^{1/2})$$

et

$$\hat{\sigma}_\varepsilon^2 = (1 + \hat{\theta}_1^2)\hat{\gamma}(0)$$

Note Si $|\hat{\rho}(1)| = 0.5$ on trouve $|\hat{\theta}_1| = 1$ et le processus n'est pas inversible

PACF : estimation et inférence

- Comme énoncé au C1, Yule-Walker nous permet d'obtenir les PACFs
- Les PACFs empiriques peuvent également s'obtenir via la régression

$$x_t = \hat{\phi}_{1,j}x_{t-j} + \cdots + \hat{\phi}_{j,j}x_{t-j} + \varepsilon_t$$

avec l'estimateur OLS (ou MLE Gaussien)

$$\hat{\phi}_{j,j}$$

donnant la corrélation partielle d'ordre j

- La théorie limite de ces estimateurs nous révèle que pour $j > p$

$$\hat{\phi}_{j,j} \xrightarrow{a.a.d} \mathcal{N}(0, n^{-1})$$

et donc qu'un intervalle de confiance à 95% construit autour de

$$\phi_{j,j} = 0$$

est simplement donné par (formule de Quenouille)

$$\pm \frac{1.96}{\sqrt{n}}$$

Plan

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE

- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

Newey-West

- Si le temps nous le permet je reviendrai sur cet estimateur

Plan

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE

- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

Rappels MLE

- Vous connaissez déjà l'estimateur du Maximum de Vraisemblance (MLE)
- Pour un ensemble de paramètres ϑ

- le MLE est convergent

$$\hat{\vartheta} \xrightarrow{p} \vartheta_0$$

- le MLE est asymptotiquement efficace

$$\mathbb{V}(\hat{\vartheta}) = I_n^{-1}(\vartheta_0)$$

- le MLE est asymptotiquement normalement distribué (pour des lois exp)

$$\sqrt{n}(\hat{\vartheta} - \vartheta_0) \xrightarrow{d} \mathcal{N}(0, I_n^{-1}(\vartheta_0))$$

où $I_n^{-1}(\vartheta_0)$ représente la matrice d'information de Fisher

- sous certaines hypothèse de régularité

Note l'élément crucial dans la construction de la vraisemblance étant le **choix de la distribution**

Densité conditionnelle Gaussienne et AR(1)

- Soit un ARMA(p, q) dont on suppose $\varepsilon_t \sim \mathcal{N}(0, \sigma_\varepsilon^2)$

$$X_t = \sum_{j=1}^p \phi_j X_{t-j} + \varepsilon_t + \sum_{j=1}^q \theta_j \varepsilon_{t-j}$$

- Si on sait que $p = 1$ et $q = 0$, on a un AR(1) pour lequel on sait que

$$X_1 \sim \mathcal{N}(0, \sigma_\varepsilon^2(1 - \phi_1^2)^{-1})$$

$$\Rightarrow f_{X_1}(x_1; \phi_1) = (\sigma_X \sqrt{2\pi})^{-1} \exp\left(-\frac{1}{2} \frac{x_1^2}{\sigma_X^2}\right)$$

$$X_2|X_1 \sim \mathcal{N}(\phi_1 X_1, \sigma_\varepsilon^2)$$

$$\Rightarrow f_{X_2|X_1}(x_2|x_1; \phi_1) = (\sigma_\varepsilon \sqrt{2\pi})^{-1} \exp\left(-\frac{1}{2} \frac{(x_2 - \phi_1 x_1)^2}{\sigma_\varepsilon^2}\right)$$

⋮

$$X_n|X_{n-1} \sim \mathcal{N}(\phi_1 X_{n-1}, \sigma_\varepsilon^2)$$

$$\Rightarrow f_{X_n|X_{n-1}}(x_n|x_{n-1}; \phi_1) = (\sigma_\varepsilon \sqrt{2\pi})^{-1} \exp\left(-\frac{1}{2} \frac{(x_n - \phi_1 x_{n-1})^2}{\sigma_\varepsilon^2}\right)$$

Maximum de Vraisemblance Gaussien Exact et AR(1)

- Des densités conditionnelles découle la fonction de vraisemblance

$$L_n(\vartheta; x_1, \dots, x_n) = \frac{(1 - \phi_1^2)^{1/2}}{(\sigma_\varepsilon \sqrt{2\pi})^n} \exp\left(-\frac{(1 - \phi_1^2)x_1^2}{2\sigma_\varepsilon^2}\right) \prod_{t=2}^n \exp\left(-\frac{\varepsilon_t^2}{2\sigma_\varepsilon^2}\right)$$

- Dont l'écriture sous forme de log-vraisemblance donne

$$\ell_n(\vartheta; x_1, \dots, x_n) = -\frac{1}{2} \ln\left(\frac{(1 - \phi_1^2)}{(\sigma_\varepsilon \sqrt{2\pi})^n}\right) - \frac{1}{2\sigma_\varepsilon^2} \left((1 - \phi_1^2)x_1^2 + \sum_{t=2}^n \varepsilon_t^2\right)$$

avec $\vartheta = (\phi_1, \sigma_\varepsilon^2)'$

- Le second terme fait intervenir : $RSS_c := \sum_{t=2}^n \varepsilon_t^2$
- On peut construire un OLS conditionnel basé sur la minisation de RSS_c
- Ou un OLS non-conditionnel basé sur $RSS_c + (1 - \phi_1^2)x_1^2$

Note Seul le MLE permet de gérer la non-linéarité engendrée par $q > 0$

$$\hat{\vartheta} = \arg \max_{\vartheta \in \mathbb{R}} \ell_n(\vartheta; x_1, \dots, x_n)$$

Maximum de Vraisemblance Gaussien Exact et $AR(p)$

- Dans le cas d'un $AR(p)$, il nous faut isoler

$$f_{X_p, X_{p-1}, \dots, X_1}(x_p, x_{p-1}, \dots, x_1; \vartheta)$$

Rappel La vraisemblance Gaussienne de $\mathbf{X}_n = (X_1, \dots, X_n)'$ s'écrit

$$\ell_n(\vartheta; X_1, \dots, X_n) = (2\pi)^{-n/2} (|\Gamma_n|)^{-1/2} \exp \left(-1/2 \mathbf{X}_n' \Gamma_n^{-1} \mathbf{X}_n \right)$$

- Pour $\mathbf{X}_p = (X_1, \dots, X_p)'$ on a

$$\Gamma_p = \begin{pmatrix} \gamma_0 & \gamma_1 & \dots & \gamma_{p-1} \\ \gamma_1 & \gamma_0 & \dots & \gamma_{p-2} \\ \vdots & \vdots & \dots & \vdots \\ \gamma_{p-1} & \gamma_{p-2} & \dots & \gamma_1 \end{pmatrix}$$

et donc la densité Gaussienne des p premières observations devient

$$f_{X_p, X_{p-1}, \dots, X_1}(x_p, x_{p-1}, \dots, x_1; \vartheta) = (2\pi)^{-p/2} (|\Gamma_p|)^{-1/2} \exp \left(-\frac{1}{2} \mathbf{X}_p' \Gamma_p^{-1} \mathbf{X}_p \right)$$

menant à la densité Gaussienne complète $f_{X_n, \dots, X_1}(x_n, \dots, x_1; \vartheta) =$

$$f_{X_p, X_{p-1}, \dots, X_1}(x_p, x_{p-1}, \dots, x_1; \vartheta) \prod_{t=p+1}^n f_{X_t | X_{t-1}, \dots, X_{t-p}}(x_t | x_{t-1}, \dots, x_{t-p}; \vartheta)$$

Maximum de Vraisemblance Gaussien conditionnel et AR(p)

- Il est possible de considérer les p observations comme déterministes
- En conditionnant la densité complète sur ces p observations on obtient

$$f_{X_n, \dots, X_{p+1} | X_1, \dots, X_p}(x_n, \dots, x_{p+1} | x_1, \dots, x_p; \vartheta) = \prod_{t=p+1}^n f_{X_t | X_{t-1}, \dots, X_{t-p}}(x_t | x_{t-1}, \dots, x_{t-p}; \vartheta)$$

- Dès lors, dans la construction de la vraisemblance, on sacrifiera

$$\mathbf{X}_p = (X_1, \dots, X_p)'$$

- Mais le problème d'optimisation s'en trouvera simplifié

$$\ell_n(\vartheta; x_p, \dots, x_n) = -\frac{1}{2} \ln \left(\frac{1}{(\sigma_\varepsilon \sqrt{2\pi})^n} \right) - \frac{1}{2\sigma_\varepsilon^2} \left(\sum_{t=p+1}^n \varepsilon_t^2 \right)$$

et il y aura équivalence avec les OLS

Maximum de Vraisemblance Gaussien conditionnel et MA(q)

- Partons de X_1 où

$$X_1 = \varepsilon_1 + \theta_1 \varepsilon_0 + \dots + \theta_q \varepsilon_{-q+1}$$

sous l'hypothèse que $\boldsymbol{\varepsilon}_0 = (\varepsilon_0, \varepsilon_{-1}, \dots, \varepsilon_{-q+1})' = 0$, qui nous donne

$$(X_1 | \boldsymbol{\varepsilon}_0 = 0) \sim \mathcal{N}(0, \sigma_\varepsilon^2) \Rightarrow f_{X_1 | \boldsymbol{\varepsilon}_0}(x_1 | \boldsymbol{\varepsilon}_0; \vartheta) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(X_1^2 - \varepsilon_1^2)}{2\sigma_\varepsilon^2}\right)$$

Pour X_2 , notons que X_1 et donc ε_1 est observable

$$X_2 = \varepsilon_2 + \theta_1 \varepsilon_1 + \dots + \theta_q \varepsilon_{-q+2}$$

ce qui nous donne

$$(X_2 | X_1 = x_1, \boldsymbol{\varepsilon}_0 = 0) \sim \mathcal{N}(\theta_1 \varepsilon_1, \sigma_\varepsilon^2)$$

$$\text{et donc } f_{X_2 | X_1, \boldsymbol{\varepsilon}_0}(x_2 | x_1, \boldsymbol{\varepsilon}_0; \vartheta) = \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(X_2 - \theta_1 \varepsilon_1)^2}{2\sigma_\varepsilon^2}\right)$$

$$= \frac{1}{\sqrt{2\pi\sigma_\varepsilon^2}} \exp\left(-\frac{(\varepsilon_2)^2}{2\sigma_\varepsilon^2}\right)$$

En itérant on obtient, de la même manière que pour les AR, la vraisemblance conditionnelle

$$\ell_n(\vartheta; x_1, \dots, x_n) = -\frac{n}{2} \ln\left(\frac{1}{(\sigma_\varepsilon \sqrt{2\pi})}\right) - \frac{1}{2\sigma_\varepsilon^2} \left(\sum_{t=1}^n \varepsilon_t^2\right)$$

Vraisemblance Gaussienne et ARMA(p, q)

- Pour $p > 0$ et $q > 0$, en se basant sur l'AI (Ch. 1) et

$$\mathbb{E}(X_{n+1} - \hat{X}_{n+1})^2 = \sigma_\varepsilon^2 r_n$$

une écriture générale de la vraisemblance Gaussienne est possible

$$L_n(\vartheta; x_1, \dots, x_n) = \frac{1}{\sqrt{(2\pi\sigma_\varepsilon^2)r_0 \dots r_{n-1}}} \exp \left(-\frac{1}{2\sigma_\varepsilon^2} \sum_{j=p+1}^n \frac{(x_j - \hat{x}_j)^2}{r_{j-1}} \right)$$

- Les estimateurs du maximum de vraisemblance (Gaussien) sont alors

$$\hat{\sigma}_\varepsilon^2 = n^{-1} S(\hat{\vartheta}_{p,q}; x_1, \dots, x_n) \text{ où } S(\hat{\vartheta}_{p,q}; x_1, \dots, x_n) = \frac{(x_j - \hat{x}_j)^2}{r_{j-1}}$$

et

$$\hat{\vartheta}_{p,q} = (\hat{\phi}_1, \dots, \hat{\phi}_p, \hat{\theta}_1, \dots, \hat{\theta}_q)'$$

sont les valeurs qui minimisent (l'opposée de) la log-vraisemblance Gaussienne concentrée

$$\ell_n(\vartheta_{p,q}; x_1, \dots, x_n) = \ln(n^{-1} S(\vartheta_{p,q}; x_1, \dots, x_n)) + n^{-1} \sum_{j=p+1}^n \ln r_{j-1}$$

Note Concentrer : résoudre sur un paramètre puis le faire dépendre des autres

Inférence et vraisemblance Gaussienne d'un ARMA(p, q)

- Pour $n \rightarrow \infty$ on peut montrer que $r_n = 1$ et

$$\hat{\vartheta} \xrightarrow{a.a.d} \mathcal{N}(\vartheta, n^{-1}\mathbb{V}(\vartheta))$$

La covariance $n^{-1}\mathbb{V}(\vartheta)$ peut être approximée via

$$\hat{H}_n(\vartheta; x_1, \dots, x_n)^{-1}$$

la Hessienne numérique évaluée par l'optimiseur à l'optimum

⇒ Les écarts-type seront donc aussi évaluables

- Pour des ARMA simples, la variance asymptotique est disponible

⇒ Pour n grand, l'usage direct de $n^{-1}\mathbb{V}(\vartheta)$ sera donc possible

Note Si n est petit, vous verrez au S2 qu'il existe des techniques de Bootstrap

Variance asymptotique et vraisemblance Gaussienne

- Pour un $\text{AR}(p)$ stationnaire, $\mathbb{V}(\vartheta)$ est identique à celle de Yule-Walker

$$\mathbb{V}(\vartheta_p) = \sigma_\varepsilon^2 \Gamma_p^{-1}$$

et plus particulièrement pour un $\text{AR}(1)$ et un $\text{AR}(2)$ on obtient

$$\mathbb{V}(\vartheta_1) = (1 - \phi_1^2) \text{ et } \mathbb{V}(\vartheta_2) = \begin{pmatrix} (1 - \phi_1^2) & -\phi_1(1 + \phi_2) \\ -\phi_1(1 + \phi_2) & (1 - \phi_2^2) \end{pmatrix}$$

- Pour un $\text{MA}(q)$ inversible, il faut adopter une écriture en $\text{AR}(q)$

$$X_t + \theta_1 X_{t-1} + \dots + \theta_q X_{t-q} = \varepsilon_t$$

et il peut être montré que $\mathbb{V}(\vartheta_q) = \sigma_\varepsilon^2 \Gamma_q^{-1}$ et e.g. pour $q = 1$ et $q = 2$

$$\mathbb{V}(\vartheta_1) = (1 - \theta_1^2) \text{ et } \mathbb{V}(\vartheta_2) = \begin{pmatrix} (1 - \theta_1^2) & \theta_1(1 - \theta_2) \\ \theta_1(1 - \theta_2) & (1 - \theta_2^2) \end{pmatrix}$$

- Pour un $\text{ARMA}(1, 1)$ stationnaire causal on peut montrer que

$$\mathbb{V}(\vartheta_{1,1}) = \frac{1 + \phi_1 \theta_1}{(\phi_1 + \theta_1)^2} \begin{pmatrix} (1 - \phi_1^2)(1 + \phi_1 \theta_1) & -(1 - \theta_1^2)(1 - \phi_1^2) \\ -(1 - \theta_1^2)(1 - \phi_1^2) & (1 - \theta_1^2)(1 + \phi_1 \theta_1) \end{pmatrix}$$

Estimation d'un AR(1) et inférence

- Considérons les rendements du Dow Jones journalier sur quelques mois

$$n = 77$$

- L'ACF et la PACF nous révèlent une faible dépendance au passé
- L'estimation par MLE d'un AR(1) nous donne

$$X_t = 0.4471X_{t-1} + \varepsilon_t$$

et le logiciel utilisé nous retourne également $\hat{\sigma}_{\phi_1} = 0.1050$

- Pour l'écart-type asymptotique, $\sigma_{\phi_1} = \sqrt{\mathbb{V}(\phi_1)n^{-1}}$, on trouve

$$\sqrt{(1 - 0.4471^2)/77} = 0.1019$$

$\Rightarrow \sigma_{\phi_1}$ et $\hat{\sigma}_{\phi_1}$ sont relativement proches

- Pour le calcul de l'intervalle de confiance à 95% on obtient

$$0.4471 \pm 1.96 \times 0.1050 = (0.2413, 0.6529)$$

ou

$$0.4471 \pm 1.96 \times 0.1019 = (0.2473, 0.6468)$$

Plan

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE

- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

Comment sélectionner un modèle

- Dans l'exemple S28, on peut se demander si l'AR(1) est approprié
 - ⇒ plusieurs réponses existent à la question de la sélection des retards
 - 1 Dans un premier temps, les examens des ACF et PACF ont été utilisés
 - ⇒ approche à la Box et Jenkins pour déterminer si MA, AR, ou ARMA
 - 2 Les critères d'information sont plus fiables pour le choix de p et q
 - ⇒ on présentera AIC, BIC et HQ
 - 3 Des tests de validation suivent ces procédures d'identifications
 - ⇒ divers tests sur les résidus

La méthodologie de Box et Jenkins

- L'approche Box et Jenkins s'inscrit dans la recherche de la parcimonie
- ⇒ comment prédire avec le minimum de paramètres
- On peut résumer la méthodologie de BJ en 4 étapes
 - 1 Transformer les données pour satisfaire l'hypothèse de stationnarité
 - 2 Sélectionner p, q petits pour décrire la série à l'aide d'un $\text{ARMA}(p, q)$
 - 3 Estimer les paramètres AR et MA du modèle
 - 4 Procéder à une validation de la spécification retenue
 - 5 Prévisions
 - L'étape 1 sera détaillée au C3 et l'étape 3 relève de la section précédente
 - L'étape 2 est appelée étape d'identification par BJ et repose sur
$$\hat{\rho}_j \text{ et } \hat{\phi}_{j,j}, \quad j = 1, 2, \dots$$
dont les comportements peuvent nous renseigner sur p et q
 - L'étape 4 est détaillée dans la section suivante

L'identification de p chez Box et Jenkins

Rappel Au C1 nous avons vu que pour un $\text{AR}(p > 0)$

$$p = \inf\{j | \phi_{j,j} = 0\}$$

- Dans la pratique il faut donc estimer $\phi_{j,j}$ pour $j = 1, 2, \dots, h$ et tester

$$H_0 : \phi_{1,1} = 0 \text{ versus } H_1 : \phi_{1,1} \neq 0$$

- D'après le S15 on sait que si $|\hat{\phi}_{1,1}| > 1.96/\sqrt{n}$ on rejette H_0

\Rightarrow si H_0 n'est pas rejeté on conclut que $p < 1$, mais si on rejette, on teste

$$H_0 : \phi_{2,2} = 0 \text{ versus } H_1 : \phi_{2,2} \neq 0$$

- De nouveau, si $|\hat{\phi}_{2,2}| > 1.96/\sqrt{n}$ on rejette H_0

\Rightarrow si on ne rejette pas, $p = 1$, mais si on rejette, on teste

$$H_0 : \phi_{3,3} = 0 \text{ versus } H_1 : \phi_{3,3} \neq 0$$

- Cette procédure se prolonge ainsi jusqu'à arriver à

$$H_0 : \phi_{h,h} = 0 \text{ versus } H_1 : \phi_{h,h} \neq 0$$

avec $h = p + 1$

L'identification de q chez Box et Jenkins

Rappel Au C1 nous avons vu que pour un $MA(q > 0)$

$$q = \inf\{j | \rho_j = 0\}$$

- La même stratégie peut donc être mise en place en estimant ρ_j
- Pour un $MA(q)$ Gaussien, la formule de Bartlett (cf. S7) nous donne

$$\mathbb{V}(\hat{\rho}_j) = \frac{1}{n} \left(1 + 2 \sum_{i=1}^q \rho_i^2 \right), \quad j = q+1, q+2, \dots$$

où dans la pratique on remplacera ρ_i par $\hat{\rho}_i$

- La procédure sera alors de commencer par estimer ρ_1 et de tester

$$H_0 : \rho_1 = 0 \text{ versus } H_1 : \rho_1 \neq 0$$

sachant que si $|\hat{\rho}_1| > 1.96/\sqrt{n}$ on rejette H_0

\Rightarrow si H_0 n'est pas rejeté on conclut que $q < 1$, mais si on rejette, on teste

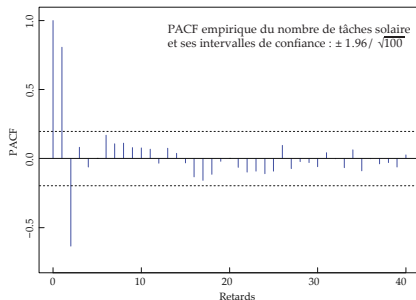
$$H_0 : \rho_2 = 0 \text{ versus } H_1 : \rho_2 \neq 0$$

sachant que si $|\hat{\rho}_2| > 1.96\sqrt{n^{-1}(1 + 2\hat{\rho}_1^2)} = 1.96\sqrt{\mathbb{V}(\hat{\rho}_2)}$ on rejette H_0

- Cette procédure se prolonge ainsi jusqu'à

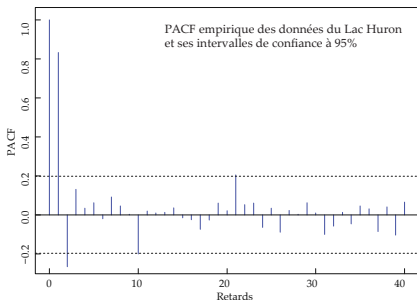
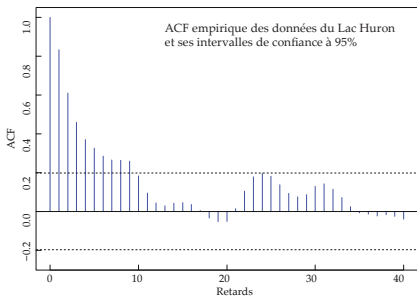
$$H_0 : \rho_h = 0 \text{ versus } H_1 : \rho_h \neq 0, \text{ avec } h = q+1$$

Exemple d'analyse de la PACF chez Box et Jenkins



- Soit la séquence du nombre de Wolf (1770-1869) et sa PACF
 - Analysons les PACFs significatives (attention, $\phi_{0,0} = \rho_0 = 1$)
 - Au delà de 2 retards ($\phi_{1,1}$ et $\phi_{2,2}$), les PACFs sont dans les intervalles
- ⇒ Cela suggère qu'un modèle candidat est serait l'AR(2)

Exemple d'analyse graphique chez Box et Jenkins



- Soit les données du Lac Huron vues au S8
 - Analysons les ACFs et PACFs significatives
 - La décroissance de l'ACF est graduelle ne suggérant pas un MA
 - Inversement, une rupture dans la PACF s'observe au delà de 2 retards
- ⇒ Cela suggère de nouveau qu'un modèle candidat serait l'AR(2)

La philosophie du critère AIC (Akaike Information Criterion)

- L'idée de l'AIC est de minimiser une divergence distributionnelle

$$d(\tilde{\vartheta}|\vartheta) = \Delta(\tilde{\vartheta}|\vartheta) - \Delta(\vartheta|\vartheta)$$

où la distance de Kullback-Leibler

$$\Delta(\tilde{\vartheta}|\vartheta) = \mathbb{E}(-2 \ln f(\mathbf{X}; \tilde{\vartheta})) = \int_{\mathbb{R}^n} -2 f(\mathbf{x}; \vartheta) \ln f(\mathbf{x}; \tilde{\vartheta}) d\mathbf{x}$$

mesure la dissimilarité entre la famille de fonctions de densités

$$\{f(\mathbf{x}; \tilde{\vartheta}), \tilde{\vartheta} \in \Theta\}$$

dont on suppose que $\mathbf{X} = (X_1, \dots, X_n)'$ est tiré, et $f(\mathbf{x}; \vartheta)$

- En effet on peut voir que par l'inégalité de Jensen,

$$\begin{aligned} d(\tilde{\vartheta}|\vartheta) &= \int_{\mathbb{R}^n} -2 f(\mathbf{x}; \vartheta) \ln \left(\frac{f(\mathbf{x}; \tilde{\vartheta})}{f(\mathbf{x}; \vartheta)} \right) d\mathbf{x} \\ &\geq -2 \ln \left(\int_{\mathbb{R}^n} \frac{f(\mathbf{x}; \tilde{\vartheta})}{f(\mathbf{x}; \vartheta)} f(\mathbf{x}; \vartheta) d\mathbf{x} \right) \\ &= -2 \ln \left(\int_{\mathbb{R}^n} f(\mathbf{x}; \tilde{\vartheta}) d\mathbf{x} \right) = 0 \end{aligned}$$

où l'inégalité de Jensen devient une égalité si $f(\mathbf{x}; \tilde{\vartheta}) = f(\mathbf{x}; \vartheta)$

La construction du critère AIC corrigé

- Bien sur $d(\tilde{\vartheta}|\vartheta)$ doit être estimé et pour cela on suppose la Normalité
- \Rightarrow Pour $\vartheta = (\vartheta_{p,q}, \sigma_\varepsilon^2)'$, $f(\cdot; \vartheta)$ est la fonction de densité de

$$\mathbf{Y} = (Y_1, \dots, Y_n)'$$

où Y_t est un ARMA(p, q) Gaussien de paramètre $\vartheta_{p,q}$ et σ_ε^2

- Supposons que $\mathbf{x} = (x_1, \dots, x_n)'$ provient d'un ARMA Gaussien avec p, q et ϑ ses vrais paramètres
- Notons $\hat{\vartheta}$ l'estimateur MLE basé sur \mathbf{X} , alors

$$-2 \ln L_n(\hat{\vartheta}; \mathbf{y}) = -2 \ln L_n(\hat{\vartheta}; \mathbf{x}) + \hat{\sigma}_\varepsilon^{-2} S(\hat{\vartheta}_{p,q}; \mathbf{y}) - n$$

et donc on peut écrire $\mathbb{E}_\vartheta(\Delta(\tilde{\vartheta}|\vartheta))$ comme

$$\mathbb{E}_\vartheta(-2 \ln L_n(\hat{\vartheta}; \mathbf{y})) = \mathbb{E}_\vartheta(-2 \ln L_n(\hat{\vartheta}; \mathbf{x})) + \mathbb{E}_\vartheta(\hat{\sigma}_\varepsilon^{-2} S(\hat{\vartheta}_{p,q}; \mathbf{y})) - n$$

dont l'approximation asymptotique du terme

$$\mathbb{E}_\vartheta(\hat{\sigma}_\varepsilon^{-2} S(\hat{\vartheta}_{p,q}; \mathbf{y})) = \frac{2(p+q+1)n}{n-p-q-2}$$

nous fourni un estimateur non-biaisé de la distance de KL espérée

$$AICc = -2 \ln L_n(\hat{\vartheta}; \mathbf{x}) + \frac{2(p+q+1)n}{n-p-q-2}$$

Critère AIC versus AICc

- A l'origine, le critère AIC fut proposé comme

$$AIC = -2 \ln L_n(\hat{\vartheta}; \mathbf{x}) + 2(p + q + 1)$$

⇒ Asymptotiquement AIC et AICc sont équivalents

Note 1 En échantillon fini, des simulations montrent que l'AIC surestime p

⇒ phénomène d'overfitting

Note 2 Si le vrai p est grand, l'AICc sous-estime souvent p

⇒ phénomène overfitting inverse

Note 3 Les deux critères sont basés sur deux éléments

⇒ “qualité d'ajustement” (vraisemblance) + “pénalité” (# de paramètres)

Critères BIC et HQ

- De nombreux autres critères existent dont le BIC et l'HQ

BIC Le Bayesian Information Criterion se distingue par sa pénalité

$$BIC = -2 \ln L_n(\hat{\vartheta}; \mathbf{x}) + (p + q + 1) \ln(n)$$

HQ Le critère d'Hannan-Quinn se distingue aussi par sa pénalité

$$HQ = -2 \ln L_n(\hat{\vartheta}; \mathbf{x}) + (p + q + 1) \ln(\ln(n))$$

- Comparaison avec AICc :

- BIC et HQ sont consistants (pas AICc) : $\tilde{\vartheta} \in \Theta$

$\Rightarrow \hat{p} \rightarrow p$ et $\hat{q} \rightarrow q$ avec une probabilité unitaire

- AICc est efficace, i.e. minimise la MSE (pas BIC ni HQ) : $\tilde{\vartheta} \notin \Theta$

\Rightarrow il minimise le risque de choisir un très mauvais modèle

Comparaison par simulation

- Soit des simulations Monte Carlo : AR(1), AR(2), MA(2), ARMA(1, 1)

⇒ 10000 simulations de $n \in \{30, 60, 180, 500\}$ observations

- On fixe $p_{\max} = q_{\max} = 4$ et on applique chacun des 3 critères

⇒ les tableaux suivants reportent les % de sélection de p et q par critère

Note V désigne le nombre de procédures d'estimation ayant abouti

DGP : AR(1)		0	1	P 2	3	4
n = 30	AIC	1%	71%	14%	7%	7%
	BIC	2%	87%	7%	2%	1%
	HQ	1%	79%	11%	5%	4%
n = 60	AIC	0%	75%	13%	7%	6%
	BIC	0%	94%	5%	1%	0%
	HQ	0%	86%	9%	3%	2%
n = 180	AIC	0%	75%	13%	7%	6%
	BIC	0%	97%	2%	0%	0%
	HQ	0%	90%	7%	2%	1%
n = 500	AIC	0%	77%	12%	6%	5%
	BIC	0%	98%	2%	0%	0%
	HQ	0%	93%	5%	1%	0%

DGP : AR(2)		0	1	P 2	3	4
n = 30	AIC	11%	28%	42%	11%	8%
	BIC	21%	39%	34%	4%	2%
	HQ	14%	33%	40%	8%	5%
n = 60	AIC	1%	8%	70%	12%	9%
	BIC	3%	22%	71%	3%	1%
	HQ	1%	14%	74%	7%	4%
n = 180	AIC	0%	0%	78%	14%	8%
	BIC	0%	0%	97%	2%	0%
	HQ	0%	0%	91%	7%	2%
n = 500	AIC	0%	0%	78%	13%	9%
	BIC	0%	0%	99%	1%	0%
	HQ	0%	0%	93%	5%	2%

DGP : MA(2)		0	1	q 2	3	4
n = 30	AIC	17%	15%	42%	14%	12%
	BIC	36%	18%	35%	6%	5%
	HQ	23%	16%	41%	11%	9%
n = 60	V	10000	9977	9902	9245	8076
	AIC	2%	5%	67%	15%	11%
	BIC	12%	13%	69%	4%	2%
n = 180	HQ	5%	8%	72%	9%	6%
	V	10000	10000	9995	9961	9802
n = 500	AIC	0%	0%	77%	13%	9%
	BIC	0%	0%	97%	3%	0%
	HQ	0%	0%	90%	7%	3%
n = 500	V	10000	10000	10000	10000	10000
	AIC	0%	0%	78%	14%	8%
	BIC	0%	0%	99%	1%	0%
n = 500	HQ	0%	0%	93%	5%	2%
	V	10000	10000	10000	10000	10000

Comparaison par simulation : ARMA(1,1)

DGP : ARMA(1,1)

n = 30							q								
		AIC	0	1	2	3	4			HQ	0	1	2	3	4
p	0	0%	1%	4%	2%	9%		0	16%	0%	0%	0%	0%	0%	
	1	4%	28%	5%	2%	4%		1	41%	0%	0%	0%	0%		
	2	9%	6%	1%	2%	1%		2	21%	0%	0%	0%	0%		
	3	8%	1%	1%	0%	0%		3	10%	0%	0%	0%	0%		
	4	13%	3%	1%	0%	0%		4	13%	0%	0%	0%	0%		
q							q								
		BIC	0	1	2	3	4			V	0	1	2	3	4
p	0	0%	1%	6%	2%	6%		0	10000	7676	7514	5159	5050		
	1	2%	39%	3%	1%	2%		1	10000	8066	7247	4955	3755		
	2	17%	4%	1%	1%	1%		2	10000	6726	2411	1937	993		
	3	6%	1%	0%	0%	0%		3	10000	6089	2432	608	248		
	4	6%	0%	0%	0%	0%		4	10000	5489	1790	452	99		
n = 60							q								
		AIC	0	1	2	3	4			HQ	0	1	2	3	4
p	0	0%	0%	0%	1%	5%		0	6%	0%	0%	0%	0%	0%	
	1	0%	44%	5%	2%	3%		1	66%	0%	0%	0%	0%	0%	
	2	1%	7%	4%	3%	3%		2	13%	0%	0%	0%	0%	0%	
	3	3%	2%	1%	2%	1%		3	7%	0%	0%	0%	0%	0%	
	4	7%	3%	1%	1%	0%		4	8%	0%	0%	0%	0%	0%	
q							q								
		BIC	0	1	2	3	4			V	0	1	2	3	4
p	0	0%	0%	1%	1%	3%		0	10000	8599	9271	7843	8215		
	1	0%	73%	3%	0%	1%		1	10000	9444	9228	8261	7667		
	2	3%	4%	1%	1%	1%		2	10000	8713	4472	4697	3861		
	3	3%	0%	0%	0%	0%		3	10000	8432	5193	2165	1520		
	4	3%	0%	0%	0%	0%		4	10000	7867	4381	1902	698		

DGP : ARMA(1,1)

n = 180							HQ									
		AIC	0	1	q	2	3	4			0	1	q	2	3	4
P	0	0%	0%	0%	0%	0%	0%	0%	p	0	0%	0%	0%	0%	0%	0%
	1	0%	46%	4%	2%	2%	2%	1		79%	0%	0%	0%	0%	0%	0%
	2	0%	5%	9%	2%	2%	2%	2		12%	0%	0%	0%	0%	0%	0%
	3	0%	2%	2%	11%	3%	3%	3		7%	0%	0%	0%	0%	0%	0%
		4	0%	2%	1%	3%	3%			4	3%	0%	0%	0%	0%	0%
BIC							V									
		0	1	q	2	3	4			0	1	q	2	3	4	
P	0	0%	0%	0%	0%	0%	0%	p	0	10000	9711	9958	9663	9911		
	1	0%	92%	1%	0%	0%	1		10000	9996	9989	9948	9935			
	2	0%	2%	2%	0%	0%	2		10000	9908	7304	7870	7451			
	3	0%	0%	0%	1%	0%	3		10000	9913	8500	6059	5577			
		4	0%	0%	0%	0%	0%			4	10000	9796	7739	5674	3250	
n = 500							HQ									
		AIC	0	1	q	2	3	4			0	1	q	2	3	4
P	0	0%	0%	0%	0%	0%	0%	0%	p	0	0%	0%	0%	0%	0%	0%
	1	0%	41%	3%	2%	1%	1	82%		0%	0%	0%	0%	0%	0%	
	2	0%	4%	9%	2%	1%	2	9%		0%	0%	0%	0%	0%	0%	
	3	0%	2%	2%	17%	5%	3	7%		0%	0%	0%	0%	0%	0%	
		4	0%	1%	1%	4%	6%			4	2%	0%	0%	0%	0%	0%
BIC							V									
		0	1	q	2	3	4			0	1	q	2	3	4	
P	0	0%	0%	0%	0%	0%	0%	p	0	10000	9991	9997	9954	9994		
	1	0%	96%	1%	0%	0%	1		10000	10000	10000	10000	10000			
	2	0%	1%	1%	0%	0%	2		10000	9998	8650	9002	8717			
	3	0%	0%	0%	0%	0%	3		10000	9997	9380	8143	7778			
		4	0%	0%	0%	0%	0%			4	10000	9994	8983	7967	5808	

Limites de l'utilisation des critères

- On considère de nouveau les données du lac Huron (cf. S8)

- Si l'économètre balaye uniquement sur $p > 0$:

⇒ Les AICc et BIC minimum sont trouvés pour $p = 2$ et on obtient

$$X_t - 1.0441X_{t-1} + 0.2503X_{t-2} = \varepsilon_t, \quad \varepsilon_t \sim \mathcal{N}(0, 0.4789)$$

avec $AICc = 213.54$ et $BIC = 217.63$

- Si l'économètre balaye sur $p > 0$ et $q > 0$:

⇒ Les AICc et BIC minimum sont trouvés pour $p = 1$ et $q = 1$ et on obtient

$$X_t - 0.7446X_{t-1} = \varepsilon_t + 0.3213\varepsilon_{t-1}, \quad \varepsilon_t \sim \mathcal{N}(0, 0.4750)$$

avec $AICc = 212.77$ et $BIC = 217.86$

Note Les critères étant très proches, difficile de déterminer le meilleur modèle

Plan

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE

- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

Identification

- Il est possible d'intégrer des variables exogènes, \mathbf{X}_t dans un ARMA
- ⇒ mais l'identification et l'estimation du processus se complexifient

- Pour comprendre cela considérons l'ARMAX suivant

$$Y_t = \alpha_0 X_t + \alpha_1 X_{t-1} + \dots + \alpha_m X_{t-m} + \nu_t$$

où ν_t suit un ARMA(p, q) indépendant $\forall t$ de X_t

- Sous l'hypothèse que $X_t \sim WN(0, \sigma_X^2)$, on a

$$\mathbb{E}(Y_t X_t) = \alpha_0 \sigma_X^2 \Rightarrow \text{Corr}(Y_t, X_t) = \alpha_0 \sigma_X \sigma_Y^{-1}$$

$$\mathbb{E}(Y_t X_{t-1}) = \alpha_1 \sigma_X^2 \Rightarrow \text{Corr}(Y_t, X_{t-1}) = \alpha_1 \sigma_X \sigma_Y^{-1}$$

$$\mathbb{E}(Y_t X_{t-2}) = \alpha_2 \sigma_X^2 \Rightarrow \text{Corr}(Y_t, X_{t-2}) = \alpha_2 \sigma_X \sigma_Y^{-1}$$

ce qui implique que $\text{Corr}(Y_t, X_{t-j})$ est proportionnelle à $\partial Y_t / (\partial X_{t-j})$

- ⇒ on peut identifier les retards de X_t entrant dans le modèle de Y_t

Pre-whitening

- Mais cette identification n'est possible que si $X_t \sim WN(0, \sigma_X^2)$

⇒ Si ce n'est pas le cas, il faut blanchir X_t de toute dépendance ...

... sans affecter sa relation avec Y_t , c'est le **pre-whitening**

- Pour comprendre cette étape considérons à présent que $X_t \sim ARMA(p, q)$

$$Y_t = \alpha(L)X_t + \nu_t, \quad \Phi_X(L)X_t = \Theta_X(L)u_t$$

où $\nu_t \sim ARMA(p, q)$ indépendant $\forall t$ de X_t , u_t est un bruit blanc et

$$u_t = \Theta_X(L)^{-1}\Phi_X(L)X_t$$

Il suffit alors de multiplier Y_t par le filtre $\Theta_X(L)^{-1}\Phi_X(L)$ pour obtenir

$$\Theta_X(L)^{-1}\Phi_X(L)Y_t = \alpha(L)\Theta_X(L)^{-1}\Phi_X(L)X_t + \Theta_X(L)^{-1}\Phi_X(L)\nu_t$$

ce qui nous donne

$$\tilde{Y}_t = \alpha(L)u_t + \tilde{\nu}_t$$

où u_t est bien un bruit blanc et les $\alpha(L)$ sont donc identifiables

Estimation en étapes

- L'estimation des ARMAX nécessite donc plusieurs étapes
 - identification du processus ARMA afférent X_t pour déterminer
$$\Theta_X(L)^{-1}\Phi_X(L)$$
 - blanchiment de X_t par application du filtre $\Theta_X(L)^{-1}\Phi_X(L)$ à X_t et Y_t
 - calcul des corrélations croisées entre \tilde{Y}_t et u_t
- ⇒ les corrélations non-nulles signalent les lags devant intégrer le modèle
 - identification du processus ARMA afférent à \tilde{Y}_t

Note 1 L'inférence sur les corrélations croisées est simplifiée puisque

$$u_t \sim WN(0, \sigma_u^2)$$

⇒ ce qui nous ramène à la formule de Bartlett simplifiée vue au S9

Fonction de transfert générale

- Soit $\mathbf{X}_t = X_{1,t}, \dots, X_{k,t}$ dont tous les éléments sont orthogonaux
- La fonction de transfert d'un ARMAX sur Y_t sera alors

$$Y_t = \frac{\alpha_1(L)}{\beta_1(L)} X_{1,t-d_1} + \dots + \frac{\alpha_k(L)}{\beta_k(L)} X_{k,t-d_k} + \frac{\Theta_Y(L)}{\Phi_Y(L)} \varepsilon_t$$

où $X_{i,t-d_i}$ indique que l'exogène rentre dans le modèle avec un retard qui lui est propre

Note 1 Les polynômes $\beta_k(L)$ ajoute de la généralité mais le plus souvent

$$\beta(L) = \Phi_Y(L)^{-1}, \quad \forall k$$

Note 2 En effet rappelons que si l'on part du modèle

$$\Phi_Y(L)Y_t = \alpha(L)\mathbf{X}_t + \Theta_Y(L)\varepsilon_t$$

et donc

$$Y_t = \alpha(L)\Phi_Y(L)^{-1}\mathbf{X}_t + \Theta_Y(L)\Phi_Y(L)^{-1}\varepsilon_t$$

Exemple de fonction de transfert

- Soit $Y_t = 0.6Y_{t-1} + 1.2X_{t-2} + \varepsilon_t$ et $X_t = 0.8X_{t-1} + u_t + 0.5u_{t-1}$

- En utilisant la représentation en fonction de transfert on obtient

$$Y_t = \frac{1.2}{1 - 0.6L} X_{t-2} + \frac{1}{1 - 0.6L} \varepsilon_t$$

- Les divisions polynômiales engendrent une décroissance infinie

$$\begin{aligned} Y_t &= 1.2(1 + 0.6L + 0.6^2L^2 + 0.6^3L^3 + \dots)X_{t-2} \\ &\quad + (1 + 0.6L + 0.6^2L^2 + 0.6^3L^3 + \dots)\varepsilon_t \\ &= 1.2X_{t-2} + 0.72X_{t-3} + 0.432X_{t-4} + 0.2592X_{t-5} + \dots \\ &\quad + \varepsilon_t + 0.6\varepsilon_{t-1} + 0.36\varepsilon_{t-2} + 0.216\varepsilon_{t-3} + 0.1296\varepsilon_{t-4} + \dots \end{aligned}$$

dans la dépendance entre Y_t et X_{t-i} , $i > 2$

- ⇒ On s'attend donc à voir une fonction d'autocorrélation croisée nulle en $i = 0, 1$ et non-nulle pour $i > 2$

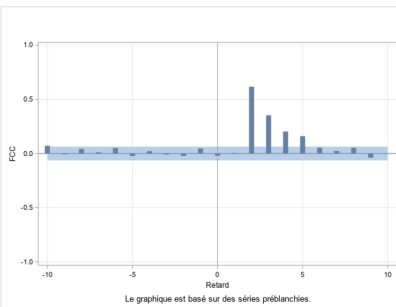
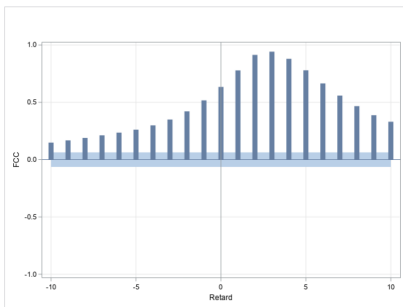
Importance du prewhitening

- Poursuivons avec l'exemple précédent et construisons \tilde{Y}_t et u_t

$$\tilde{Y}_t = (1 + 0.5L)^{-1}(1 - 0.8L)Y_t$$

$$u_t = (1 + 0.5L)^{-1}(1 - 0.8L)X_t$$

- Comparons à présent $\rho(\tilde{Y}_t, u_{t-i})$ et $\rho(Y_t, X_{t-i})$



Plan

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE
- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

Tests de validation

- Les procédures d'identification ne suffisent pas ...

... et doivent être couplées à des tests de validation sur les résidus :

- Test de significativité et nullité jointe (pour rappel)
- Coefficient de détermination
- Test de nullité de moyenne des résidus
- Tests de Ljung-Box et de Box-Pierce
- Test de Jarque-Bera

⇒ En effet, si le modèle est bien spécifié on devrait avoir $\hat{\varepsilon}_t \sim WN(0, \hat{\sigma}^2)$

Note Ces tests pourraient également s'appliquer sur les observées

Test de significativité et nullité jointe

- Sous certaines hypothèses de régularité on a vu que le MLE

$$\sqrt{n}(\hat{\vartheta} - \vartheta_0) \xrightarrow{d} \mathcal{N}(0, \mathbb{V}(\hat{\vartheta}))$$

⇒ cela autorise la construction de tests inférenciels usuels car

$$\sqrt{n}((\hat{\vartheta} - \vartheta_0)/\sigma_{\hat{\vartheta}}) \sim \mathcal{N}(0, 1)$$

- Dans la pratique $\sigma_{\hat{\vartheta}}$ est inconnu et on considère $\hat{\sigma}_{\hat{\vartheta}}^2 \sim \chi^2(n-1)$

Rappel Si $X \sim \mathcal{N}(0, 1)$ et $Y \sim \chi^2(\nu)$, alors $Z = X(\sqrt{Y/\nu})^{-1} \sim t(\nu)$

⇒ Le test de Student est alors ainsi formulé :

$$\frac{\hat{\vartheta} - \vartheta_{H0}}{\hat{\sigma}_{\hat{\vartheta}}/n} \sim t(n-1)$$

- Le test de nullité jointe de Fisher s'applique de façon standard

⇒ si l'on souhaite tester

$$H_0 : \vartheta_1 = \dots = \vartheta_{p+q} = 0$$

contre $H_1 : \exists j$ tel que $\vartheta_j \neq 0$ la statistique de Fisher est

$$\frac{(SST - RSS)/(p+q)}{RSS/(n - (p+q) - 1)} \sim F(p+q, n - (p+q) - 1)$$

Coefficients de détermination

- Les coefficients de détermination usuels sont donnés par

$$R^2 = 1 - \frac{\sum_{t=1}^n \hat{\varepsilon}_t^2}{\sum_{t=1}^n X_t^2} \quad (1)$$

$$\bar{R}^2 = 1 - \frac{n-1}{n-p-q} \frac{\sum_{t=1}^n \hat{\varepsilon}_t^2}{\sum_{t=1}^n X_t^2} \quad (2)$$

où l'on préférera \bar{R}^2 qui tient compte des retards AR et MA

Test de nullité de moyenne des résidus

- Si $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, on s'attend à ce que $\mathbb{E}(\hat{\varepsilon}_t) = 0$, et donc

$$\bar{\varepsilon} = n^{-1} \sum_{t=1}^n \hat{\varepsilon}_t \xrightarrow{p} 0$$

Par application du théorème central limite on a donc

$$n^{1/2} \frac{\bar{\varepsilon}}{\sigma_{\bar{\varepsilon}}} \xrightarrow{d} \mathcal{N}(0, 1)$$

- On donc peut tester la nullité de $\bar{\varepsilon}$ en construisant

$$\bar{\varepsilon} \pm \Phi_{1-\alpha/2} n^{-1/2} \hat{\sigma}_{\varepsilon}$$

Tests de Ljung-Box et de Box-Pierce

- Si $\varepsilon_t \sim WN(0, \sigma_\varepsilon^2)$, on s'attend à ce que $\gamma(j) = \rho(j) = 0, \forall j > 0$
 - Plutôt qu'étudier chaque IC autour de $\rho(j)$ comme au S8...
- ... on peut construire une seule statistique (de type portmanteau) :

$$Q = n \sum_{j=1}^h \hat{\rho}^2(j)$$

- La statistique de Box-Pierce étudie le comportement limite de Q

$$Q_{BP} = n \sum_{j=1}^h \hat{\rho}^2(j) \sim \chi^2(h - p - q)$$

si on test $H_0 : \rho_1 = \dots = \rho_h = 0$ contre $H_1 : \exists j$ tel que $\rho_j \neq 0$

- La statistique de Ljung-Box est un raffinement de Q où

$$Q_{LB} = n(n+2) \sum_{j=1}^h \hat{\rho}^2(j)/(n-j) \sim \chi^2(h - p - q)$$

si on test $H_0 : \rho_1 = \dots = \rho_h = 0$ contre $H_1 : \exists j$ tel que $\rho_j \neq 0$

Note Si ces tests sont appliqués sur les observés, les distributions sont $\chi^2(h)$

Test de Jarque-Bera

- Si on suppose que $\varepsilon_t \sim \text{i.i.d. } \mathcal{N}(0, \sigma_\varepsilon^2)$, on peut tester cette hypothèse
- Le test de Jarque-Bera permet cela à travers une unique statistique

$$S_{JB} = \frac{n}{6} S_k + \frac{n}{24} (K_u - 3)^2 \xrightarrow{d} \chi^2(2)$$

où S_k et K_u représentent les coefficients de Skewness et Kurtosis resp.

\Rightarrow Si $S_{JB} \geq \chi_{1-\alpha}^2(2)$ on rejette H_0 de normalité des résidus au seuil de $\alpha\%$

Méthode du Maximum de Vraisemblance

- Partons d'un exemple : soit un échantillon $X_t = X_1, \dots, X_n \sim P(\vartheta)$
 - $P(\vartheta)$ dénote la distribution de Poisson dont la fonction de masse est

$$\Pr(X_i = x) = \frac{\exp(-\vartheta)\vartheta^x}{x!}, \quad \vartheta > 0, \quad \forall x \in \mathbb{N}$$

- Soit une réalisation de l'échantillon $x_t = x_1, \dots, x_n$
- La probabilité d'observer cette réalisation est

$$\Pr((X_1 = x_1) \cap \dots \cap (X_n = x_n))$$

- L'indépendance des tirages donne l'équivalence avec le produit des probabilités marginales

$$\Pr((X_1 = x_1) \cap \dots \cap (X_n = x_n)) = \prod_{i=1}^n \Pr(X_i = x_i)$$

Plan

- 1 Introduction à l'estimation
- 2 Newey-West
- 3 MLE
- 4 Sélection de modèle
- 5 ARIMAX
- 6 Tests de validation
- 7 Rappels sur le MLE

L'estimateur du Maximum de Vraisemblance

- En remplaçant par la fonction de masse de la loi de Poisson on obtient

$$\Pr((X_1 = x_1) \cap \dots \cap (X_n = x_n)) = \prod_{i=1}^n \frac{e^{-\vartheta} \vartheta^{x_i}}{x_i!} = e^{-n\vartheta} \frac{\vartheta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

- Il s'agit donc d'une fonction dépendant de x_1, \dots, x_n et de ϑ
- ϑ est un paramètre inconnu mais on observe x_1, \dots, x_n
- Par la suite on notera :

$$L_n(\vartheta; x_1, \dots, x_n) = \Pr((X_1 = x_1) \cap \dots \cap (X_n = x_n))$$

- Le principe du maximum de vraisemblance est le suivant :
 - Trouver le ϑ qui maximise la probabilité d'apparition de x_1, \dots, x_n
- L'estimateur du maximum de vraisemblance est donc :

$$\hat{\vartheta} = \arg \max_{\vartheta \in \mathbb{R}^+} L_n(\vartheta; x_1, \dots, x_n)$$

L'estimateur du Maximum de la log-Vraisemblance

- Dans le cas de l'exemple reposant sur la loi de Poisson on a

$$\hat{\vartheta} = \arg \max_{\vartheta \in \mathbb{R}^+} e^{-n\vartheta} \frac{\vartheta^{\sum_{i=1}^n x_i}}{\prod_{i=1}^n x_i!}$$

- La formule est complexe et la présence d'un produit n'arrange rien
- Simplifions le programme de maximisation en considérant la log-vraisemblance

$$\hat{\vartheta} = \arg \max_{\vartheta \in \mathbb{R}^+} \ln L_n(\vartheta; x_1, \dots, x_n)$$

- Dans le cadre de notre exemple la log-vraisemblance est

$$\ln L_n(\vartheta; x_1, \dots, x_n) = -n\vartheta + \ln(\vartheta) \sum_{i=1}^n x_i - \ln \left(\prod_{i=1}^n x_i! \right)$$

Conditions nécessaire et suffisante

■ La condition nécessaire répond à la question

- Le problème admet-il une solution ?

⇒ Pour répondre on annule la dérivée première par rapport à ϑ

$$\left. \frac{\partial \ln L_n(\vartheta; x_1, \dots, x_n)}{\partial \vartheta} \right|_{\hat{\vartheta}} = -n + \hat{\vartheta}^{-1} \sum_{i=1}^n x_i = 0 \iff \hat{\vartheta} = n^{-1} \sum_{i=1}^n x_i$$

- Ici, la log vraisemblance est maximisée par la moyenne empirique

■ La condition suffisante répond à la question

- Cette solution est-elle un maximum ?

⇒ Pour répondre on regarde le signe de la dérivée seconde par rapport à ϑ

$$\left. \frac{\partial^2 \ln L_n(\vartheta; x_1, \dots, x_n)}{\partial \vartheta^2} \right|_{\hat{\vartheta}} = -\hat{\vartheta}^{-2} \sum_{i=1}^n x_i < 0$$

- Négatif donc bien un maximum

Log-Vraisemblance Gaussienne

- Dans l'exemple, il s'agissait de variables aléatoires discrètes
- Dans le cas de variables aléatoires continues, l'intuition est la même
 - Néanmoins, on raisonnera sur la densité de la loi jointe des variables

$$L_n(\vartheta; x_1, \dots, x_n) = f_{X_1, \dots, X_n}(x_1, \dots, x_n; \vartheta)$$

- Soit une séquence $X_n \sim$ i.i.d. (μ, σ^2) selon une loi normale
- La densité de la loi normale implique 2 paramètres $\vartheta = (\mu, \sigma^2)'$

$$\begin{aligned} L_n(\vartheta; x_1, \dots, x_n) &= \prod_{i=1}^n (\sigma\sqrt{2\pi})^{-1} e^{-\frac{(x_i - \mu)^2}{2\sigma^2}} \\ &= (2\pi\sigma^2)^{-n/2} \exp\left(-\frac{\sum_{i=1}^n (x_i - \mu)^2}{2\sigma^2}\right) \end{aligned}$$

$$\ln L_n(\vartheta; x_1, \dots, x_n) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (x_i - \mu)^2$$

L'Estimateur du Maximum de Vraisemblance

■ Estimateur du maximum de vraisemblance

$$\hat{\vartheta} = \arg \max_{\vartheta \in \mathbb{R}^+} \ln L_n(\vartheta; x_1, \dots, x_n)$$

■ Hypothèses

- $\vartheta = (\mu, \sigma^2)'$ est identifiable : $\forall \vartheta^*, \vartheta$ avec $\vartheta^* \neq \vartheta$, les lois jointes de x_1, \dots, x_n sont différentes
- Condition nécessaire du gradient :

$$g_n(\hat{\vartheta}; x_1, \dots, x_n) = \left. \frac{\partial \ln L_n(\vartheta; x_1, \dots, x_n)}{\partial \vartheta} \right|_{\hat{\vartheta}} = 0$$

- Condition suffisante de la hessienne :

$$H_n(\hat{\vartheta}; x_1, \dots, x_n) = \left. \frac{\partial^2 \ln L_n(\vartheta; x_1, \dots, x_n)}{\partial \vartheta^2} \right|_{\hat{\vartheta}} < 0$$

Condition nécessaire du MLE gaussien

- Notons $\ln L_n(\vartheta; x_1, \dots, x_n) = \ell_n(\vartheta; x)$ et commençons par le gradient :

$$\begin{aligned} \frac{\partial \ell_n(\vartheta; x)}{\partial \vartheta} &= \begin{pmatrix} \frac{\partial \ell_n(\vartheta; x)}{\partial \mu} \\ \frac{\partial \ell_n(\vartheta; x)}{\partial \sigma^2} \end{pmatrix} = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (x_i - \mu)^2 \end{pmatrix} \\ \Rightarrow \left. \frac{\partial \ell_n(\vartheta; x)}{\partial \vartheta} \right|_{\hat{\vartheta}} &= \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \hat{\vartheta} = \begin{pmatrix} \hat{\mu} = n^{-1} \sum_{i=1}^n x_i = \bar{x} \\ \hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2 \end{pmatrix} \end{aligned}$$

- Le programme de maximisation a donc une solution

- Les réalisations du ML sont $\hat{\mu} = n^{-1} \sum_{i=1}^n x_i = \bar{x}$ et

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

i.e. variance empirique non-corrigée

- Les estimateurs du ML sont $\hat{\mu} = n^{-1} \sum_{i=1}^n X_i = \bar{X}$ et

$$\hat{\sigma}^2 = n^{-1} \sum_{i=1}^n (X_i - \bar{X})^2$$

Condition suffisante du MLE gaussien

- La solution est-elle bien un maximum ?

$$\frac{\partial^2 \ell_n(\vartheta; x)}{\partial \vartheta \partial \vartheta'} = \begin{pmatrix} \frac{\partial^2 \ell_n(\vartheta; x)}{\partial \mu^2} & \frac{\partial^2 \ell_n(\vartheta; x)}{\partial \mu \partial \sigma^2} \\ \frac{\partial^2 \ell_n(\vartheta; x)}{\partial \mu \partial \sigma^2} & \frac{\partial^2 \ell_n(\vartheta; x)}{\partial \sigma^4} \end{pmatrix}$$

- On obtient alors

$$\left. \frac{\partial^2 \ell_n(\vartheta; x)}{\partial \vartheta \partial \vartheta'} \right|_{\hat{\vartheta}} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & -\frac{1}{\hat{\sigma}^4} \sum_{i=1}^n (x_i - \hat{\mu}) \\ -\frac{1}{\hat{\sigma}^4} \sum_{i=1}^n (x_i - \hat{\mu}) & \frac{n}{2\hat{\sigma}^4} - \frac{1}{\hat{\sigma}^6} \sum_{i=1}^n (x_i - \hat{\mu})^2 \end{pmatrix}$$

- D'après l'étude du gradient, on sait que $n \times \hat{\mu} = \sum_{i=1}^n x_i$ et donc

$$-\frac{1}{\hat{\sigma}^4} \sum_{i=1}^n (x_i - \hat{\mu}) = -\frac{1}{\hat{\sigma}^4} \sum_{i=1}^n x_i + \frac{1}{\hat{\sigma}^4} n \times \hat{\mu} = \frac{1}{\hat{\sigma}^4} \sum_{i=1}^n x_i - \frac{1}{\hat{\sigma}^4} \sum_{i=1}^n x_i = 0$$

- De plus, $n \times \hat{\sigma}^2 = \sum_{i=1}^n (x_i - \hat{\mu})^2$, ce qui donne

$$\left. \frac{\partial^2 \ell_n(\vartheta; x)}{\partial \vartheta \partial \vartheta'} \right|_{\hat{\vartheta}} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & \frac{n}{2\hat{\sigma}^4} - \frac{n\hat{\sigma}^2}{\hat{\sigma}^6} \end{pmatrix} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix}$$

Condition suffisante du MLE gaussien

- Pour conclure, il faut montrer que la hessienne est définie négative

$$\frac{\partial^2 \ell_n(\vartheta; x)}{\partial \vartheta \partial \vartheta'} \bigg|_{\hat{\vartheta}} = \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix}$$

- Pour cela on s'intéresse aux mineurs principaux, Δ_1 et Δ_2 . Le premier mineur est

$$\Delta_1 = -\frac{n}{\hat{\sigma}^2} < 0$$

- Le second mineur est

$$\Delta_2 = \det \begin{pmatrix} -\frac{n}{\hat{\sigma}^2} & 0 \\ 0 & -\frac{n}{2\hat{\sigma}^4} \end{pmatrix} = -\frac{n}{\hat{\sigma}^2} \times -\frac{n}{2\hat{\sigma}^4} - 0 > 0$$

- Les mineurs principaux étant de signes opposés, la hessienne est bien définie négative et la solution du programme est bien un maximum

Le score

- Le score ressemble au gradient mais en diffère pour la raison suivante :
 - Le gradient est déterministe car basé sur les réalisations :

$$\frac{\partial \ell_n(\vartheta; x_1, \dots, x_n)}{\partial \vartheta}$$

- Le score est une version stochastique du gradient car basé sur les variables aléatoires :

$$S_n(\vartheta; X) = \frac{\partial \ell_n(\vartheta; X_1, \dots, X_n)}{\partial \vartheta}$$

- Le score étant une variable aléatoire, il convient de s'intéresser à ces moments et notamment son espérance
 - L'espérance nous intéresse afin de calculer la variance
 - La variance nous intéresse car elle permet de calculer la matrice d'information de Fisher

La hessienne stochastique

- De même que pour le gradient, on peut considérer une version stochastique de la hessienne

- La hessienne déterministe est basée sur les réalisations :

$$H_n(\vartheta, x) = \frac{\partial^2 \ell_n(\vartheta; x_1, \dots, x_n)}{\partial \vartheta \partial \vartheta'}$$

- La hessienne stochastique est basés sur les variables aléatoires :

$$H_n(\vartheta, X) = \frac{\partial^2 \ell_n(\vartheta; X_1, \dots, X_n)}{\partial \vartheta \partial \vartheta'}$$

- La hessienne stochastique étant une variable aléatoire elle a des moments :
- l'espérance de la hessienne nous permet de calculer la matrice d'information de Fisher

L'information de Fisher

- La matrice d'information de Fisher peut se calculer de plusieurs manières

Remark

La quantité d'information de Fisher associée à l'échantillon est une constante définie par la variance du score ou l'espérance de l'opposée de la hessienne stochastique :

$$I_n(\vartheta) = \mathbb{V}(S_n(\vartheta; X)) = \mathbb{E}(S_n^2(\vartheta; X)) - \mathbb{E}(S_n(\vartheta; X))^2$$

ou

$$I_n(\vartheta) = \mathbb{E}(-H_n(\vartheta, X))$$

L'information de Fisher et MLE Gaussien

- Repartons du MLE Gaussien et calculons l'information de Fisher :

$$S_n(\vartheta; X) = \begin{pmatrix} 0 \\ 0 \end{pmatrix} \Rightarrow \hat{\vartheta} = \begin{pmatrix} \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \mu) \\ -\frac{n}{2\sigma^2} + \frac{1}{2\sigma^4} \sum_{i=1}^n (X_i - \mu)^2 \end{pmatrix}$$

$$H_n(\vartheta; X) = \begin{pmatrix} -\frac{n}{\sigma^2} & -\frac{1}{\sigma^4} \sum_{i=1}^n (X_i - \mu) \\ -\frac{1}{\sigma^4} \sum_{i=1}^n (X_i - \mu) & \frac{n}{2\sigma^4} - \frac{1}{\sigma^6} \sum_{i=1}^n (X_i - \mu)^2 \end{pmatrix}$$

- Les deux méthodes peuvent être utilisées (ici l'espérance de la Hessienne)

$$\begin{aligned} I_n(\vartheta) &= \mathbb{E}(-H_n(\vartheta, X)) \\ &= \mathbb{E} \begin{pmatrix} \frac{n}{\sigma^2} & \frac{1}{\sigma^4} \sum_{i=1}^n (X_i - \mu) \\ \frac{1}{\sigma^4} \sum_{i=1}^n (X_i - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^n (X_i - \mu)^2 \end{pmatrix} \end{aligned}$$

L'information de Fisher et MLE Gaussien

- Les quantités déterministes n'étant pas affectées par l'espérance on a

$$I_n(\vartheta) = \begin{pmatrix} \frac{n}{\sigma^4} \sum_{i=1}^n \mathbb{E}(X_i - \mu) & -\frac{n}{2\sigma^4} + \frac{1}{\sigma^6} \sum_{i=1}^n \mathbb{E}((X_i - \mu)^2) \end{pmatrix}$$

- Or, $\mathbb{E}(X_i) = \mu$ donc $\mathbb{E}(X_i - \mu) = 0$
- De plus, par définition, $\mathbb{E}((X_i - \mu)^2) = \sigma^2$ ce qui nous donne

$$I_n(\vartheta) = \begin{pmatrix} \frac{n}{\sigma^2} & 0 \\ 0 & \frac{n}{2\sigma^4} \end{pmatrix}$$

- La borne informationnelle de Cramer-Rao définissant l'efficacité du MLE Gaussien est donc :

$$I_n^{-1}(\vartheta_0) = \begin{pmatrix} \frac{\sigma_0^2}{n} & 0 \\ 0 & \frac{2\sigma_0^4}{n} \end{pmatrix}$$

Propriétés du maximum de vraisemblance

■ Commençons par poser 3 hypothèses dites de régularité

Hypothèse 1 la fonction de densité $f_X(\vartheta; x_i)$ est trois fois différentiable par rapport à ϑ et ses dérivées sont continues et finies $\forall x, \vartheta$

Hypothèse 2 les espérances des dérivées première et seconde de $\ln f_X(\vartheta; X_i)$ par rapport à ϑ existent

Hypothèse 3 la vraie valeur de ϑ , i.e. ϑ_0 , appartient à un ensemble compact Θ

Note Par ensemble compact il faut comprendre un ensemble fermé et petit dont on ne peut pas s'échapper

Propriétés limites du maximum de vraisemblance et inférence

- Sous cet ensemble d'hypothèses il est possible de montrer

- que le MLE est convergent

$$\hat{\vartheta} \xrightarrow{p} \vartheta_0$$

- que le MLE est asymptotiquement efficace

$$\mathbb{V}(\hat{\vartheta}) = I_n^{-1}(\vartheta_0)$$

- que le MLE est asymptotiquement normalement distribué

$$\sqrt{n}(\hat{\vartheta} - \vartheta_0) \xrightarrow{d} \mathcal{N}(0, I_n^{-1}(\vartheta_0))$$

Maximum de vraisemblance conditionnelle

- Soit un modèle économétrique du type $Y_t = g(\vartheta; X_t) + \varepsilon_t$
- Une approche par MLE nécessite de considérer la distribution conditionnelle de Y sachant les réalisations de X

$$f_{Y|X}(y|x; \vartheta)$$

Remark (Vraisemblance conditionnelle)

Les fonctions de vraisemblance et log-vraisemblance conditionnelle d'un échantillon $\{y_t, x_t\}_{t=1}^n$ sont définies par

$$L_n(\vartheta; y|x) = \prod_{t=1}^n f_{Y|X}(y_t|x_t; \vartheta), \quad \text{et} \quad \ell_n(\vartheta; y|x) = \sum_{t=1}^n \ln f_{Y|X}(y_t|x_t; \vartheta)$$

MLE et modèle de régression linéaire

- Dans le cadre simple du modèle $Y_t = \beta X_t + \varepsilon_t \sim \text{i.i.d.}$
- En supposant la normalité des erreurs, i.e. $\varepsilon_t \sim \mathcal{N}(0, \sigma^2)$, si $X_i = x_i$, on obtient que $Y_i|x_i \sim \mathcal{N}(\beta x_i, \sigma^2)$
- On obtient alors la densité conditionnelle de Y_i suivante

$$f_{Y|X}(y_t|x_t; \vartheta) = \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{y_i - \beta x_i}{2\sigma}\right)^2, \quad \vartheta = (\beta, \sigma^2)'$$

- Les fonctions de ML et log-ML conditionnelles sont alors

$$L_n(\vartheta; y|x) = \prod_{i=1}^n \frac{1}{\sigma\sqrt{2\pi}} \exp\left(-\frac{y_i - \beta x_i}{2\sigma}\right)^2$$

et

$$\ell_n(\vartheta; y|x) = -\frac{n}{2} \ln(\sigma^2) - \frac{n}{2} \ln(2\pi) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \beta x_i)^2$$

Application du MLE : exercice

- Soit un échantillon $(X_1, \dots, X_n) \sim \text{i.i.d.}$ selon une distribution exponentielle de paramètre ϑ^{-1}

- La fonction de densité d'une loi exponentielle est $\vartheta^{-1} \exp(-\vartheta^{-1}X)$

- La log-vraisemblance de l'échantillon (x_1, \dots, x_n) est alors

$$\ell_n(\vartheta; x) = -n \ln(\vartheta) - \vartheta^{-1} \sum_{t=1}^n x_t$$

- L'estimateur du log-ML est alors

$$\frac{\partial \ell_n(\vartheta; x)}{\partial \vartheta} = -\frac{n}{\vartheta} + \frac{1}{\vartheta^2} \sum_{t=1}^n X_t = 0 \Rightarrow \hat{\vartheta} = n^{-1} \sum_{t=1}^n X_t$$

- Sachant que la loi exponentielle de paramètre λ a pour espérance λ^{-1} et pour variance λ^{-2} , $\mathbb{E}(X_t)$ et $\mathbb{V}(X_t)$ sont données par

$$\mathbb{E}(X_t) = \vartheta_0, \quad \mathbb{V}(X_t) = \vartheta_0^2$$

Application du MLE : solutions

- Calculez $\mathbb{E}(\hat{\vartheta})$

$$\mathbb{E}(\hat{\vartheta}) = \mathbb{E}\left(n^{-1} \sum_{t=1}^n X_t\right) = n^{-1} \sum_{t=1}^n \mathbb{E}(X_t) = \frac{n \times \vartheta_0}{n} = \vartheta_0$$

- Calculez $\mathbb{V}(\hat{\vartheta})$

$$\mathbb{V}(\hat{\vartheta}) = \mathbb{V}\left(n^{-1} \sum_{t=1}^n X_t\right) = n^{-2} \sum_{t=1}^n \mathbb{V}(X_t) = \frac{n \times \vartheta_0^2}{n^2} = \frac{\vartheta_0^2}{n}$$

- Que pouvez-vous conclure ?

- L'estimateur est sans biais et asymptotiquement convergent car

$$\lim_{n \rightarrow \infty} \mathbb{V}(\hat{\vartheta}) = 0$$

et donc

$$\hat{\vartheta} \xrightarrow{P} \vartheta_0$$