Estimation des modèles ARMA

Plan

Plan

Le problème de l'estimation

- Soit un modèle ARMA(p, q) :

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

  avec ε_t un bruit blanc.

- Les chapitres précédents ont montré comment calculer les moments du processus (autocovariances, autocorrélations, prévisions linéaires) en fonction des paramètres.

- **Problème :** Comment estimer les paramètres $(c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)$ à partir d'un échantillon d'observations $(y_1, y_2, \ldots, y_T)$?

Principe du maximum de vraisemblance

- On note $\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$ le vecteur des paramètres.
- On suppose que $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$.
- On calcule la densité jointe de l'échantillon observé :

$$f_{Y_T, Y_{T-1}, \ldots, Y_1}(y_T, y_{T-1}, \ldots, y_1; \boldsymbol{\theta})$$

  vue comme une fonction de $\boldsymbol{\theta}$ pour les données observées.

- L'**estimateur du maximum de vraisemblance** (EMV) est la valeur $\hat{\boldsymbol{\theta}}$ qui maximise cette fonction, c'est-à-dire la valeur des paramètres pour laquelle l'échantillon observé est le plus probable.

Hypothèse de normalité

- L'hypothèse de normalité sur ε_t est forte, mais l'estimation résultante reste pertinente même si elle est violée.
- Si le vrai processus est non gaussien, les estimateurs obtenus en maximisant la vraisemblance gaussienne restent **convergents**. On parle alors d'estimateur de **quasi-maximum de vraisemblance**.
- En revanche, les erreurs standard calculées sous l'hypothèse de normalité peuvent ne pas être correctes si les données sont non gaussiennes.

Décomposition en erreurs de prévision

- La densité jointe peut être factorisée en utilisant la règle de Bayes :

$$f_{Y_T,\ldots,Y_1}(y_T,\ldots,y_1;\boldsymbol{\theta}) = f_{Y_1}(y_1;\boldsymbol{\theta}) \cdot \prod_{t=2}^{T} f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\boldsymbol{\theta})$$

- La **log-vraisemblance** est donc :

$$\mathcal{L}(\boldsymbol{\theta}) = \log f_{Y_1}(y_1;\boldsymbol{\theta}) + \sum_{t=2}^{T} \log f_{Y_t|Y_{t-1}}(y_t|y_{t-1};\boldsymbol{\theta})$$

- Cette décomposition est connue sous le nom de **décomposition en erreurs de prévision** (*prediction-error decomposition*).

Plan

Le modèle AR(1) gaussien

- On considère le processus AR(1) gaussien :

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t$$

avec $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$ et $|\phi| < 1$.

- Le vecteur de paramètres est $\boldsymbol{\theta} = (c, \phi, \sigma^2)'$.
- L'espérance du processus stationnaire est $\mu = c/(1-\phi)$.
- La variance du processus stationnaire est $\sigma^2/(1-\phi^2)$.

Densité de la première observation

- Puisque ε_t est gaussien, Y_1 est également gaussien avec :

$$\mathbb{E}[Y_1] = \mu = \frac{c}{1-\phi}, \quad \mathbb{V}[Y_1] = \frac{\sigma^2}{1-\phi^2}$$

- La densité de Y_1 est donc :

$$f_{Y_1}(y_1; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2/(1-\phi^2)}} \exp\left[\frac{-\{y_1 - c/(1-\phi)\}^2}{2\sigma^2/(1-\phi^2)}\right]$$

Densités conditionnelles

- Conditionnellement à $Y_{t-1} = y_{t-1}$, on a :

$$Y_t = c + \phi Y_{t-1} + \varepsilon_t \sim N(c + \phi y_{t-1}, \sigma^2)$$

- La densité conditionnelle est :

$$f_{Y_t|Y_{t-1}}(y_t|y_{t-1}; \boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\right]$$

- Cette expression est valable pour $t = 2, 3, \ldots, T$.

Log-vraisemblance exacte

- La log-vraisemblance exacte de l'AR(1) gaussien est :

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = &-\frac{1}{2}\log(2\pi) - \frac{1}{2}\log\left(\frac{\sigma^2}{1-\phi^2}\right) - \frac{\{y_1 - c/(1-\phi)\}^2}{2\sigma^2/(1-\phi^2)} \\ &- \frac{T-1}{2}\log(2\pi) - \frac{T-1}{2}\log(\sigma^2) \\ &- \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\end{aligned}$$

- On regroupe les termes pour obtenir :

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = &-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log(1-\phi^2) \\ &- \frac{(1-\phi^2)}{2\sigma^2}\left(y_1 - \frac{c}{1-\phi}\right)^2 - \sum_{t=2}^{T} \frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}\end{aligned}$$

Log-vraisemblance conditionnelle

- Une alternative consiste à conditionner sur la première observation y_1 et à maximiser :

$$\mathcal{L}_c(\boldsymbol{\theta}) = -\frac{T-1}{2}\log(2\pi) - \frac{T-1}{2}\log(\sigma^2) - \sum_{t=2}^{T}\frac{(y_t - c - \phi y_{t-1})^2}{2\sigma^2}$$

- La maximisation par rapport à c et ϕ revient à minimiser :

$$\sum_{t=2}^{T}(y_t - c - \phi y_{t-1})^2$$

- C'est un problème de **moindres carrés ordinaires** (MCO) : régression de y_t sur une constante et y_{t-1}.

Estimateurs conditionnels du maximum de vraisemblance

- Les estimateurs conditionnels de c et ϕ sont donnés par la formule des MCO :

$$\begin{bmatrix} \hat{c} \\ \hat{\phi} \end{bmatrix} = \begin{bmatrix} T-1 & \sum y_{t-1} \\ \sum y_{t-1} & \sum y_{t-1}^2 \end{bmatrix}^{-1} \begin{bmatrix} \sum y_t \\ \sum y_{t-1} y_t \end{bmatrix}$$

  où les sommes portent sur $t = 2, 3, \ldots, T$.

- L'estimateur conditionnel de σ^2 est la variance résiduelle de la régression :

$$\hat{\sigma}^2 = \frac{1}{T-1} \sum_{t=2}^{T} (y_t - \hat{c} - \hat{\phi} y_{t-1})^2$$

- Pour T grand, les estimateurs exact et conditionnel convergent vers la même distribution asymptotique (si $|\phi| < 1$).

Vraisemblance exacte et conditionnelle : comparaison

- La vraisemblance exacte nécessite le terme supplémentaire $\frac{1}{2}\log(1-\phi^2) - \frac{(1-\phi^2)}{2\sigma^2}(y_1-\mu)^2$ lié à la première observation.
- La vraisemblance conditionnelle ignore ce terme et traite y_1 comme déterministe.
- Si T est grand, la contribution de la première observation est négligeable.
- Si $|\phi| < 1$, les deux approches donnent des estimateurs **convergents**. Si $|\phi| > 1$, la vraisemblance conditionnelle ne fournit pas d'estimateurs convergents car la densité de Y_1 n'est pas correctement spécifiée.

Expression vectorielle de la vraisemblance (1/6)

- On peut dériver la vraisemblance d'une manière alternative en considérant le vecteur des T observations :

$$\mathbf{y} = (y_1, y_2, \ldots, y_T)'$$

- Ce vecteur peut être vu comme une unique réalisation d'un vecteur gaussien de dimension T :

$$\mathbf{y} \sim N(\boldsymbol{\mu}, \boldsymbol{\Omega})$$

- Le vecteur d'espérances est $\boldsymbol{\mu} = (\mu, \mu, \ldots, \mu)'$ avec $\mu = c/(1-\phi)$.

- La matrice de variance-covariance $\boldsymbol{\Omega}$ est une matrice $(T \times T)$ dont l'élément (i, j) est l'autocovariance $\gamma(|i-j|)$.

Expression vectorielle de la vraisemblance (2/6)

- Pour l'AR(1), l'autocovariance est $\gamma(h) = \sigma^2\phi^{|h|}/(1-\phi^2)$, donc :

$$\boldsymbol{\Omega} = \sigma^2 \mathbf{V}$$

où

$$\mathbf{V} = \frac{1}{1-\phi^2}\begin{bmatrix} 1 & \phi & \phi^2 & \cdots & \phi^{T-1} \\ \phi & 1 & \phi & \cdots & \phi^{T-2} \\ \phi^2 & \phi & 1 & \cdots & \phi^{T-3} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \phi^{T-1} & \phi^{T-2} & \phi^{T-3} & \cdots & 1 \end{bmatrix}$$

- $\mathbf{V}$ est une matrice de Toeplitz symétrique définie positive.

Expression vectorielle de la vraisemblance (3/6)

- La densité du vecteur gaussien $\mathbf{y}$ s'écrit :

$$f_{\mathbf{Y}}(\mathbf{y};\boldsymbol{\theta}) = (2\pi)^{-T/2}|\boldsymbol{\Omega}|^{-1/2}\exp\left[-\tfrac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})\right]$$

- La log-vraisemblance est donc :

$$\boxed{\mathcal{L}(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) + \frac{1}{2}\log|\boldsymbol{\Omega}^{-1}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})}$$

- Cette expression fait intervenir l'inverse et le déterminant de la matrice $\boldsymbol{\Omega}$ de taille $(T \times T)$. En apparence, le calcul est coûteux, mais la structure de $\boldsymbol{\Omega}$ permet de le simplifier.

Expression vectorielle de la vraisemblance (4/6)

- On introduit la matrice triangulaire inférieure $\mathbf{L}$ de taille $(T \times T)$:

$$\mathbf{L} = \begin{bmatrix} \sqrt{1-\phi^2} & 0 & 0 & \cdots & 0 & 0 \\ -\phi & 1 & 0 & \cdots & 0 & 0 \\ 0 & -\phi & 1 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\phi & 1 \end{bmatrix}$$

- On peut montrer que :

$$\mathbf{L}'\mathbf{L} = \mathbf{V}^{-1}$$

et donc $\mathbf{\Omega}^{-1} = \sigma^{-2}\mathbf{V}^{-1} = \sigma^{-2}\mathbf{L}'\mathbf{L}$.

Expression vectorielle de la vraisemblance (5/6)

- Puisque $\mathbf{L}$ est triangulaire, son déterminant est le produit des éléments diagonaux :

$$|\mathbf{L}| = \sqrt{1-\phi^2} \cdot \underbrace{1 \cdot 1 \cdots 1}_{T-1} = \sqrt{1-\phi^2}$$

- Donc $|\mathbf{L}'\mathbf{L}| = |\mathbf{V}^{-1}| = 1-\phi^2$, et :

$$\frac{1}{2}\log|\mathbf{\Omega}^{-1}| = \frac{1}{2}\log(\sigma^{-2T} \cdot |\mathbf{V}^{-1}|) = -\frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log(1-\phi^2)$$

- On définit le vecteur transformé $\tilde{\mathbf{y}} = \mathbf{L}(\mathbf{y}-\boldsymbol{\mu})$, de sorte que :

$$(\mathbf{y}-\boldsymbol{\mu})'\mathbf{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu}) = \frac{1}{\sigma^2}\tilde{\mathbf{y}}'\tilde{\mathbf{y}}$$

Expression vectorielle de la vraisemblance (6/6)

- Les composantes de $\tilde{\mathbf{y}} = \mathbf{L}(\mathbf{y} - \boldsymbol{\mu})$ sont :

$$\tilde{y}_1 = \sqrt{1-\phi^2}\,(y_1 - \mu), \quad \tilde{y}_t = (y_t - \mu) - \phi(y_{t-1} - \mu) \text{ pour } t \geq 2$$

- En substituant $\mu = c/(1-\phi)$:

$$\tilde{y}_t = y_t - c - \phi y_{t-1} \quad \text{pour } t \geq 2$$

  Ce sont les **erreurs de prévision** !

- La log-vraisemblance s'écrit alors :

$$\boxed{\mathcal{L}(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log(1-\phi^2) - \frac{1}{2\sigma^2}\sum_{t=1}^{T}\tilde{y}_t^2}$$

- On retrouve exactement l'expression de la log-vraisemblance exacte obtenue par la

Plan

Le modèle AR(p) gaussien

- On considère le processus AR(p) gaussien :

$$Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \cdots + \phi_p Y_{t-p} + \varepsilon_t$$

avec $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$.

- Le vecteur de paramètres est $\boldsymbol{\theta} = (c, \phi_1, \phi_2, \ldots, \phi_p, \sigma^2)'$.

- L'espérance du processus stationnaire est :

$$\mu = \frac{c}{1 - \phi_1 - \phi_2 - \cdots - \phi_p}$$

Construction de la vraisemblance

- Les p premières observations $(y_1, \ldots, y_p)$ suivent une loi gaussienne multivariée $N(\boldsymbol{\mu}_p, \sigma^2 \mathbf{V}_p)$.

- Pour $t > p$, conditionnellement aux p observations précédentes :

$$Y_t | Y_{t-1}, \ldots, Y_{t-p} \sim N(c + \phi_1 y_{t-1} + \cdots + \phi_p y_{t-p},\ \sigma^2)$$

- La log-vraisemblance exacte est :

$$\begin{aligned}\mathcal{L}(\boldsymbol{\theta}) = &-\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) + \frac{1}{2}\log|\mathbf{V}_p^{-1}| \\ &- \frac{1}{2\sigma^2}(\mathbf{y}_p - \boldsymbol{\mu}_p)'\mathbf{V}_p^{-1}(\mathbf{y}_p - \boldsymbol{\mu}_p) \\ &- \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}\end{aligned}$$

Estimateurs conditionnels pour l'AR(p)

- La log-vraisemblance conditionnelle (sur les p premières observations) est :

$$\mathcal{L}_c(\boldsymbol{\theta}) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T} \frac{(y_t - c - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p})^2}{2\sigma^2}$$

- La maximisation revient à minimiser :

$$\sum_{t=p+1}^{T} (y_t - c - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p})^2$$

- C'est la somme des carrés des résidus d'une régression MCO de y_t sur une constante et ses p valeurs retardées.

Estimateur de la variance des innovations

- L'estimateur conditionnel de σ^2 est le résidu quadratique moyen :

$$\hat{\sigma}^2 = \frac{1}{T-p} \sum_{t=p+1}^{T} (y_t - \hat{c} - \hat{\phi}_1 y_{t-1} - \hat{\phi}_2 y_{t-2} - \cdots - \hat{\phi}_p y_{t-p})^2$$

- Les estimateurs exact et conditionnel ont la même distribution asymptotique.

- **Résultat important :** Pour un processus AR(p), les estimateurs conditionnels du maximum de vraisemblance sont identiques aux estimateurs des MCO. L'estimation d'un AR(p) est donc particulièrement simple.

Plan

Processus non gaussiens

- Que se passe-t-il si le processus n'est pas gaussien ?
- La régression MCO de y_t sur une constante et ses p retards fournit une estimation convergente de la **projection linéaire** :

$$\hat{\mathbb{E}}(Y_t|Y_{t-1}, Y_{t-2}, \ldots, Y_{t-p})$$

  pourvu que le processus soit ergodique pour les moments d'ordre 2.
- Cette régression MCO maximise aussi la **log-vraisemblance gaussienne conditionnelle**. Même si le processus n'est pas gaussien, maximiser cette fonction fournit des estimateurs convergents.

Estimateur de quasi-maximum de vraisemblance

- Un estimateur obtenu en maximisant une vraisemblance mal spécifiée (par exemple, gaussienne alors que les données ne le sont pas) est appelé estimateur de **quasi-maximum de vraisemblance** (QMLE).

- **Propriété :** Le QMLE fournit des estimateurs **convergents** des paramètres $(\phi_1, \ldots, \phi_p)$.

- Cependant, les erreurs standard calculées sous l'hypothèse gaussienne peuvent ne pas être correctes. Il faut utiliser un estimateur robuste de la matrice de variance-covariance.

- En pratique, si les données sont non gaussiennes, une transformation préalable (par exemple, le logarithme) peut rapprocher la distribution de la normalité.

Transformations Box-Cox

- Pour une variable positive Y_t, Box et Cox (1964) proposent la famille de transformations :

$$Y_t^{(\lambda)} = \begin{cases} \dfrac{Y_t^{\lambda} - 1}{\lambda} & \text{si } \lambda \neq 0 \\ \log Y_t & \text{si } \lambda = 0 \end{cases}$$

- On choisit λ de sorte que $Y_t^{(\lambda)}$ soit bien approximé par un processus ARMA gaussien.

- En pratique, pour les séries économiques qui croissent dans le temps (PIB, prix), on utilise souvent :

$$y_t = \log X_t - \log X_{t-1}$$

c'est-à-dire le taux de croissance en log.

Plan

Le modèle MA(1) gaussien

- On considère le processus MA(1) gaussien :

$$Y_t = \mu + \varepsilon_t + \theta\varepsilon_{t-1}$$

avec $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$.

- Le vecteur de paramètres est $\boldsymbol{\theta} = (\mu, \theta, \sigma^2)'$.

- Contrairement à l'AR, la vraisemblance du MA ne se réduit **pas** à un problème de moindres carrés.

- Deux approches : la vraisemblance **conditionnelle** (plus simple) et la vraisemblance **exacte** (plus précise en petit échantillon).

Vraisemblance conditionnelle du MA(1) (1/2)

- On conditionne sur la valeur initiale $\varepsilon_0 = 0$. Sachant ε_{t-1}, la densité de Y_t est :

$$f_{Y_t|\varepsilon_{t-1}}(y_t|\varepsilon_{t-1};\boldsymbol{\theta}) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[\frac{-(y_t - \mu - \theta\varepsilon_{t-1})^2}{2\sigma^2}\right]$$

- Sachant $\varepsilon_0 = 0$, on déduit ε_1 de l'observation y_1 :

$$\varepsilon_1 = y_1 - \mu$$

- Plus généralement, les innovations sont calculées récursivement :

$$\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}$$

pour $t = 1, 2, \ldots, T$, en partant de $\varepsilon_0 = 0$.

Vraisemblance conditionnelle du MA(1) (2/2)

- La log-vraisemblance conditionnelle est :

$$\mathcal{L}_c(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T}\frac{\varepsilon_t^2}{2\sigma^2}$$

  où $\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}$ avec $\varepsilon_0 = 0$.

- La log-vraisemblance est une fonction **non linéaire** de μ et θ : pas de solution analytique explicite.

- La maximisation requiert une **optimisation numérique**.

- **Condition d'inversibilité :** L'approximation conditionnelle est valide si $|\theta| < 1$. Si l'optimisation conduit à $|\hat{\theta}| > 1$, il faut recommencer avec $\hat{\theta}^{-1}$.

Effet de la condition initiale

- En développant la récurrence $\varepsilon_t = y_t - \mu - \theta\varepsilon_{t-1}$:

$$\varepsilon_t = (y_t - \mu) - \theta(y_{t-1} - \mu) + \theta^2(y_{t-2} - \mu) - \cdots \\ + (-1)^{t-1}\theta^{t-1}(y_1 - \mu) + (-1)^t\theta^t\varepsilon_0$$

- Si $|\theta| < 1$, l'effet de $\varepsilon_0 = 0$ décroît géométriquement : $(-1)^t\theta^t\varepsilon_0 \to 0$.

- Pour T suffisamment grand, l'approximation $\varepsilon_0 = 0$ est inoffensive.

- Si $|\theta|$ est proche de 1, les premières innovations ε_t sont mal estimées, ce qui peut affecter l'estimation en petit échantillon.

Vraisemblance exacte du MA(1) (1/6)

- La vraisemblance exacte traite les T observations comme un vecteur gaussien $\mathbf{y} = (y_1, \ldots, y_T)'$ de loi $N(\boldsymbol{\mu}, \boldsymbol{\Omega})$.

- La matrice de variance-covariance $\boldsymbol{\Omega}$ a une structure **tridiagonale** :

$$\boldsymbol{\Omega} = \sigma^2 \begin{bmatrix} 1+\theta^2 & \theta & 0 & \cdots & 0 \\ \theta & 1+\theta^2 & \theta & \cdots & 0 \\ 0 & \theta & 1+\theta^2 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1+\theta^2 \end{bmatrix}$$

- La log-vraisemblance exacte est :

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\log|\boldsymbol{\Omega}| - \frac{1}{2}(\mathbf{y}-\boldsymbol{\mu})'\boldsymbol{\Omega}^{-1}(\mathbf{y}-\boldsymbol{\mu})$$

Vraisemblance exacte du MA(1) (2/6)

- On cherche la **factorisation triangulaire** $\mathbf{\Omega} = \mathbf{A}\mathbf{D}\mathbf{A}'$ où :
  - $\mathbf{A}$ est triangulaire inférieure avec des 1 sur la diagonale,
  - $\mathbf{D}$ est diagonale avec des éléments strictement positifs.
- Puisque $\mathbf{\Omega}$ est tridiagonale, $\mathbf{A}$ est **bidiagonale** : seuls les éléments diagonaux et sous-diagonaux sont non nuls.

- On note $S_t = 1 + \theta^2 + \theta^4 + \cdots + \theta^{2(t-1)}$ la somme géométrique. On a $S_1 = 1$ et :

$$S_t = \frac{1 - \theta^{2t}}{1 - \theta^2} \quad \text{si } \theta^2 \neq 1$$

Vraisemblance exacte du MA(1) (3/6)

- On obtient $\mathbf{A}$ et $\mathbf{D}$ par **élimination de Gauss** sur $\mathbf{\Omega}/\sigma^2$.

- **Étape 1 :** Le pivot est $d_1 = (1+\theta^2) = S_2/S_1$. On élimine le terme sous-diagonal θ en soustrayant $\frac{\theta}{1+\theta^2}$ fois la première ligne. Cela donne $a_{21} = \frac{\theta}{1+\theta^2} = \frac{\theta S_1}{S_2}$.

- **Étape 2 :** Le nouveau pivot est :

$$d_2 = (1+\theta^2) - \frac{\theta^2}{1+\theta^2} = \frac{(1+\theta^2)^2 - \theta^2}{1+\theta^2} = \frac{1+\theta^2+\theta^4}{1+\theta^2} = \frac{S_3}{S_2}$$

  Le multiplicateur est $a_{32} = \frac{\theta}{d_2} = \frac{\theta S_2}{S_3}$.

Vraisemblance exacte du MA(1) (4/6)

- En poursuivant, on montre par récurrence que pour $t = 1, \ldots, T$:

$$\boxed{d_t = \frac{S_{t+1}}{S_t}} \quad \text{et} \quad \boxed{a_{t+1,t} = \frac{\theta\, S_t}{S_{t+1}}}$$

- Explicitement, les matrices sont (en factorisant σ^2 dans $\mathbf{D}$) :

$$\mathbf{A} = \begin{bmatrix} 1 & 0 & 0 & \cdots & 0 \\ \frac{\theta S_1}{S_2} & 1 & 0 & \cdots & 0 \\ 0 & \frac{\theta S_2}{S_3} & 1 & \cdots & 0 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & 1 \end{bmatrix}, \quad \mathbf{D} = \sigma^2 \begin{bmatrix} \frac{S_2}{S_1} & & & & \\ & \frac{S_3}{S_2} & & & \\ & & \ddots & & \\ & & & \frac{S_{T+1}}{S_T} & \end{bmatrix}$$

Vraisemblance exacte du MA(1) (5/6)

- Le vecteur transformé $\tilde{\mathbf{y}} = \mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ a pour composantes :

$$\tilde{y}_1 = y_1 - \mu, \qquad \tilde{y}_t = (y_t - \mu) - \frac{\theta\, S_{t-1}}{S_t}\, \tilde{y}_{t-1} \quad \text{pour } t \geq 2$$

- Puisque $|\mathbf{A}| = 1$ (triangulaire avec des 1 sur la diagonale) et $\boldsymbol{\Omega} = \mathbf{A}\mathbf{D}\mathbf{A}'$:

$$|\boldsymbol{\Omega}| = |\mathbf{D}| = \sigma^{2T} \prod_{t=1}^{T} \frac{S_{t+1}}{S_t} = \sigma^{2T}\, \frac{S_{T+1}}{S_1} = \sigma^{2T}\, S_{T+1}$$

  car le produit est télescopique.

- De plus :

$$(\mathbf{y} - \boldsymbol{\mu})' \boldsymbol{\Omega}^{-1} (\mathbf{y} - \boldsymbol{\mu}) = \tilde{\mathbf{y}}' \mathbf{D}^{-1} \tilde{\mathbf{y}} = \sum_{t=1}^{T} \frac{\tilde{y}_t^2}{d_t}$$

Vraisemblance exacte du MA(1) (6/6)

- La log-vraisemblance exacte s'écrit alors :

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_t) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_t}$$

avec $d_t = \sigma^2 S_{t+1}/S_t$ et $S_t = 1 + \theta^2 + \cdots + \theta^{2(t-1)}$.

- Cette expression est valide pour **toute** valeur de θ (pas seulement $|\theta| < 1$).

- On montre que si $\theta = \hat{\theta}$ maximise cette expression, alors $\hat{\theta}^{-1}$ donne la même valeur. On retient la solution inversible : $|\hat{\theta}| < 1$.

- Pour la vraisemblance conditionnelle, cette propriété n'est pas garantie.

Plan

Le modèle MA(q) gaussien

- On considère le processus MA(q) gaussien :

$$Y_t = \mu + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \theta_2 \varepsilon_{t-2} + \cdots + \theta_q \varepsilon_{t-q}$$

  avec $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$.

- Le vecteur de paramètres est $\boldsymbol{\theta} = (\mu, \theta_1, \ldots, \theta_q, \sigma^2)'$.

- La matrice de variance-covariance $\boldsymbol{\Omega}$ est une matrice bande de largeur q : les autocovariances γ_k sont nulles pour $k > q$.

Vraisemblance conditionnelle du MA(q)

- On pose $\varepsilon_0 = \varepsilon_{-1} = \cdots = \varepsilon_{-q+1} = 0$.

- Les innovations sont calculées par récurrence :

$$\varepsilon_t = y_t - \mu - \theta_1 \varepsilon_{t-1} - \theta_2 \varepsilon_{t-2} - \cdots - \theta_q \varepsilon_{t-q}$$

  pour $t = 1, 2, \ldots, T$.

- La log-vraisemblance conditionnelle est :

$$\mathcal{L}_c(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}$$

- Cette approximation est valide si toutes les racines de $1 + \theta_1 z + \cdots + \theta_q z^q = 0$ sont de module > 1 (inversibilité).

Vraisemblance exacte du MA(q)

- La structure bande de $\boldsymbol{\Omega}$ permet d'utiliser la factorisation triangulaire $\boldsymbol{\Omega} = \mathbf{A}\mathbf{D}\mathbf{A}'$.
- $\mathbf{A}$ est une matrice triangulaire inférieure bande : $a_{ij} = 0$ pour $i > q + j$.
- Les éléments de $\tilde{\mathbf{y}} = \mathbf{A}^{-1}(\mathbf{y} - \boldsymbol{\mu})$ se calculent récursivement par résolution d'un système triangulaire.
- La log-vraisemblance exacte est :

$$\mathcal{L}(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{1}{2}\sum_{t=1}^{T}\log(d_{tt}) - \frac{1}{2}\sum_{t=1}^{T}\frac{\tilde{y}_t^2}{d_{tt}}$$

- Contrairement à la vraisemblance conditionnelle, l'expression exacte est valide pour toute valeur de $(\theta_1, \ldots, \theta_q)$.

Plan

Le modèle ARMA(p, q) gaussien

- Le modèle ARMA(p, q) gaussien s'écrit :

$$Y_t = c + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \varepsilon_t + \theta_1 \varepsilon_{t-1} + \cdots + \theta_q \varepsilon_{t-q}$$

  avec $\varepsilon_t \sim$ i.i.d. $N(0, \sigma^2)$.

- Le vecteur de paramètres est $\boldsymbol{\theta} = (c, \phi_1, \ldots, \phi_p, \theta_1, \ldots, \theta_q, \sigma^2)'$.

- Le nombre de paramètres est $p + q + 2$.

Vraisemblance conditionnelle du ARMA(p, q)

- On fixe les conditions initiales :
  - $\mathbf{y}_0 = (y_0, y_{-1}, \ldots, y_{-p+1})'$ aux valeurs observées ou à l'espérance,
  - $\boldsymbol{\varepsilon}_0 = (\varepsilon_0, \varepsilon_{-1}, \ldots, \varepsilon_{-q+1})' = \mathbf{0}$.
- Les innovations sont calculées récursivement :

$$\varepsilon_t = y_t - c - \phi_1 y_{t-1} - \cdots - \phi_p y_{t-p} - \theta_1 \varepsilon_{t-1} - \cdots - \theta_q \varepsilon_{t-q}$$

  pour $t = 1, 2, \ldots, T$.

- La log-vraisemblance conditionnelle est :

$$\mathcal{L}_c(\boldsymbol{\theta}) = -\frac{T}{2}\log(2\pi) - \frac{T}{2}\log(\sigma^2) - \sum_{t=1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}$$

Conditions initiales alternatives

- **Approche Box-Jenkins :** On fixe $y_s = c/(1 - \phi_1 - \cdots - \phi_p)$ pour $s \leq 0$ et $\varepsilon_s = 0$ pour $s \leq 0$.

- **Approche alternative :** On fixe les ε à zéro et les y à leurs valeurs observées. On commence l'itération à $t = p + 1$:

$$\varepsilon_p = \varepsilon_{p-1} = \cdots = \varepsilon_{p-q+1} = 0$$

La log-vraisemblance conditionnelle est alors :

$$\mathcal{L}_c(\boldsymbol{\theta}) = -\frac{T-p}{2}\log(2\pi) - \frac{T-p}{2}\log(\sigma^2) - \sum_{t=p+1}^{T} \frac{\varepsilon_t^2}{2\sigma^2}$$

- Ces approximations sont valides si toutes les racines de $\phi(z) = 0$ et $\theta(z) = 0$ sont de module > 1.

Vraisemblance exacte

- L'approche la plus rigoureuse utilise le **filtre de Kalman** pour calculer la vraisemblance exacte.
- Le filtre de Kalman fournit, de manière récursive, les prévisions optimales et les erreurs de prévision, ce qui permet de construire la vraisemblance par décomposition en erreurs de prévision.
- Alternativement, on peut utiliser la **factorisation triangulaire** de la matrice de variance-covariance $\mathbf{\Omega}$, comme dans le cas MA.
- Les deux approches donnent le même résultat mais diffèrent en termes d'implémentation informatique.

Plan

Le problème d'optimisation

- La log-vraisemblance est une fonction non linéaire de $\boldsymbol{\theta}$. Il faut trouver :

$$\hat{\boldsymbol{\theta}} = \arg\max_{\boldsymbol{\theta}} \mathcal{L}(\boldsymbol{\theta})$$

- **Exception :** pour un processus AR pur, les estimateurs ont une solution analytique (MCO).

- Dans le cas général (MA ou ARMA), il faut recourir à des méthodes d'**optimisation numérique**.

- Idée : calculer numériquement $\mathcal{L}(\boldsymbol{\theta})$ pour différentes valeurs de $\boldsymbol{\theta}$ et chercher la valeur qui maximise cette fonction.

Recherche sur grille (*grid search*)

- La méthode la plus simple : évaluer $\mathcal{L}(\boldsymbol{\theta})$ sur une grille de valeurs de $\boldsymbol{\theta}$.
- **Exemple :** Pour un AR(1) avec $c = 0$ et $\sigma^2 = 1$, on évalue $\mathcal{L}(\phi)$ pour $\phi \in \{-0.9, -0.8, \ldots, 0.8, 0.9\}$.
- On raffine la grille autour du maximum.
- **Avantage :** Simple et permet de visualiser la surface de vraisemblance.
- **Inconvénient :** Devient impraticable quand le nombre de paramètres augmente (malédiction de la dimension).

Gradient et méthode de plus forte pente

- On définit le **gradient** de la log-vraisemblance :

$$\mathbf{g}(\boldsymbol{\theta}) = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

  C'est un vecteur qui pointe dans la direction d'augmentation la plus rapide de $\mathcal{L}$.

- **Méthode de plus forte pente** (*steepest ascent*) :

$$\boldsymbol{\theta}^{(n+1)} = \boldsymbol{\theta}^{(n)} + \alpha_n \mathbf{g}(\boldsymbol{\theta}^{(n)})$$

  où $\alpha_n > 0$ est le **pas** de l'algorithme.

- On itère jusqu'à convergence : $\|\boldsymbol{\theta}^{(n+1)} - \boldsymbol{\theta}^{(n)}\| < \epsilon$ ou $\|\mathbf{g}(\boldsymbol{\theta}^{(n)})\| < \epsilon$.

Calcul du gradient

- Le gradient peut être calculé de deux manières.

- **Analytiquement :** On différencie $\mathcal{L}(\boldsymbol{\theta})$ par rapport à chaque élément de $\boldsymbol{\theta}$. C'est possible pour les modèles AR et MA, mais les expressions deviennent complexes pour les modèles ARMA.

- **Numériquement :** On approxime les dérivées partielles par différences finies :

$$\frac{\partial \mathcal{L}}{\partial \theta_i} \approx \frac{\mathcal{L}(\boldsymbol{\theta} + h\mathbf{e}_i) - \mathcal{L}(\boldsymbol{\theta} - h\mathbf{e}_i)}{2h}$$

  où $\mathbf{e}_i$ est le i-ème vecteur de la base canonique et h est un petit incrément.

- Le gradient numérique est facile à programmer mais introduit une erreur d'approximation.

Maxima locaux et globaux

- Si la log-vraisemblance est **unimodale**, les méthodes itératives convergent vers le maximum global, quel que soit le point de départ.

- En général, $\mathcal{L}(\boldsymbol{\theta})$ peut avoir plusieurs **maxima locaux**. L'algorithme converge alors vers le maximum local le plus proche du point de départ.

- **Stratégies pratiques :**
  - Essayer plusieurs points de départ différents
  - Utiliser d'abord une recherche sur grille grossière, puis affiner avec une méthode de gradient
  - Comparer les valeurs de $\mathcal{L}(\hat{\boldsymbol{\theta}})$ obtenues à partir de différents points de départ

Plan

La fonction de score

- On appelle **fonction de score** le gradient de la log-vraisemblance :

$$\mathbf{s}(\boldsymbol{\theta}) = \frac{\partial \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}}$$

  C'est un vecteur de dimension $(p + q + 2) \times 1$.

- Sous des conditions de régularité (échange dérivée/intégrale), le score a une espérance nulle évaluée en la vraie valeur :

$$\mathbb{E}_{\boldsymbol{\theta}_0}[\mathbf{s}(\boldsymbol{\theta}_0)] = \mathbf{0}$$

- Intuition : à la vraie valeur des paramètres, la log-vraisemblance est en moyenne à son maximum.

Matrice d'information de Fisher

- La **matrice d'information de Fisher** est définie par :

$$\mathcal{I}(\boldsymbol{\theta}) = \mathbb{V}_{\boldsymbol{\theta}}[\mathbf{s}(\boldsymbol{\theta})] = \mathbb{E}_{\boldsymbol{\theta}}\left[\mathbf{s}(\boldsymbol{\theta})\mathbf{s}(\boldsymbol{\theta})'\right]$$

- Sous les conditions de régularité, on a l'**égalité de l'information** :

$$\mathcal{I}(\boldsymbol{\theta}) = -\mathbb{E}_{\boldsymbol{\theta}}\left[\frac{\partial^2 \mathcal{L}(\boldsymbol{\theta})}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\right] = -\mathbb{E}_{\boldsymbol{\theta}}[\mathbf{H}(\boldsymbol{\theta})]$$

  où $\mathbf{H}(\boldsymbol{\theta})$ est la matrice **hessienne** de la log-vraisemblance.

- La matrice d'information mesure la quantité d'information que l'échantillon contient sur les paramètres.

Borne de Cramér-Rao

- **Borne de Cramér-Rao :** Pour tout estimateur sans biais $\hat{\boldsymbol{\theta}}$ de $\boldsymbol{\theta}$:

$$\mathbb{V}(\hat{\boldsymbol{\theta}}) \geq \mathcal{I}(\boldsymbol{\theta})^{-1}$$

  au sens des matrices (la différence est semi-définie positive).

- L'estimateur du maximum de vraisemblance **atteint** cette borne asymptotiquement : c'est l'estimateur le plus précis (asymptotiquement) parmi les estimateurs réguliers.

- Plus la courbure de la log-vraisemblance est forte autour de $\boldsymbol{\theta}_0$ (information de Fisher élevée), plus l'estimation est précise.

Distribution asymptotique de l'EMV

- Sous des conditions de régularité (stationnarité, ergodicité, identifiabilité, $\boldsymbol{\theta}_0$ intérieur à l'espace des paramètres), l'EMV vérifie :

- **Convergence :**

$$\hat{\boldsymbol{\theta}} \xrightarrow{p} \boldsymbol{\theta}_0 \quad \text{quand } T \to \infty$$

- **Normalité asymptotique :**

$$\boxed{\sqrt{T}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\big(\mathbf{0},\, \mathcal{I}(\boldsymbol{\theta}_0)^{-1}\big)}$$

- En d'autres termes, pour T grand :

$$\hat{\boldsymbol{\theta}} \overset{a}{\sim} N\left(\boldsymbol{\theta}_0,\, \frac{1}{T}\mathcal{I}(\boldsymbol{\theta}_0)^{-1}\right)$$

Erreurs standard et intervalles de confiance

- En pratique, on estime la matrice de variance-covariance en remplaçant la matrice d'information par la **hessienne observée** :

$$\widehat{\mathbb{V}}(\hat{\boldsymbol{\theta}}) = \left[-\mathbf{H}(\hat{\boldsymbol{\theta}})\right]^{-1} = \left[-\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\boldsymbol{\theta}=\hat{\boldsymbol{\theta}}}\right]^{-1}$$

- L'**erreur standard** du i-ème paramètre est :

$$\mathsf{se}(\hat{\theta}_i) = \sqrt{\left[\widehat{\mathbb{V}}(\hat{\boldsymbol{\theta}})\right]_{ii}}$$

- Un **intervalle de confiance** à 95% pour θ_i est :

$$\hat{\theta}_i \pm 1{,}96 \times \mathsf{se}(\hat{\theta}_i)$$

Estimateur OPG

- Une alternative à la hessienne est l'estimateur **OPG** (*outer product of gradients*). On décompose la log-vraisemblance en contributions individuelles :

$$\mathcal{L}(\boldsymbol{\theta}) = \sum_{t=1}^{T} \ell_t(\boldsymbol{\theta})$$

- La matrice d'information est estimée par :

$$\widehat{\mathcal{I}}_{\text{OPG}} = \sum_{t=1}^{T} \frac{\partial \ell_t(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \frac{\partial \ell_t(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}$$

- Cet estimateur est facile à calculer (pas besoin de dérivées secondes), mais peut être moins précis en petit échantillon.

Cas du quasi-maximum de vraisemblance

- Si le modèle est mal spécifié (par exemple, on maximise une vraisemblance gaussienne alors que les erreurs ne sont pas gaussiennes), l'égalité de l'information **ne tient plus** :

$$\mathbf{A} = -\mathbb{E}[\mathbf{H}(\boldsymbol{\theta}_0)] \neq \mathbf{B} = \mathbb{E}\big[\mathbf{s}(\boldsymbol{\theta}_0)\mathbf{s}(\boldsymbol{\theta}_0)'\big]$$

- La distribution asymptotique du QMLE est alors :

$$\sqrt{T}\left(\hat{\boldsymbol{\theta}} - \boldsymbol{\theta}_0\right) \xrightarrow{d} N\big(\mathbf{0},\ \mathbf{A}^{-1}\mathbf{B}\,\mathbf{A}^{-1}\big)$$

- La matrice $\mathbf{A}^{-1}\mathbf{B}\,\mathbf{A}^{-1}$ est appelée **matrice sandwich** (ou estimateur de White/Huber). Il faut l'utiliser pour obtenir des erreurs standard robustes.

Estimation de la matrice sandwich

- En pratique, on estime :

$$\hat{\mathbf{A}} = -\frac{1}{T}\mathbf{H}(\hat{\boldsymbol{\theta}}) = -\frac{1}{T}\frac{\partial^2 \mathcal{L}}{\partial \boldsymbol{\theta}\, \partial \boldsymbol{\theta}'}\bigg|_{\hat{\boldsymbol{\theta}}}$$

- Et :

$$\hat{\mathbf{B}} = \frac{1}{T}\sum_{t=1}^{T} \frac{\partial \ell_t(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}} \frac{\partial \ell_t(\hat{\boldsymbol{\theta}})}{\partial \boldsymbol{\theta}'}$$

- L'estimateur robuste de la variance est alors :

$$\widehat{\mathbb{V}}_{\text{rob}}(\hat{\boldsymbol{\theta}}) = \frac{1}{T}\hat{\mathbf{A}}^{-1}\hat{\mathbf{B}}\,\hat{\mathbf{A}}^{-1}$$

- Si le modèle est correctement spécifié, $\hat{\mathbf{A}} \approx \hat{\mathbf{B}}$ et on retrouve l'estimateur classique.

Test de Wald

- On souhaite tester l'hypothèse linéaire $H_0 : \mathbf{R}\boldsymbol{\theta} = \mathbf{r}$ où $\mathbf{R}$ est une matrice $(r \times k)$ de rang r et $k = p + q + 2$.

- La **statistique de Wald** est :

$$W = (\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r})' \left[\mathbf{R}\,\widehat{\mathbb{V}}(\hat{\boldsymbol{\theta}})\,\mathbf{R}'\right]^{-1} (\mathbf{R}\hat{\boldsymbol{\theta}} - \mathbf{r})$$

- Sous H_0 et asymptotiquement :

$$W \xrightarrow{d} \chi^2(r)$$

- **Avantage :** Ne nécessite que l'estimation du modèle **non contraint**.

Test de Wald : cas scalaire

- Pour tester $H_0 : \theta_i = \theta_i^0$ (un seul paramètre), la statistique de Wald se simplifie en la t-**statistique** au carré :

$$W = \left(\frac{\hat{\theta}_i - \theta_i^0}{\mathsf{se}(\hat{\theta}_i)} \right)^2 \xrightarrow{d} \chi^2(1)$$

- De manière équivalente, la *t-statistique* :

$$t = \frac{\hat{\theta}_i - \theta_i^0}{\mathsf{se}(\hat{\theta}_i)} \xrightarrow{d} N(0,1)$$

- **Exemple :** Pour tester si un coefficient AR ou MA est significativement différent de zéro, on calcule $t = \hat{\theta}_i / \mathsf{se}(\hat{\theta}_i)$ et on rejette H_0 si $|t| > 1{,}96$ au seuil de 5%.

Test du rapport de vraisemblance

- On estime le modèle sous H_0 (contraint, $\tilde{\boldsymbol{\theta}}$) et sous H_1 (non contraint, $\hat{\boldsymbol{\theta}}$).
- La **statistique du rapport de vraisemblance** (*likelihood ratio*) est :

$$LR = 2\left[\mathcal{L}(\hat{\boldsymbol{\theta}}) - \mathcal{L}(\tilde{\boldsymbol{\theta}})\right]$$

- Sous H_0 et asymptotiquement :

$$LR \xrightarrow{d} \chi^2(r)$$

- **Avantage :** Ne nécessite pas le calcul de la matrice de variance-covariance.
- **Inconvénient :** Requiert l'estimation des deux modèles (contraint et non contraint).

Test du rapport de vraisemblance : exemple

- **Test d'un AR(1) contre un AR(2) :**
  - Modèle non contraint : $Y_t = c + \phi_1 Y_{t-1} + \phi_2 Y_{t-2} + \varepsilon_t$
  - Modèle contraint ($H_0 : \phi_2 = 0$) : $Y_t = c + \phi_1 Y_{t-1} + \varepsilon_t$
- On calcule :

$$LR = 2\left[\mathcal{L}_{\mathsf{AR}(2)}(\hat{\boldsymbol{\theta}}) - \mathcal{L}_{\mathsf{AR}(1)}(\tilde{\boldsymbol{\theta}})\right]$$

- On rejette H_0 si $LR > \chi^2_{1,0,95} = 3{,}84$ (seuil à 5%, 1 degré de liberté).
- Ce test est directement applicable à la sélection de l'ordre d'un modèle ARMA.

Test du multiplicateur de Lagrange

- Le test du **multiplicateur de Lagrange** (ou **test du score**) utilise uniquement l'estimation sous H_0 :

$$LM = \mathbf{s}(\tilde{\boldsymbol{\theta}})' \mathcal{I}(\tilde{\boldsymbol{\theta}})^{-1} \mathbf{s}(\tilde{\boldsymbol{\theta}})$$

où $\mathbf{s}(\tilde{\boldsymbol{\theta}})$ est le score évalué à l'estimateur contraint.

- Sous H_0 et asymptotiquement :

$$LM \xrightarrow{d} \chi^2(r)$$

- **Intuition :** Si H_0 est vraie, le score $\mathbf{s}(\tilde{\boldsymbol{\theta}})$ doit être proche de zéro. Un score élevé (en norme) fournit une preuve contre H_0.

- **Avantage :** Ne nécessite que l'estimation du modèle **contraint** (le plus simple).

Comparaison des trois tests

- Les trois tests (Wald, LR, LM) sont **asymptotiquement équivalents** sous H_0 : ils ont la même distribution limite $\chi^2(r)$.

- En échantillon fini, ils peuvent donner des résultats différents. On montre que :

$$W \geq LR \geq LM$$

  Le test de Wald rejette le plus souvent, le test LM le moins.

- **Choix pratique :**
  - Test de Wald : facile si le modèle non contraint est déjà estimé,
  - Test LR : le plus couramment utilisé pour comparer des modèles emboîtés,
  - Test LM : utile quand le modèle contraint est beaucoup plus simple.

Critères d'information

- Pour comparer des modèles **non emboîtés**, on utilise des critères d'information qui pénalisent la complexité.

- **Critère d'Akaike** (AIC, 1973) :

$$\text{AIC} = -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + 2k$$

où k est le nombre de paramètres estimés.

- **Critère bayésien de Schwarz** (BIC, 1978) :

$$\text{BIC} = -2\mathcal{L}(\hat{\boldsymbol{\theta}}) + k \log T$$

- On sélectionne le modèle qui **minimise** le critère. Le BIC pénalise davantage la complexité ($\log T > 2$ dès que $T \geq 8$) et sélectionne des modèles plus parcimonieux.

Critères d'information : propriétés

- **AIC :** Minimise asymptotiquement l'erreur quadratique moyenne de prévision. Tend à sélectionner des modèles légèrement sur-paramétrés.

- **BIC :** Sélectionne le vrai modèle avec probabilité tendant vers 1 quand $T \to \infty$ (**consistant**). Tend à sélectionner des modèles sous-paramétrés en petit échantillon.

- **En pratique :**
  - Si l'objectif est la **prévision**, l'AIC est souvent préféré.
  - Si l'objectif est l'**identification** du vrai modèle, le BIC est plus adapté.
  - On estime plusieurs modèles ARMA(p, q) pour $p, q \in \{0, 1, \ldots, p_{\max}\}$ et on retient celui qui minimise le critère choisi.

Plan

Approche de Box-Jenkins

- Box et Jenkins (1976) proposent une procédure en quatre étapes :
- **(1) Transformation :** Transformer les données pour obtenir une série approximativement stationnaire (différenciation, logarithme).
- **(2) Identification :** Choisir les ordres p et q à l'aide des autocorrélations et autocorrélations partielles empiriques.
- **(3) Estimation :** Estimer les paramètres $\phi(L)$ et $\theta(L)$ par maximum de vraisemblance.
- **(4) Diagnostic :** Vérifier que le modèle estimé est compatible avec les données observées.

Autocorrélations empiriques

- L'**autocorrélation empirique** d'ordre j est :

$$\hat{\rho}_j = \frac{\hat{\gamma}_j}{\hat{\gamma}_0}$$

où

$$\hat{\gamma}_j = \frac{1}{T} \sum_{t=j+1}^{T} (y_t - \bar{y})(y_{t-j} - \bar{y}), \quad \bar{y} = \frac{1}{T} \sum_{t=1}^{T} y_t$$

- Si les données proviennent d'un processus MA(q), alors $\rho_j = 0$ pour $j > q$. On s'attend donc à ce que $\hat{\rho}_j \approx 0$ pour $j > q$.

- Sous l'hypothèse de bruit blanc, $\hat{\rho}_j$ est approximativement distribué selon $N(0, 1/T)$. L'intervalle de confiance à 95% est $\pm 2/\sqrt{T}$.

Autocorrélations partielles empiriques

- L'**autocorrélation partielle** empirique d'ordre m est le dernier coefficient $\hat{\alpha}_m^{(m)}$ dans la régression :

$$y_{t+1} = \hat{c} + \hat{\alpha}_1^{(m)} y_t + \hat{\alpha}_2^{(m)} y_{t-1} + \cdots + \hat{\alpha}_m^{(m)} y_{t-m+1} + \hat{e}_t$$

- Si les données proviennent d'un processus AR(p), alors l'autocorrélation partielle est nulle pour $m > p$:

$$\alpha_m^{(m)} = 0 \quad \text{pour } m > p$$

- Sous cette hypothèse, $\mathbb{V}(\hat{\alpha}_m^{(m)}) \approx 1/T$ et l'intervalle de confiance à 95% est $\pm 2/\sqrt{T}$.

Règles d'identification

- **Processus MA(q) :** Les autocorrélations $\hat{\rho}_j$ sont significativement non nulles pour $j \leq q$, puis deviennent nulles au-delà. Les autocorrélations partielles décroissent progressivement.

- **Processus AR(p) :** Les autocorrélations partielles $\hat{\alpha}_m^{(m)}$ sont significativement non nulles pour $m \leq p$, puis deviennent nulles au-delà. Les autocorrélations décroissent progressivement (mélange d'exponentielles ou de sinusoïdes amorties).

- **Processus ARMA(p, q) :** Les deux fonctions décroissent progressivement. L'identification est plus difficile et demande souvent d'essayer plusieurs spécifications.

Philosophie de la parcimonie

- Box et Jenkins insistent sur le principe de **parcimonie** : utiliser le moins de paramètres possible.

- Un modèle avec trop de paramètres :
  - s'ajuste bien aux données historiques (sur-ajustement),
  - mais prévoit mal hors échantillon.
- La découverte que des modèles ARMA avec de petites valeurs de p et q produisent souvent de meilleures prévisions que les grands modèles macroéconométriques a été un résultat marquant.

- En pratique, des valeurs de p et q inférieures ou égales à 2 ou 3 suffisent pour la plupart des applications.

Résumé

- Le **maximum de vraisemblance** est le principe d'estimation dominant.
- Pour les processus **AR purs**, l'estimation se réduit aux MCO.
- Pour les processus **MA** et **ARMA**, la vraisemblance est non linéaire et nécessite une optimisation numérique.
- La vraisemblance **conditionnelle** simplifie les calculs mais nécessite l'inversibilité. La vraisemblance **exacte** est plus robuste.
- L'EMV est **asymptotiquement normal** avec une variance atteignant la borne de Cramér-Rao.
- Les tests de **Wald**, du **rapport de vraisemblance** et du **multiplicateur de Lagrange** permettent de tester des hypothèses sur les paramètres.
- Les critères **AIC** et **BIC** complètent l'identification par les autocorrélations pour la sélection de modèles.