UPPA

Data Mining

Modèles ARMA

Rebecca Courtenay
M2 Big Data

M. Kojadinovic

2022 - 2023

Table des matières

1 Introduction **2**

2 Série temporelle **3**

3 ARMA **4**
3.1 AR : Autorégressive 4
3.1.1 Caractérisation du processus 4
3.2 MA : Moyenne mobile 5
3.2.1 Caractérisation du processus 5
3.3 ARMA 5
3.3.1 Caractérisation du processus 5

4 Processus non stationnaire **6**
4.1 ARIMA 6
4.2 SARIMA 7

5 Application **8**
5.1 Stationnarisation 8
5.2 Choix du modèles 8
5.3 Qualité prédictive du modèle 9

6 Sources **10**

Chapitre 1

Introduction

Les modèles ARMA sont les principaux modèles de séries temporelles. On les utilise notamment pour prédire les valeurs futures d'une série à l'aide des valeurs du passés, cela permet d'anticiper l'évolution d'un phénomène. On peut retrouver cette méthode dans la prédiction de fin de vie d'un système ou dans la prédiction de la météo. C'est un outil de prédiction qui permet de voir la fluctuation des valeurs sur une échelle de temps.

Dans ce document, nous allons d'abord voir ce qu'est une série temporelle avant de nous focaliser sur la méthode ARMA. Pour finir, nous verrons deux modèles : ARIMA et SARIMA, qui sont des combinaisons du processus ARMA avec, en plus, un développement de différenciation.

Chapitre 2

Série temporelle

Une série temporelle est une série de données indexée par le temps. Ce temps peut avoir le cycle que l'on souhaite : secondes, minutes, mois, saisons de l'année et bien d'autres. Cependant, cela reste un ensemble fini de valeurs numériques. Ces séries sont utilisées pour voir l'allure d'une variable dans le temps. Cela peut permettre de prédire l'allure de cette variable avec une échelle de temps.

Nous pouvons décomposer une série temporelle. Nous pouvons la traduire mathématiquement tel quel :

$$X_t = T_t + S_t + \epsilon_t$$

Avec :
- X_t : la série temporelle ;
- T_t : la tendance - explique le comportement de la série au cours du temps (croissant/décroissant) ;
- s_t : la saisonnalité - explique la présence d'un phénomène périodique au cours du temps (par exemple : il peut avoir des piques au même moment de l'année) ;
- ϵ_t : le résidu - représente la partie non expliqué par la décomposition.

Nous nous intéressons au résidu car celui-ci permet de voir si toutes les composantes temporelles sont bien expliquées. Cela à l'aide d'un modèle que nous allons utiliser pour expliquer notre série temporelle.
Il peut avoir deux types de résidu :
- Stationnaire : les composantes sont bien explquées car le processus n'évolue pas avec le temps (la moyenne et la variance sont constantes) ;
- Non stationnaire : les composantes sont mal expliquées.

Les modèles ARMA permettent donc d'expliquer le résidu de la décomposition avec des méthodes tel que AR (autorégressive) et MA (moyenne mobile). Et lorsque nous avons un résidu non stationnaire, nous pouvons utliser des modèles contenant la méthode ARMA et un processus de différenciation en leur sein, tel que ARIMA et SARIMA.
Le but est d'avoir, à la fin, un résidu stationnaire aléatoire et décorrélé.
C'est pourquoi, les modèles ARMA sont utilisés pour expliquer les séries temporelles.

Chapitre 3

ARMA

Le modèle ARMA n'est défini uniquement s'il est stationnaire. C'est-à-dire que les propriétés statistiques telle que l'espérence des variables, soit indépendante du temps.

Ce modèle est composé de deux parties :
— Autorégressive (AR) ;
— Moyenne mobile (MA).

3.1 AR : Autorégressive

Un processus x_t est un déroulement autorégressif d'ordre p si on peut expliquer une relation linéaire entre les X_t, sa valeur à l'instant T, et les p termes prédécents X_t. Cet méthode est noté $AR(p)$.
Mathématiquement, c'est-à-dire, si :

$$\forall t, X_t = c + \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t$$

Avec :
— $(\varphi_1, ..., \varphi_p)$: paramètres du modèle (des réels) ;
— ϵ_t : les erreurs ;
— c : une constante.
La constate peut être omise car généralement nous avons des processus centrés (avec une espérance nul). Si celui-ci ne l'est pas, nous pouvons le centrer : $X_t - \mu_X$.

Dans ce cas, le passé de la série explique la série à l'instant t.

3.1.1 Caractérisation du processus

De plus ce qui caractérise cet méthode est d'avoir une autocorrélation partielles qui :
— Qui s'annule à partir du rang p+1
— N'est pas nul à l'ordre p
— Dont les autocorrélations simples décroissrons rapidement vers 0 (de manière exponentielle ou sinusoïdale amortie)

3.2 MA : Moyenne mobile

Un processus X_t est un déroulement de moyenne mobile (moving average) d'ordre q si on peut exprimer sa valeur à l'instant T comme une combinaison linéaire d'erreur aléatoire. Cet méthode est noté $MA(q)$.
Mathématiquement, c'est-à-dire, si :

$$\forall t, X_t = \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Avec :
— $(\theta_1, ..., \theta_q)$: paramètres du modèle (des réels) ;
— ϵ_t : les erreurs.
Dans ce cas, on insinue qu'une perturbation (erreur) puis des pertubartions futures ont engendré des informations.

3.2.1 Caractérisation du processus

De plus, ce qui caractérise ce processus est d'avoir une autocorrélation simple qui :
— S'annule à partir du rang q+1 ;
— N'est pas nul à l'ordre q ;
— Dont les autocorrélations partielles décroissrons rapidement vers 0.

3.3 ARMA

Le modèle ARMA (Auto Regressive Moving Average) est donc un mixte entre ces deux modèles. Il est noté $ARMA(p, q)$.
Le processus X_t est une combinaison linéaire du passé de X_t jusqu'à $t - p$ et une combinaison linéaire du bruit blanc (l'erreur) et du q instant de ce bruit blanc. C'est-à-dire :

$$\forall t, X_t = \sum_{i=1}^{p} \varphi_i X_{t-i} + \epsilon_t + \sum_{i=1}^{q} \theta_i \epsilon_{t-i}$$

Avec :
— $(\varphi_1, ..., \varphi_p)$ et $(\theta_1, ..., \theta_q)$: paramètres du modèle (des réels).

3.3.1 Caractérisation du processus

Concernant ce déroulement, il n'y a pas de caractérisation simple du processus X_t comme avec l'autorégression et la moyene mobile.

Chapitre 4

Processus non stationnaire

Comme dit précédemment, le modèle ARMA ne peut modéliser que des séries temporelles stationnaires.
Pour palier ce problème, il existe deux modèles qui intègrent un processus de différenciation permettant de rendre notre série stationnaire.

4.1 ARIMA

Lorsque notre série temporelle présente ne qu'une tendance, non constante, nous pouvons utiliser la méthode ARIMA. Celui-ci permet de retirer les tendances que la série présente pour les stationnariser.

Cet méthode présente la même structure que celui d'ARMA mais au lieu de travailler sur le processus X_t nous travaillons sur celui $\nabla^d X_t$. C'est pourquoi nous rajoutons un I dans le nom du modèle, I pour intégration.
Avec ∇^d opérateur de différenciation tel que :

$$\nabla^d = (I - B)^d$$

Où B est l'applicateur retard, c'est-à-dire $\frac{X_t}{X_{t-1}}$.

Et d est l'ordre de différenciation.

Nous appelons donc le modèle ARIMA : $ARIMA(p, d, q)$.

Par exemple, si nous travaillons avec une tendance linéaire de la forme :

$$X_t = \alpha + \beta_t + \epsilon_t$$

Nous aurons :

$$X_t - X_{t-1} = \beta + \epsilon_t - \epsilon_{t-1}$$

De ce fait, la dépendance temporelle est enlevée et la différence $X_t - X_{t-1}$ est stationnaire.
En général, une tendance est supprimée le nombre de fois qu'il y a d'ordre mais toujours au temps T moins le temps T-1. De ce fait, il y a une différenciation d fois notre série temporelle avant de pouvoir le modéliser comme un ARMA.

4.2 SARIMA

Si la série temporelle présente une tendance ainsi qu'une saisonnalité, on peut utiliser la méthode SARIMA. Ce modèle est comme celui d'ARIMA mais prends en compte la saisonnalité en plus. En effet, il rajoute une autre différenciation en saisonnalité.
De ce fait, il y aura une abstraction de l'effet saisonnier.

Ici aussi, nous appliquons ce déroulement de différenciation avec un ordre s. Nous appelons donc le modèle ARIMA : $SARIMA(p, d, q, s)$.

Et, en plus d'ajouter une différenciation, il ajoute un polynôme AR saisonnier ainsi qu'un MA saisonnier.

Chapitre 5

Application

Afin d'avoir un modèle d'une série temporelle qui se ramène à une méthode ARMA, il nous faut d'abord savoir si nous avons des séries temporelles qui présentent une tendance et/ou saisonnalité. Si cela est le cas, il nous faut stationnariser nos données. Puis nous pourrons commencer à estimer des modèles et choisir celui qui donne les meilleures prédictions.

5.1 Stationnarisation

Il existe plusieurs techniques dont :
— Décomposition saisonnière : retranche la tendance et la saisonnalité à la série temporelle ;
— Différenciation : directement intégré dans méthode ARIMA et SARIMA ;
— Méthode de Box-cox : stationnarise en "variance".

Pour savoir si nous devons différencier notre série, on peut calculer l'ACF (autocorrelation function). En utilisant les caractéristiques des méthodes vu dans le chapitre 3, nous savons que lorsque nous avons une autocorrélation qui décroit rapidement, nous sommes dans la situation d'une série stationnaire. Ici, l'ACF peut avoir en sortie l'estimation de l'autocorrélation simple, cela voudra donc dire que notre processus est stationnaire. Or, si en sortir nous avons une estimation qui décroit lentement vers 0, nous serons dans un cas non stationnaire. De ce fait, si il est non stationnaire, nous appliquons une différenciation de tendance d'ordre q et une différenciation de saisonnalité d'ordre s. Ces deux transformations sont appliquées de manière itératives, nous appliquons cette différenciation jusqu'à que notre ACF décroit rapidement vers 0.
La fonction de l'autocorrélation est tel que :

$$p_k = \frac{\gamma_t}{\sigma^2}$$

Avec :

$$\gamma_t = \mathbb{E}((X_t - \mu)(X_{t+k} - \mu)$$

5.2 Choix du modèles

Une fois que nous avons stationnarisé nos données, nous pouvons commencer à chercher le meilleur modèle ARMA.

Pour cela, il existe diverse méthode pour estimer AR et MA :

— Critère d'information : utilise des algorithmes de recherche automatique qui estime les ordres du modèles pouvant minimiser un critère d'information (compromis entre la partitionnement du modèle et son ajustement) ;
— Heuristique : calcul quelles sont les valeurs de l'autocorrélation simple significatives pour en déduire la structure de la partie MA et nous pouvons faire de même avec les valeurs l'autocorrélation partielle pour en déduire le partie AR (démarche itérative).

Puis, il est préférable d'utiliser la méthode du maximum de vraisemblance afin d'estimer les modèles potentiels que nous venons d'identifier. Mais n'ayant pas de solution littéral, il est préférable d'ensuite appliquer un algortihme d'optimisation qu'on initialise sur les résultats que l'on peut obtenir avec des équations tel que Yule Walker (équations qui établissent une correspondance directe entre les paramètres du modèle et ses autocovariances et qui utilisé dans le cas de l'autorégression) ou l'algortihme des innovations (utilisé dans le cas des moyennes mobiles).

Ensuite, nous vérifions nos modèles afin de choisir le meilleur. Pour cela, nous regardons deux paramètres :

— Significativité des paramètres estimés : se base sur un test du student pour chaque paramètre ;
  — $H0$: le processus est un $ARMA(p-1,q)$;
  — $H1$: le processus est un $ARMA(p,q)$.
— Blancheur du résidu : test de blancheur tel que le test Ljung-Bix qui teste l'autocorrélation d'ordre supérieur à 1.

Or, s'il y a plus de un modèle qui est validé par ces critères, nous choisirons le modèle qui a un critère d'information plus faible.

Ainsi, nous pouvons lancer la phase de prévision.

5.3 Qualité prédictive du modèle

Pour finir, nous cherchons à être sûr de la qualité de prédiction du modèle que nous venons de choisir. Pour cela, nous prenons un jeu de données train (celui sur lequel on entraîne notre modèle) et un jeu de données test. Puis, nous calculons des critères d'erreur tel que :

— RMSE (erreur quadratique moyenne) : $\sqrt{\frac{1}{T}\sum_{t=1}^{T}(x_t - \hat{x_t})^2}$;
— MAPE (erreur relative en moyenne) : $\frac{1}{T}\sum_{t=1}^{T} \| \frac{x_t - \hat{x_t}}{x_t} \|$.

Chapitre 6

Sources

DataScientest :
https ://datascientest.com/series-temporelles
https ://datascientest.com/arima-series-temporelles

Wikipedia :
https ://fr.wikipedia.org/wiki/Op
https ://fr.wikipedia.org/wiki/ARMA

OpenClassrooms :
https ://openclassrooms.com/en/courses/4525371-analysez-et-modelisez-des-series-temporelles
365DataScience :
https ://365datascience.com/tutorials/time-series-analysis-tutorials/arma-model/