**Ανάλυση Χρονοσειρών**

**Υπολογιστική Εργασία 2018-2019**

|  |  |  |
| --- | --- | --- |
| Σαμπαζιώτης Σάββας | 7974 | sampazio@ece.auth.gr |
| Κεχαγιάς Ανδρέας | 8160 | andreaskp@ece.auth.gr |

**Εισαγωγή** Εισαγωγή github Links για github kai code

Το κείμενο αυτό αποτελεί την αναφορά στην υπολογιστική εργασία του μαθήματος Ανάλυσης Χρονοσειρών, του ακαδημαϊκού έτους 2018-2019. Παρακάτω παρουσιάζονται οι δύο διαφορετικές μελέτες που προδιαγράφονται στην εκφώνηση της εργασίας, πάνω στη δεδομένη χρονοσειρά ***dat19.mat.***



Η παραπάνω χρονοσειρά διαχωρίστηκε συνολικά σε 11 μη-επικαλυπτόμενα παράθυρα των 250 δειγμάτων τα οποία παρουσιάζονται παρακάτω.



**Μελέτη Α**

1. **Απαλοιφή Τάσης**

Πρωταρχική ανάγκη της ανάλυσης είναι η μετατροπής της κάθε χρονοσειράς σε στάσιμο σήμα. Για κάθε χρονοσειρά, δοκιμάστηκαν διάφορες μέθοδοι μετασχηματισμού. Η τελικές στάσιμες χρονοσειρές προκύπτουν από την μέθοδο του Κινούμενου Μέσου τάξης 21.



Ο ακριβής μετασχηματισμός που εφαρμόστηκε περιγράφεται στον παρακάτω ψευδοκώδικα. Η υλοποίηση και παρουσίαση γίνεται στο m-script *“convert2Stationary.m”*

|  |  |
| --- | --- |
| Αλγόριθμος | Σχόλια |
|  | *Φόρτωση χρονοσειράς dat19* |
|  | *Αφαίρεση εξομαλυμένου σήματος (εκτιμώμενη τάση) από το πρωταρχικό.* |
| 1. *Για κάθε*     1. *b=k\*250* | *Διαχωρισμός Χρονοσειράς σε 11 μη-επικαλυπτόμενα παράθυρα χρονοσειρών , 250 δειγμάτων η κάθε μία.* |
|  | *Αφαίρεση εξωκείμενων τιμών βάσει του κανόνα της διασποράς: Τα δείγματα εκτός των ορίων , αντικαθίστανται με τιμές από γραμμική παρεμβολή των γειτονικών του.* |
|  | *Απαλοιφή μέσου όρου* |
|  | *Κανονικοποίηση κατά διασπορά* |
| 1. *Τέλος* |  |

Όπως παρατηρεί κανείς, επιλέχθηκε να γίνει η απαλοιφή της συνολικής τάσης και έπειτα να γίνει ο διαχωρισμός στα παράθυρα των 250 χρονοσειρών. Με αυτόν τον τρόπο, ελαχιστοποιήθηκε το πλήθος των δειγμάτων που θα «παραμελούνταν» από το φιλτράρισμα, καθώς η τεχνική Κινούμενου Μέσου τάξης , παράγει τοπικές τιμές στο διάστημα και μόνο, αφήνοντας δείγματα αφιλτράριστα.

**Εναλλακτικές τεχνικές που εξετάσθηκαν**

Όπως αναφέρθηκε αρχικά, πραγματοποιήθηκαν και επιπλέον δοκιμές απαλοιφής της τάσης, όπως εκτίμηση της τάσης με MA φιλτράρισμα μικρότερης τάξης, η εφαρμογή πρώτων και δεύτερων διαφορών, μέχρι και η εξέταση στασιμότητας με το Augmented Dickey Fuller Test***.*** Ωστόσο, το ADF test αποτελούσε ιδιαίτερη διερεύνηση για τη επιλογή κατάλληλου πλήθους lags, με αποτέλεσμα να εγκαταλειφτεί ως μετρική στασιμότητας.

Το βασικό δίλλημα ήταν μεταξύ των απλών πρώτων διαφορών και αυτή του κινούμενου μέσου. Εν τέλει, επιλέχθηκε η ΜΑ βάσει της παρακάτω παρατήρησης στις δειγματικές μερικές αυτοσυσχετίσεις που προκύπτουν:



* Παρατηρούμε, πως στη περίπτωση των πρώτων διαφορών, η χρονοσειρά είναι πλέον στάσιμη μεν, αλλά με ενισχυμένο θόρυβο δε. Το γεγονός αυτό αντικατοπτρίζεται και στην δειγματική αυτοσυσχετίσεις και μερική αυτοσυσχέτιση, η οποία προσεγγίζει αυτές των idd διαδικασιών.
* Αντίθετα, η MA(21) τεχνική, περιορίζει τον θόρυβο και αναδεικνύει με όμορφο τρόπο την χρήσιμη πληροφορία την οποία αποζητάμε να προβλέψουμε και να περιγράψουμε. Επίσης, η μερική αυτοσυσχέτιση φανερώνει την τάξη AR μοντέλου που θα μελετηθεί στην συνέχεια.

Βάσει των παραπάνω συμπερασμάτων, επιλέχθηκε η MA μέθοδος, καθώς αναμένεται να διευκολύνει σημαντικά την επιλογή κατάλληλου γραμμικού μοντέλου μετέπειτα.

Θεωρητικά, αναμένεται η χρήση ενός ARMA(p,q) μοντέλου με υψηλό p και με πρώτες διαφορές για απαλοιφή τάσης, να έχει παρόμοια απόδοση με AR μοντέλου με MA απαλοιφή τάσης. Ωστόσο, για τον προαναφερθέντα λόγο, επιλέχθηκε η τακτική που θα διευκολύνει την μοντελοποίηση μετέπειτα.

1. **Αυτοσυσχέτιση Χρονοσειρών**

Παρακάτω παρουσιάζονται οι συναρτήσεις δειγματικών αυτοσυσχετίσεων των 11 χρονοσειρών.



Η φθίνουσα αυτοσυσχέτιση έχει συντελεστές μεγάλης σημαντικότητας για υστέρηση . Έτσι, η υπόθεση πως κάποια από τις χρονοσειρές είναι λευκός θόρυβος απορρίπτεται με μεγάλη ευκολία.

Για λόγους πληρότητας παραθέτουμε και την Δειγματική Μερική Αυτοσυσχέτιση των σημάτων. Παρατηρούμε πως για υστέρηση , η μερική αυτοσυσχέτιση είναι –έστω και οριακά– στατιστικώς ασήμαντη.



1. **Επιλογή Γραμμικού Μοντέλου**

Ως μία πρώτη εκτίμηση του γραμμικού μοντέλου, αξιοποιήθηκε η τακτική της μερικής αυτοσυσχέτισης. Όπως προϊδεάστηκε προηγουμένως, η μερική αυτοσυσχέτιση που διαθέτουμε μετά την απαλοιφή της τάσης εμφανίζει με ξεκάθαρο τρόπο την τάξη ενός AR μοντέλου.

Συνοπτικά, το επιλεγμένο μοντέλο της κάθε χρονοσειράς και τα κριτήρια απόδοσης αυτών είναι τα εξής:

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Χρονοσειρά | Τάξη  Μοντέλου | Προκύπτουσες Παράμετροι | AICC | NRMSE |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

**Επιλογή Τάξης Μοντέλου – Επιπλέον Σχόλια**

* H μέση τιμή έχει απαλειφτεί και είναι πλέον, κατά το de-trending.
* Παρατηρούμε πως για όλες σχεδόν τις χρονοσειρές επιλέχθηκαν AR{2} μοντέλα.
* Σε πολλαπλές χρονοσειρές, παρόμοιες στατιστικά σημαντικές αυτοσυσχετίσεις εμφανίζονται για πολύ μεγαλύτερες υστερήσεις. Τέτοιο παράδειγμα είναι η χρονοσειρά για υστέρηση (βλ γράφημα ΧΧ). Ωστόσο, λόγο της μεγάλης τάξης μοντέλου που εξ’ υπονοείται καθώς και οι επιπλέον στατιστικές ασήμαντες μερικές αυτοσυσχετίσεις που μεσολαβούν στα lags 3,4 και 5, αποφασίστηκε τέτοιου είδους απομακρυσμένες συντελεστές μερικές αυτοσυσχέτισης να αγνοηθούν.



**Σχόλια –Συμπεράσματα**

Οι Μερικές Αυτοσυσχετίσεις του συνόλου διαγραμμάτων ΧΧ, δείχνουν πως οι διάφορες χρονοσειρές δεν έχουν ιδιαίτερη διαφορά στην Δειγματική Μερική Αυτοσυσχέτιση. Το γεγονός αυτό ανακλάται και στις προκύπτουσες παραμέτρους, όπου επίσης φαίνονται να μην έχουν ιδιαίτερη διακύμανση από μοντέλο σε μοντέλο. Ενδεχομένως, αυτό να αποτελεί ένδειξη του ότι ο μηχανισμός από τον οποίο παράγονται τα δεδομένα είναι αιτιοκρατικό χρονοαμετάβλητο σύστημα.

**Επιπλέον Ανάλυση: Χρήση των Κριτηρίων Πληροφορίας**

Έχοντας επιλέξει ως θεματική εργασία τα «Κριτήρια Πληροφορίας και τις Ιδιότητές» τους, νιώσαμε την επιθυμία να τα αξιοποιήσουμε για την επιλογή μοντέλου και να συγκρίνουμε με την μέθοδο της PACF. Επίσης, αποτελούν μία καλή ευκαιρία να εξετάσουμε **αν τυχόν υπάρχει ARMA μοντέλο** το οποίο αποδίδει καλύτερα από τα AR που αναφέρονται παραπάνω.

Το κριτήριο πληροφορίας που επιλέχθηκε είναι το . Επιλέξαμε την έκδοση αυτή του Akaike Κριτηρίου, για το λόγο του ότι το πλήθος των δειγμάτων εκπαίδευσης είναι αρκετά μικρό. Επίσης, όπως το δείξαμε στην θεματική εργασίας μας, η χρήση των Bayesian IC για επιλογή ARMA μοντέλων για τόσο μικρό πλήθος εκπαίδευσης δεν είναι ιδιαίτερα αξιόπιστη.

Στο παρακάτω πίνακα διαγραμμάτων παρουσιάζονται τα αποτελέσματα του για πολλαπλές τιμές τάξης MA και AR.

ΤΟΔΟ διάγραμμα

Παρατηρούμε πως η επιπλέον χρήση του ΜΑ στο μοντέλο δεν προσθέτει κάποια αξία στην απόδοση έναντι της αυξημένης πολυπλοκότητας που συνεπάγεται. Επίσης, τα κριτήρια πληροφορίας έρχονται σε συμφωνία με την αρχική επιλογή μοντέλου βάσει της PACF.

Τα τελικά μοντέλα που θα αξιοποιηθούν είναι αυτά του πίνακα ΧΧ.

1. **Εφαρμογή AR{5} μοντέλου.**

Όπως προδιαγράφεται στην εκφώνηση της εργασίας, εκπαιδεύτηκε AR{5} μοντέλο με σύνολο εκπαίδευσης τα πρώτα 150 δείγματα της κάθε χρονοσειράς, και αξιολογήθηκε η απόδοση του μοντέλου στα υπόλοιπα 100 δείγματα –σύνολο αξιολόγησης.

Παρακάτω σε παρουσιάζονται σε πίνακα τα σφάλματα πρόβλεψης NRMSE, των ανά χρονοσειρά.

|  |  |
| --- | --- |
| Χρονοσειρά |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |
|  |  |

Δεδομένου πως στο προηγούμενο ερώτημα δηλώθηκε πως καλύτερο μοντέλο αποτελεί ανά χρονοσειρά είναι το , αξίζει να συγκριθεί η απόδοση του με το Ωστόσο, οφείλεται να γίνει επανεκπαίδευση του μοντέλου με τις Cross-Validation συνθήκες που ορίστηκαν για το AR{5}, όπου δεδομένα εκμάθησης είναι το 60% των δειγμάτων και δεδομένα ελέγχου το υπόλοιπο 40%.

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| Χρονοσειρά |  |  | Βελτίωση % | Προτιμητέο Μοντέλο |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |
|  |  |  |  |  |

Παρατηρούμε πως στην πλειοψηφία των περιπτώσεων το AR{2} μοντέλο έχει καλύτερη απόδοση από το AR{5}. Μερικές εξαιρέσεις αποτελούν οι χρονοσειρές 7,9, και 11, όπου αναδεικνύουν το AR{5} με βελτίωση του NRMSE έναντι του AR{2} κατά . Αντιθέτως, παρατηρούμε πως το AR{2} έχει βελτίωση στην απόδοση έναντι του AR{5} από 10% έως και 40%.

**Μελέτη B**

Στο κομμάτι αυτό της εργασίας παρουσιάζεται η γραμμική και μη-γραμμική ανάλυση επί της χρονοσειράς με υποδειγματοληψία. Η χρονοσειρά αυτή παρουσιάζει την ισοτιμία του νομίσματος σε εβδομαδιαία βάση.



Αμέσως παρατηρεί κανείς μία διαφορετική «καθαρότητα» στην χρονοσειρά, καθώς ημερήσιος θόρυβος έχει πρακτικά φιλτραριστεί.

1. **Εφαρμογή Γραμμικών Μοντέλων.**

**Απαλοιφή τάσης**

Στα πλαίσια αυτής της χρονοσειράς δοκιμάστηκε η εκτίμηση της τάσης με πολυώνυμο. Όπως διαπιστώθηκε αρκετά γρήγορα η χρήση πολυωνύμου έναντι Φίλτρου Κινούμενου Μέσου, έχει σχεδόν πανομοιότυπα αποτελέσματα.



Η τελική χρονοσειρά που θα αξιοποιηθεί είναι αυτή από την απαλοιφή με χρήση πολυωνύμου.



**Αυτοσυσχέτιση – Μερική Αυτοσυσχέτιση**

Παρακάτω παρουσιάζεται η Δειγματική Συνάρτηση Αυτοσυσχέτισης και Δειγματική Συνάρτηση Μερικής Αυτοσυσχέτισης.

****

****

Ομοίως με την Μελέτη Α, παρατηρούμε ισχυρές ενδείξεις για χρήση μοντέλου AR, βάσει της PACF. Συγκεκριμένα, παρατηρούμε προς έκπληξη μας πως το μοντέλο AR{1} αποτελεί πρώτη επιλογή μοντέλου, καθώς δεν υπάρχουν ισχυροί συντελεστές μερικής αυτ/σης για υστέρηση .

Εξαίρεση στην παραπάνω παρατήρηση αποτελούν

* O ο οποίος περνάει οριακά το νέο όριο στατιστικής σημαντικότητας.
* O ο οποίος υποδεικνύει μία **έντονη αυτοσυσχέτιση ανά τρίμηνο.** Πληροφορία ενδεχομένως αρκετά σημαντική για οικονομετρικές εφαρμογές.

**Επιλογή Κατάλληλου Μοντέλου – Εφαρμογή AR{5}**

Όπως αναφέρθηκε παραπάνω, βάσει της μερικής αυτοσυσχέτισης, κατάλληλο μοντέλο αποτελεί το AR{1}. Βάσει του NRMSE που προκύπτει από τα δεδομένα ελέγχου (Cross Validation 60%-40% όπως στο A.4), τα μοντέλα AR{1} και AR{5} αποδίδουν τα παρακάτω αποτελέσματα.

|  |  |  |
| --- | --- | --- |
| Μοντέλο | NRMSE | Παράμετροι Μοντέλου |
| AR{1} | 0.1635 |  |
| AR{5} | 0.1776 |  |

Παρατηρούμε ελαφρώς καλύτερη απόδοση πρόβλεψης για το απλό μοντέλο AR{1} έναντι του AR{5}.





1. **Μη Γραμμική Ανάλυση**

Στην ενότητα αυτή θα παρουσιάζουμε την μη-γραμμική ανάλυση που πραγματοποιήθηκε επί της χρονοσειράς.

1. **Γραφήματα Διασποράς**

Πρώτο ζητούμενο ήταν τα διαγράμματα διασποράς μεταξύ των στις 2 διαστάσεις και των στις 3 διαστάσεις.

Αρχικά, σε ότι αφορά το διάγραμμα διασποράς μεταξύ της πρώτης υστέρησης, παρατηρούμε αμέσως μία γραμμική αναλογία. Αυτό εκφράζει έντονη συσχέτιση μεταξύ των δειγμάτων, όπως και έχει αναδειχτεί στην γραμμική ανάλυση της ενότητας Β.1.



Αναφορικά με την συσχέτιση των παρατηρεί κανείς στα scatterplots τα εξής:

 **

Παρόμοια αυτοσυσχέτιση παρατηρούμε και για υστέρηση αν και λιγότερο έντονη. Το γεγονός προβλήθηκε τελείως αντίθετα στην μερική αυτοσυσχέτιση του Β.1 και για αυτό δεν εξετάστηκαν τέτοια μοντέλα.

Ωστόσο, προβάλλοντας τον διάγραμμα διασποράς και στις 3 διαστάσεις παρατηρεί κανείς πως η γραμμική αυτή συσχέτιση όντως ισχύει μεταξύ των σημείων



Αν η εργασία αφορούσε εφαρμογές τύπου Principal Component Analysis (PCA), ή άλλα σχετικά εργαλεία μείωσης διαστάσεων δεδομένων και feature selection, η αντιμετώπιση θα είχε ως εξής: Θα εφαρμόζαμε κατάλληλο (γραμμικό) μετασχηματισμό και θα απεικονίζαμε την παραπάνω φαινομενικά τρισδιάστατη συσχέτιση ως αυτό που πραγματικά είναι: Μία δισδιάστατη γραμμική συσχέτιση μεταξύ δύο νέων μεταβλητών , κατανεμημένες κατά παρόμοιο τρόπο όπως στον αρχικό χώρο καταστάσεων.

Ωστόσο, στα πλαίσια της ανάλυσης χρονοσειρών διατίθενται εργαλεία όπου με παρόμοια τοπολογική προσέγγιση δίνουν την δυνατότητα ανάλυσης και εκπαίδευσης μοντέλων.

1. **Ανακατασκευή του State-Space – Διάσταση Εμβύθισης**

Έγινε χρήση του αλγορίθμου False Nearest Neighbors, για εξακρίβωση της διάστασης εμβύθισης m.



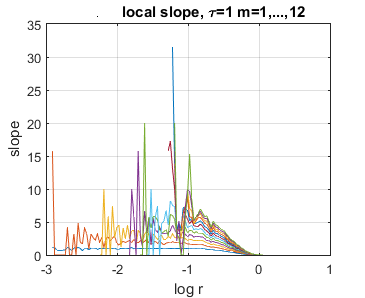
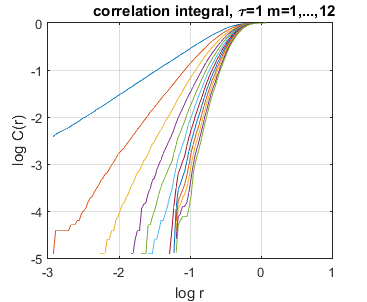
Ο αλγόριθμος εφαρμόστηκε με τις εξής παραμέτρους:

* Μέγιστη διάσταση εμβίθυσης: 10
* Ποσοστιαίο Κατώφλι Μεταβολής απόστασης (παράμετρος escape/f) = 10%

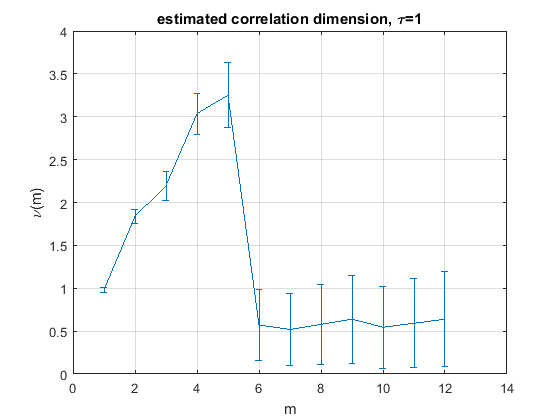
Η βέλτιστη διάσταση εμβύθισης για την οποία πρακτικά δεν έχουμε πλέον αυτοτομές είναι **.**

1. **Διάσταση Συσχέτισης**

Σε αυτό το ζητούμενο, ζητείται η διάσταση συσχέτισης της του φασικού πορτρέτου για διαφορετικές διαστάσεις εμβύθισης. Παρακάτω βλέπουμε το log-log διάγραμμα του αθροίσματος συσχέτισης έναντι της ακτίνας υπερσφαίρας r (διαγραμμα ΧΧ). Δίπλα βλέπουμε την εκτιμώμενη διάσταση συσχέτισης, έναντι της ίδιας ακτίνας r (διάγραμμα ΧΧ)



Τέλος, στο παρακάτω διάγραμμα παρουσιάζεται η διακύμανση της κλίσης v(r), για κάθε m



Παρατηρούμε πως για διάσταση εμβύθισης , η διάσταση συσχέτισης συγκλίνει στο διάστημα . Το γεγονός ότι η κλίση είναι μη-ακέραιος αριθμός δηλώνει πως έχουμε στα χέρια μας κάποιο παράξενο ελκυστή. Ωστόσο, αξίζει να σημειωθεί πως η υποψήφια διάσταση συσχέτισης του ελκυστή αυτού είναι το άνω όριο της v(m) για m>6 που ισούται με (περίπου) μονάδα. Συνεπώς, **η διάστασης συσχέτισης εκφράζει ευθεία** **όπως μας προϊδεάσει το διάγραμμα** **διασποράς.** Ωστόσο, οι μη-γραμμικότητες του συστήματος και η πιθανή χαοτική του φύση, το καθιστούν κάθε άλλο παρά ιδανικό σύστημα.

1. **Αναζήτηση ΚΝΝ Τοπικού Μοντέλου**

Τέλος, πραγματοποιήθηκε αναζήτηση τοπικού μοντέλου τύπου k-nearest neighbor για αποτελεσματική περιγραφή του συστήματος. Χρησιμοποιήθηκε λόγος δεδομένων εκπαίδευσης – δεδομένων ελέγχου 60%-40% όπως και στα προηγούμενα σημεία της εργασίας.

**Zeroth Order Prediction**

Έγινε χρήση της localpredictnrmse2, για διάφορες διαστάσεις εμβύθισης, και πλήθος Κ-κοντινών γειτόνων . Συγκεκριμένα, οι με την οποία καλέστηκε η συνάρτηση είναι

localpredictnrmse2(y, 0.4\*N, tau, m, Tmax, Κ)

* : Σύνολο δεδομένων
* : το 40% των δεδομένων αξιοποιούνται ως σύνολο ελέγχου.
* : Υστέρηση
* : Διάσταση εμβύθισης: Αναζήτηση από 1 έως 10
* ορίζοντας πρόβλεψης.
* Κ = 1

Το αποδοτικότερο μοντέλο ως προς την παράμετρο κατά NRMSE είναι για , με , το χείριστο από όλα τα μοντέλα που έχουν εξεταστεί μέχρι στιγμής.





**Local Average Prediction**

Με παρόμοιο τρόπο έγινε αξιολόγηση και LAP μοντέλων για K-NN από 2 έως 20, διάσταση εμβύθισης 1 έως 10.

****

Το βέλτιστο μοντέλο βάσει το 60%-40% Cross validation είναι αυτό με και , με .



Το LAP μοντέλο είναι αισθητά πιο αποδοτικό από το απλό 0th order local prediction, και το επιλεγόμενο μοντέλο ως αποτέλεσμα της μη γραμμικής ανάλυσης.

**Τελική Σύγκριση**

Στα πλαίσια του Β μέρους της εργασίας, καταλήξαμε σε 2 μοντέλα: