ΙΟΝΙΟ ΠΑΝΕΠΙΣΤΗΜΙΟ
ΤΜΗΜΑ ΠΛΗΡΟΦΟΡΙΚΗΣ

ΣΤΟΧΑΣΤΙΚΗ ΑΝΑΛΥΣΗ ΔΕΔΟΜΕΝΩΝ

Απαλλακτική Εργασία

Μάρκου Δήμητρα Π2019170

Στοχαστική Διαδικασία Markov

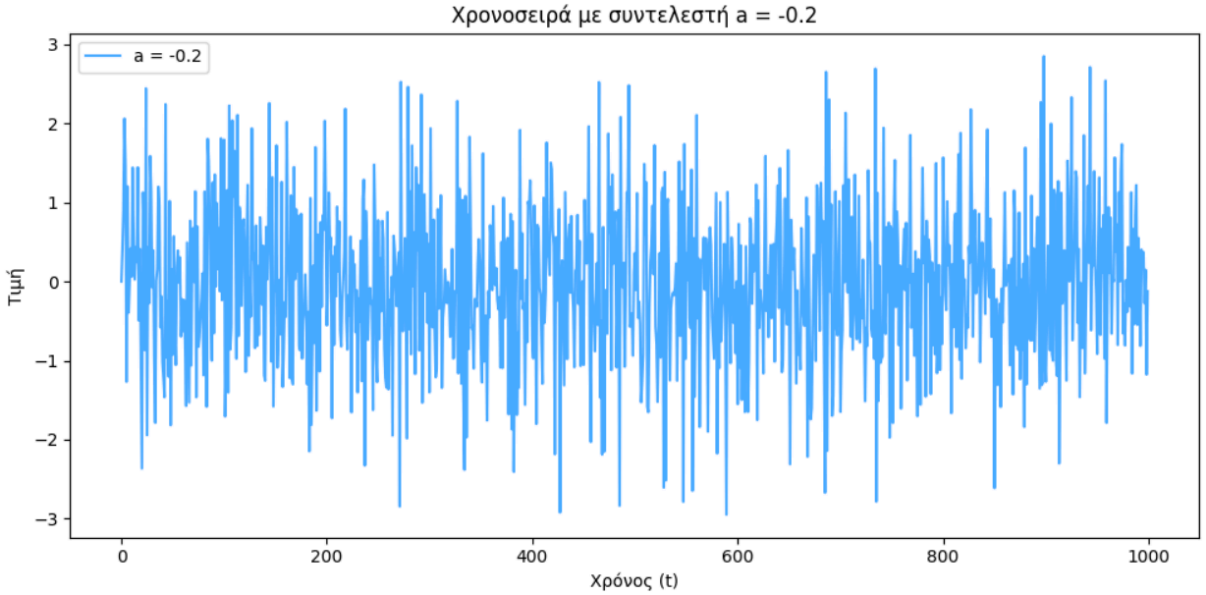
$$X_t = a \cdot X_{t-1} + Z_t$$

Ο παραπάνω ορισμός αντιπροσωπεύει μια αυτοπαλίνδρομη διαδικασία πρώτης τάξης (AR(1)). Υποδηλώνει ότι η τρέχουσα τιμή της σειράς X_t παράγεται πολλαπλασιάζοντας την προηγούμενη τιμή X_{t-1} με τον συντελεστή a και στη συνέχεια προσθέτοντας έναν όρο σφάλματος λευκού θορύβου Z_t. Ο όρος λευκού θορύβου Z_t εισάγει την τυχαιότητα, εξασφαλίζοντας ότι η διαδικασία είναι στοχαστική και όχι ντετερμινιστική.

- X_t είναι η τιμή της χρονοσειράς τη χρονική στιγμή t.
- a είναι ο αυτοπαλίνδρομος συντελεστής, ο οποίος καθορίζει την επίδραση της προηγούμενης τιμής X_{t-1} στην τρέχουσα τιμή X_t.
- Z_t είναι ο όρος λευκού θορύβου τη χρονική στιγμή t, ο οποίος είναι ένας τυχαίος όρος σφάλματος που θεωρείται ότι κατανέμεται κανονικά με μέσο όρο 0 και σταθερή διακύμανση.

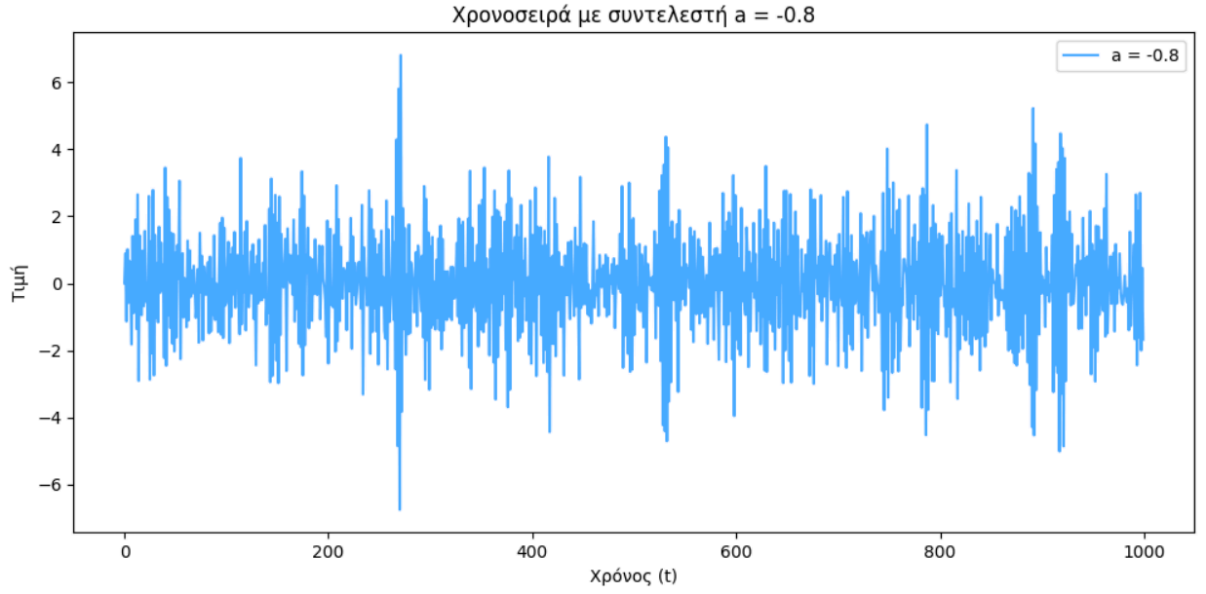
Χρονοσειρά με $a = -0,2$:

Η χρονοσειρά έχει μέτριο αρνητικό αυτοπαλινδρομικό συντελεστή a, που σημαίνει ότι κάθε τιμή επηρεάζεται από περίπου 20% της προηγούμενης τιμής προς την αντίθετη κατεύθυνση, συν κάποιο τυχαίο θόρυβο.



Χρονοσειρά με $a = -0,8$:

Η χρονοσειρά έχει ισχυρότερο αρνητικό αυτοπαλινδρομικό συντελεστή a, που σημαίνει ότι κάθε τιμή επηρεάζεται από περίπου 80% της προηγούμενης τιμής προς την αντίθετη κατεύθυνση, συν κάποιο τυχαίο θόρυβο. Αυτός ο υψηλότερος συντελεστής έχει ως αποτέλεσμα πιο έντονες διακυμάνσεις σε σύγκριση με την πρώτη σειρά.



Στοχαστική Διαδικασία Markov: Κώδικας

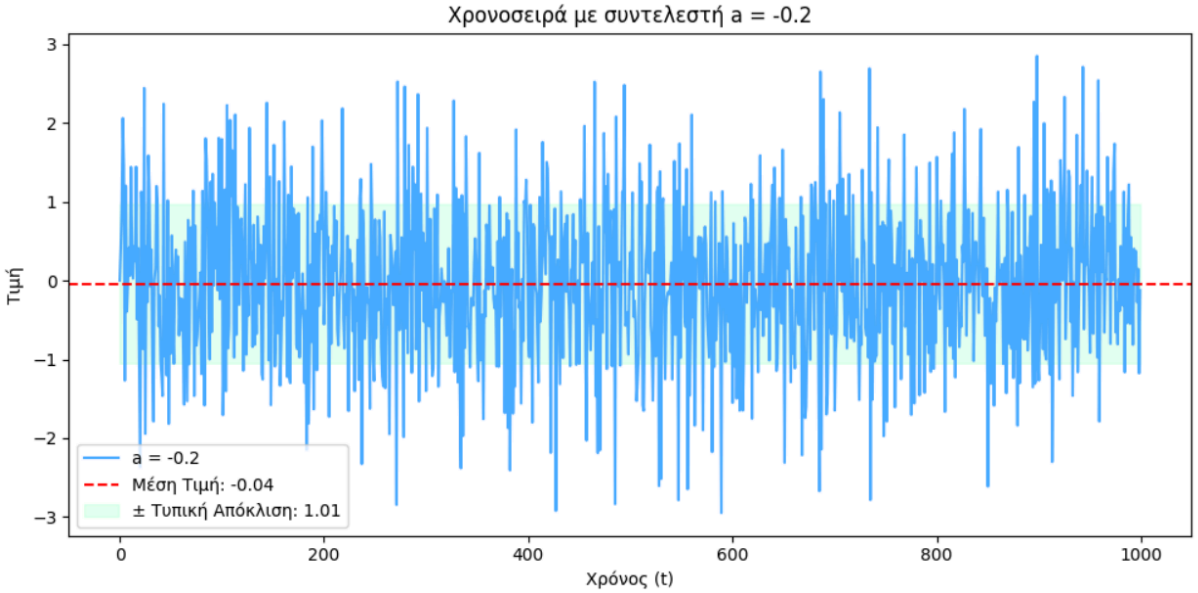
```
import numpy as np

# Παράμετροι
n = 1000  # Μήκος Χρονοσειράς
a1 = -0.2
a2 = -0.8
np.random.seed(0)

# Δημιουργία Λευκού Θορύβου
Zt1 = np.random.normal(0, 1, n)
Zt2 = np.random.normal(0, 1, n)

# Αρχικοποίηση Χρονοσειρών
X1 = np.zeros(n)
X2 = np.zeros(n)

for t in range(1, n):
     X1[t] = a1 * X1[t-1] + Zt1[t]
     X2[t] = a2 * X2[t-1] + Zt2[t]
```

Ο παραπάνω κώδικας αρχικά εισάγει τη βιβλιοθήκη NumPy, έπειτα θέτει τις παραμέτρους $n = 1000$, η οποία δηλώνει το μήκος της χρονοσειράς, καθώς και τους συντελεστές $a1 = -0.2$ και $a2 = -0.8$. Η `np.random.seed(0)` θέτει το seed για τη γεννήτρια τυχαίων αριθμών της NumPy, ώστε να διασφαλίζεται η επαναληψιμότητα των παραγόμενων τυχαίων αριθμών. Για τη δημιουργία λευκού θορύβου χρησιμοποιείται η `np.random.normal(0, 1, n)`. Στη συνέχεια γίνεται αρχικοποίηση των χρονοσειρών με την `np.zeros(n)` για μήκος ν . Ο βρόγχος της επανάληψης πολλαπλασιάζει την προηγούμενη τιμή της κάθε χρονοσειράς με τον συντελεστή και προσθέτει τον λευκό θόρυβο.

Μέση Τιμή, Διακύμανση και Τυπική Απόκλιση

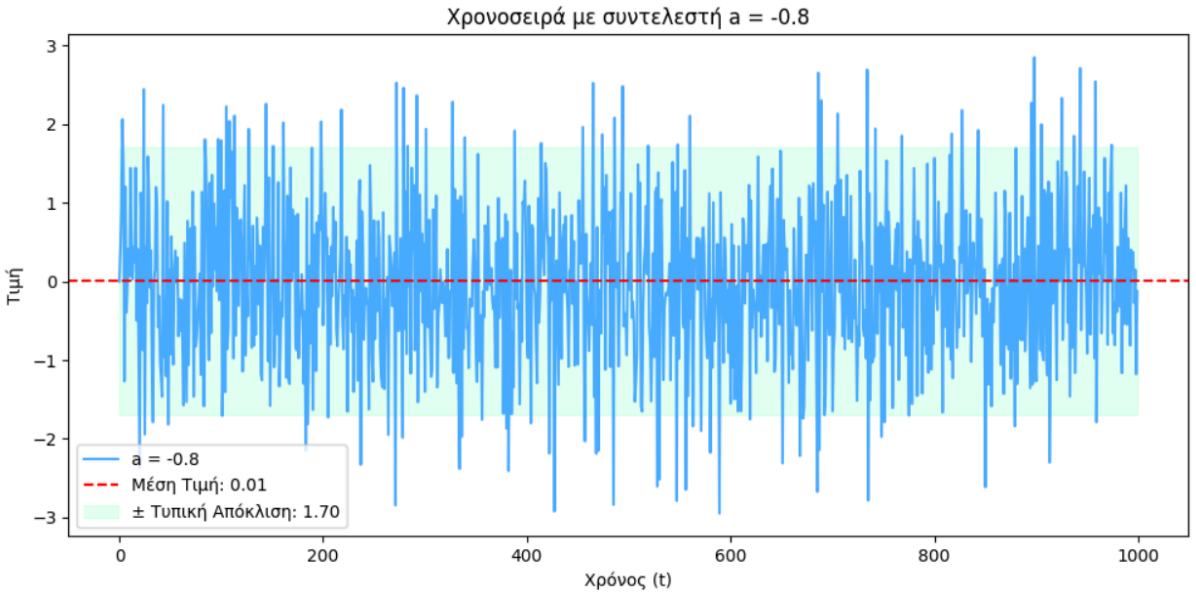
Χρονοσειρά με $a = -0.2$:

- Μέση Τιμή (Mean): -0.039204407048413166
- Διακύμανση (Variance): 1.0226585877909087
- Τυπική Απόκλιση (Standard Deviation): 1.0112658343832786



Χρονοσειρά με $a = -0.8$:

- Μέση Τιμή (Mean): 0.006509721478170432
- Διακύμανση (Variance): 2.8934121695352677
- Τυπική Απόκλιση (Standard Deviation): 1.7010032832229536



Μέση Τιμή, Διακύμανση και Τυπική Απόκλιση: Ερμηνεία

Χρονοσειρά με $a = -0.2$:

- **Μέση Τιμή** (Mean): -0.039204407048413166

Η μέση τιμή της χρονοσειράς είναι κοντά στο μηδέν, αυτό υποδηλώνει ότι κατά μέσο όρο, οι τιμές της χρονοσειράς κυμαίνονται γύρω από το μηδέν. Αυτό είναι αναμενόμενο για μια χρονοσειρά με μικρό αρνητικό συντελεστή, καθώς η επίδραση των προηγούμενων τιμών στην τρέχουσα τιμή είναι σχετικά ασθενής.

- **Διακύμανση** (Variance): 1.0226585877909087

Η διακύμανση μετρά τη διασπορά των τιμών της χρονοσειράς από τη μέση τιμή. Μια διακύμανση γύρω στο 1 υποδηλώνει μέτρια μεταβλητότητα των δεδομένων. Οι τιμές δεν έχουν μεγάλη διασπορά από τη μέση τιμή.

- **Τυπική Απόκλιση** (Standard Deviation): 1.0112658343832786

Η τυπική απόκλιση, υποδηλώνει επίσης μέτρια μεταβλητότητα. Παρέχει ένα μέτρο της μέσης απόστασης των τιμών της χρονοσειράς από τη μέση τιμή. Μια τυπική απόκλιση περίπου 1 είναι συνεπής με την παρατηρούμενη μεταβλητότητα.

Χρονοσειρά με $a = -0.8$:

- **Μέση Τιμή** (Mean): 0.006509721478170432

Η μέση τιμή είναι πολύ κοντά στο μηδέν, όπου υποδηλώνει ότι οι τιμές της χρονοσειράς κυμαίνονται κατά μέσο όρο γύρω από το μηδέν. Αυτό είναι τυπικό για μια χρονοσειρά όπου ο ισχυρός αρνητικός συντελεστής εξασφαλίζει ότι οι μεγάλες αποκλίσεις προς μια κατεύθυνση ακολουθούνται από σημαντικές διορθώσεις προς την αντίθετη κατεύθυνση.

- **Διακύμανση** (Variance): 2.8934121695352677

Η διακύμανση είναι σημαντικά υψηλότερη από εκείνη της πρώτης χρονοσειράς. Αυτό υποδηλώνει υψηλότερο βαθμό μεταβλητότητας στα δεδομένα, το οποίο είναι αναμενόμενο λόγω του ισχυρότερου αρνητικού συντελεστή. Οι τιμές τείνουν να διαφέρουν περισσότερο από τη μέση τιμή σε σύγκριση με την πρώτη σειρά.

- **Τυπική Απόκλιση** (Standard Deviation): 1.7010032832229536

Η τυπική απόκλιση είναι επίσης υψηλότερη, αντανακλώντας την αυξημένη μεταβλητότητα της χρονοσειράς. Μια τυπική απόκλιση της τάξης του 1,7 σημαίνει ότι οι τιμές της σειράς απέχουν κατά μέσο όρο 1,7 μονάδες από τη μέση τιμή, υποδηλώνοντας σημαντικότερες διακυμάνσεις.

Η χρονοσειρά με $a = -0,2$ παρουσιάζει μέτρια μεταβλητότητα γύρω από τη μέση τιμή, με τις τιμές να παραμένουν γενικά πιο κοντά στη μέση τιμή, ενώ η χρονοσειρά με $a = -0,8$ παρουσιάζει μεγαλύτερη μεταβλητότητα, με τιμές πιο απομακρυσμένες από τον μέσο όρο, γεγονός που οδηγεί σε πιο έντονες διακυμάνσεις. Η διαφορά στους συντελεστές a επηρεάζει τον βαθμό εξάρτησης από προηγούμενες τιμές, επηρεάζοντας έτσι τη μεταβλητότητα και τη σταθερότητα της χρονοσειράς. Ένα υψηλότερο μέγεθος του συντελεστή οδηγεί σε μεγαλύτερη μεταβλητότητα και μεγαλύτερες διορθώσεις μετά από αποκλίσεις, όπως φαίνεται στη δεύτερη σειρά.

Μέση Τιμή, Διακύμανση και Τυπική Απόκλιση: Κώδικας

```
# Υπολογισμός Μέσης Τιμής, Διακύμανσης και Τυπικής Απόκλισης

mean_X1 = np.mean(X1)
variance_X1 = np.var(X1)
std_dev_X1 = np.std(X1)
print("Για την χρονοσειρά με συντελεστή a = -0.2:")
print("Μέση Τιμή (Mean):", mean_X1)
print("Διακύμανση (Variance):", variance_X1)
print("Τυπική Απόκλιση (Standard Deviation):", std_dev_X1, "\n")

mean_X2 = np.mean(X2)
variance_X2 = np.var(X2)
std_dev_X2 = np.std(X2)
print("Για την χρονοσειρά με συντελεστή a = -0.8")
print("Μέση Τιμή (Mean):", mean_X2)
print("Διακύμανση (Variance):", variance_X2)
print("Τυπική Απόκλιση (Standard Deviation):", std_dev_X2)
```

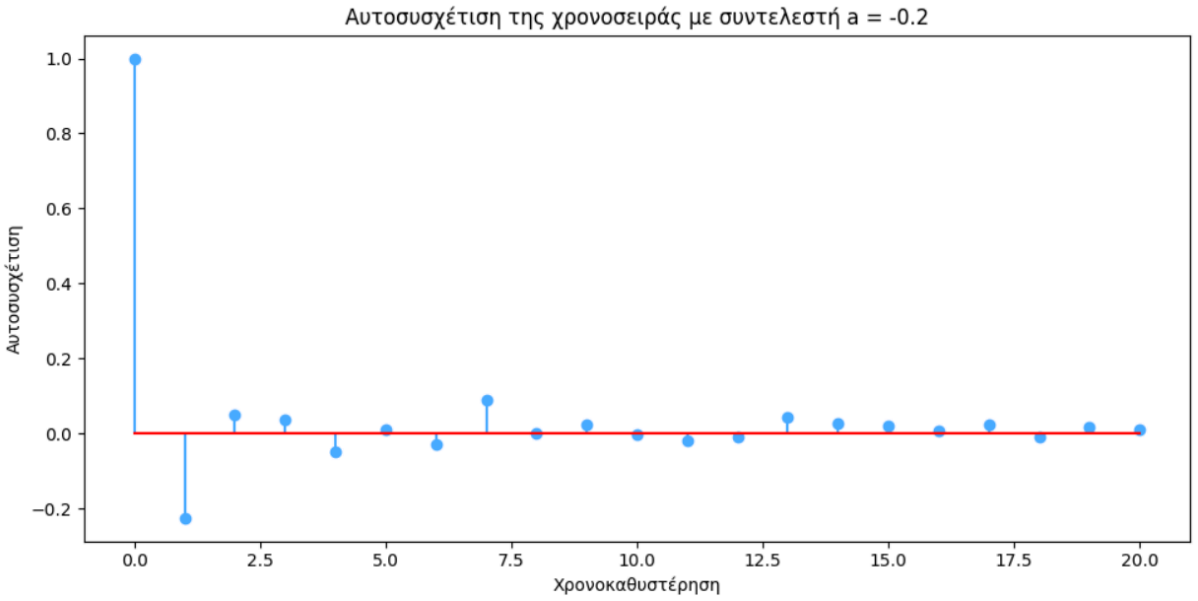
Ο παραπάνω κώδικας υπολογίζει τη μέση τιμή, τη διακύμανση και την τυπική απόκλιση με τη βιβλιοθήκη NumPy.

Αυτοσυσχέτιση - Autocorrelation

Η συνάρτηση αυτοσυσχέτισης μας προσφέρει πληροφορίες σχετικά με τη "μνήμη" μιας χρονοσειράς.

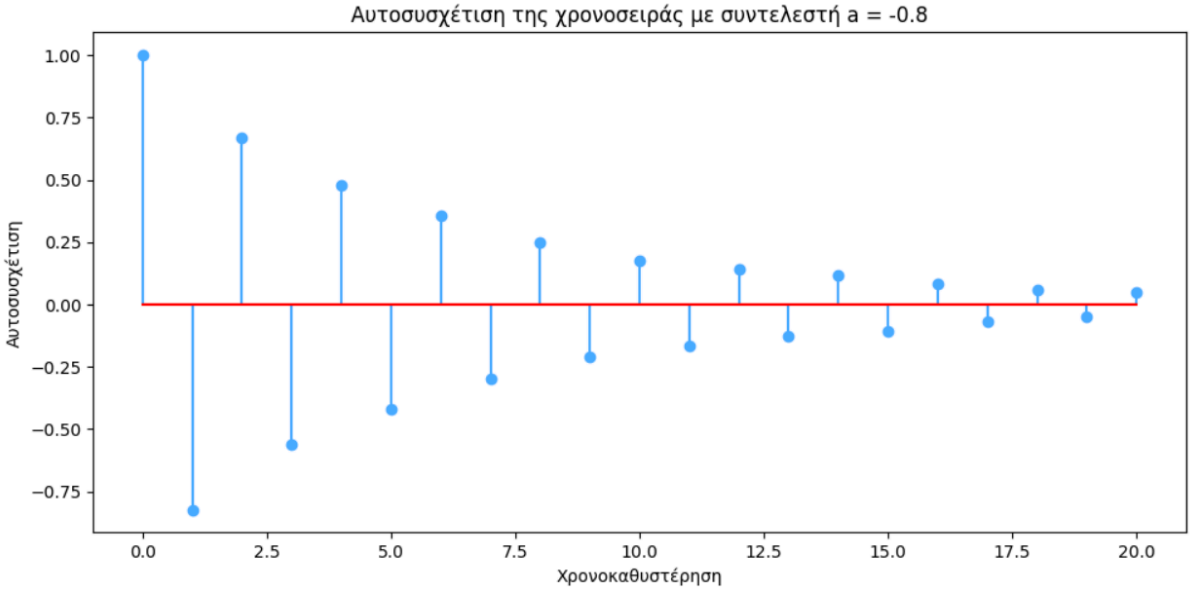
Χρονοσειρά με $a = -0.2$:

Η συνάρτηση αυτοσυσχέτισης παρουσιάζει μια σχετικά γρήγορη εξασθένηση. Στη στιγμή όπου η χρονοκαθυστέρηση ισούται με 1, παρατηρείται μια αρνητική αυτοσυσχέτιση, η οποία είναι αναμενόμενη αφού ο συντελεστής $a = -0.2$ εισάγει μια ήπια αρνητική εξάρτηση. Για τις επόμενες χρονικές καθυστερήσεις οι τιμές αυτοσυσχέτισης θα μειωθούν γρήγορα, πλησιάζοντας το μηδέν με μικρή ταλάντωση γύρω του, υποδεικνύοντας ότι η επιρροή των προηγούμενων τιμών μειώνεται γρήγορα.



Χρονοσειρά με $a = -0.8$:

Η συνάρτηση αυτοσυσχέτισης παρουσιάζει βραδύτερη εξασθένηση, αντανακλώντας ισχυρότερη εξάρτηση από παρελθοντικές τιμές. Στη στιγμή όπου η χρονοκαθυστέρηση ισούται με 1, παρατηρούμε μεγάλη αρνητική αυτοσυσχέτιση, λόγω της ισχυρής αρνητικής ανατροφοδότησης που εισάγεται από τον συντελεστή $a = -0,8$. Για τις επόμενες χρονικές καθυστερήσεις οι τιμές αυτοσυσχέτισης θα παρουσιάζουν ένα ταλαντευόμενο διάγραμμα γύρω από το μηδέν, προσεγγίζοντας το σταδιακά, αντανακλώντας την εναλλασσόμενη επιρροή των παρελθουσών αλλά θα φθίνει πιο αργά σε σύγκριση με την πρώτη χρονοσειρά.



Αυτοσυσχέτιση - Autocorrelation: Ερμηνεία

Γενικότερα, οι προβλέψεις που βασίζονται σε πρόσφατες τιμές είναι πιο ακριβείς, αλλά η χρονοσειρά είναι πιο επιρρεπής σε τυχαίες διακυμάνσεις, ενώ οι χρονοσειρές με μεγάλη μνήμη μπορεί να είναι πιο προβλέψιμες για μεγαλύτερους ορίζοντες, αλλά οι ισχυρές αρνητικές συσχετίσεις μπορεί να οδηγήσουν σε πιο έντονες ταλαντώσεις.

Η χρονοσειρά με συντελεστή $a = -0.2$ παρουσιάζει ραγδαία φθίνουσα ροή, υποδεικνύοντας ότι η επιρροή των προηγούμενων τιμών μειώνεται γρήγορα. Αυτό υποδηλώνει ότι η χρονοσειρά έχει βραχεία μνήμη, πράγμα που σημαίνει ότι η τρέχουσα τιμή εξαρτάται μόνο ασθενώς από τις τιμές του μακρινού παρελθόντος.

Η χρονοσειρά με συντελεστή $a = -0.8$ παρουσιάζει πιο αργή πτώση σε σύγκριση με την προηγούμενη χρονοσειρά, υποδεικνύοντας ότι η επιρροή των προηγούμενων τιμών επιμένει περισσότερο. Αυτό υποδηλώνει ότι η χρονοσειρά έχει μεγαλύτερη μνήμη, δηλαδή ότι η τρέχουσα τιμή εξαρτάται περισσότερο από τις τιμές του παρελθόντος.

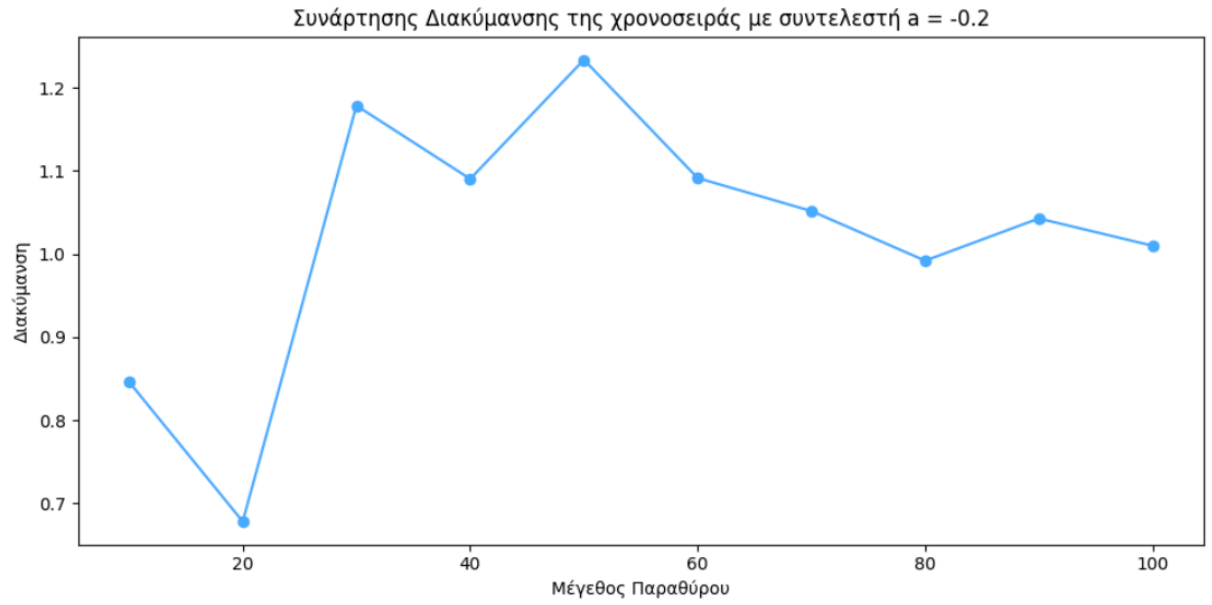
Αυτοσυσχέτιση - Autocorrelation: Κώδικας

```
# Υπολογισμός αυτοσυχέτισης - autocorrelation
acf_X1 = sm.tsa.acf(X1, nlags=20)
acf_X2 = sm.tsa.acf(X2, nlags=20)
```

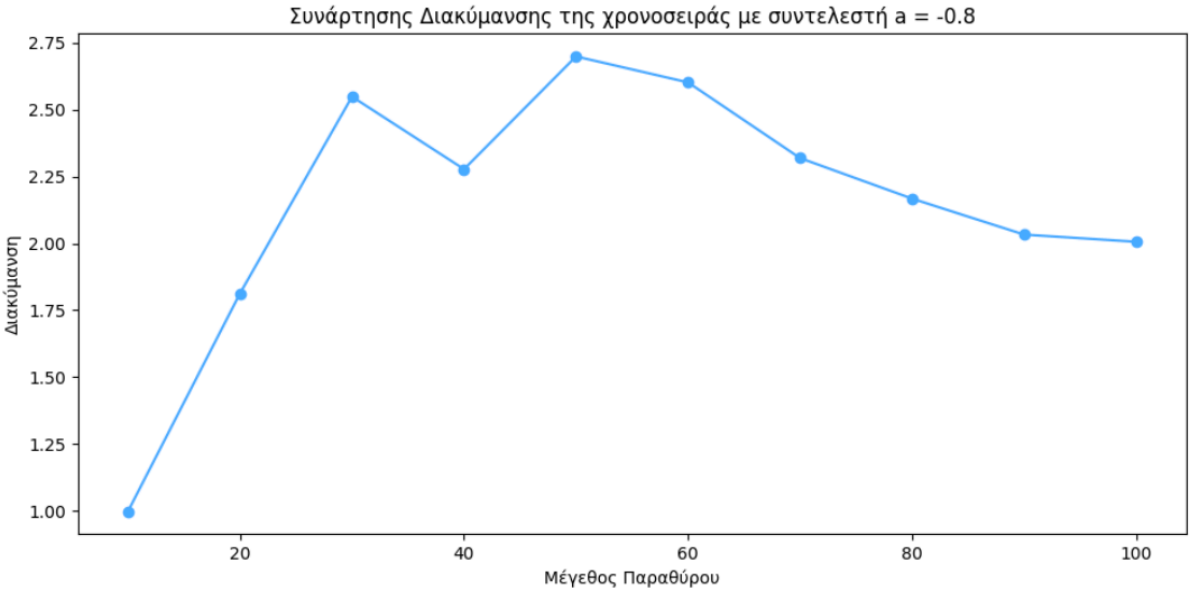
Για τον υπολογισμό της αυτοσυσχέτισης αξιοποιήθηκε η βιβλιοθήκη `statsmodels.api`. Προσπάθησα να αξιοποιήσω τη συνάρτηση `curve_fit` από τη βιβλιοθήκη `scipy.optimize`, δυστυχώς χωρίς επιτυχία.

Συνάρτηση Διακύμανσης

Χρονοσειρά με $a = -0.2$:



Χρονοσειρά με $a = -0.8$:



Συνάρτηση Διακύμανσης: Ερμηνεία

Η συνάρτηση διακύμανσης μετρά πώς διαμορφώνεται η μεταβλητότητα της χρονοσειράς ως συνάρτηση του μεγέθους του παραθύρου. Εξετάζοντας τη διακύμανση σε διαφορετικά μεγέθη παραθύρου, κατανοούμε πώς συμπεριφέρεται η χρονοσειρά σε μικρές και μεγάλες περιόδους.

Για μια χρονοσειρά με βραχεία μνήμη η διακύμανση παραμένει σχετικά σταθερή με την αύξηση του μεγέθους του παραθύρου. Αυτό υποδηλώνει ότι η επιρροή των προηγούμενων τιμών μειώνεται γρήγορα και η χρονοσειρά συμπεριφέρεται περισσότερο σαν λευκός θόρυβος. Για τη χρονοσειρά με $a = -0,2$ η διακύμανση δεν αυξάνεται σημαντικά με το μέγεθος του παραθύρου, υποδεικνύοντας μικρότερη μνήμη και λιγότερο επίμονη μεταβλητότητα.

Όταν η χρονοσειρά έχει μεγάλη μνήμη, η διακύμανση αυξάνεται με το μέγεθος του παραθύρου. Αυτό οφείλεται στο γεγονός ότι η σωρευτική επίδραση των προηγούμενων τιμών συνεχίζει να επηρεάζει τη σειρά καθώς αυξάνεται το μέγεθος του παραθύρου. Για τη χρονοσειρά με συντελεστή $a = -0,8$, η διακύμανση αυξάνεται πιο αισθητά με το μέγεθος του παραθύρου, υποδηλώνοντας μεγαλύτερη μνήμη και πιο επίμονη μεταβλητότητα.

Συνάρτηση Διακύμανσης: Κώδικας

```
# Υπολογισμός Συνάρτησης Διακύμανσης
window_sizes = np.arange(10, 101, 10)
rolling_variance_X1 = [np.var(X1[:w]) for w in window_sizes]
rolling_variance_X2 = [np.var(X2[:w]) for w in window_sizes]
```

Ο παραπάνω κώδικας αξιοποιεί τη συνάρτηση `np.arange(10, 101, 10)` της NumPy βιβλιοθήκης. Η συνάρτηση αυτή παράγει έναν πίνακα μεγεθών παραθύρων που ξεκινούν από το 10 έως το 100 (συμπεριλαμβανομένου), με βήμα 10.

Συμπεράσματα Εργασίας

Η εργασία αυτή έγινε ανάλυση σε δύο στοχαστικές διαδικασίες που μοντελοποιούνται από αυτοπαλίνδρομες χρονοσειρές με διαφορετικούς συντελεστές, $a1 = -0,2$ και $a2 = -0,8$. Μέσω της εξέτασης στατιστικών μετρικών, συναρτήσεων αυτοσυσχέτισης και αναλύσεων συναρτησης διακύμανσης, αποδείξαμε πώς ο αυτοπαλίνδρομος συντελεστής επηρεάζει σημαντικά τη συμπεριφορά και τα χαρακτηριστικά μιας χρονοσειράς. Ένας μικρότερος συντελεστής οδηγεί σε ήπιες διακυμάνσεις και χαμηλότερη μεταβλητότητα, υποδηλώνοντας βραχεία μνήμη, ενώ ένας μεγαλύτερος συντελεστής παράγει πιο έντονες διακυμάνσεις και υψηλότερη μεταβλητότητα, αντανακλώντας μεγαλύτερη μνήμη και μεγαλύτερη επιμονή. Τα ευρήματα αυτά υπογραμμίζουν τη σημασία της κατανόησης των θεμελιωδών παραμέτρων των στοχαστικών διαδικασιών, οι οποίες μπορούν να έχουν βαθιές επιπτώσεις σε διάφορες εφαρμογές του πραγματικού κόσμου, συμπεριλαμβανομένων των χρηματοπιστωτικών αγορών, της πρόβλεψης του καιρού και της οικονομικής μοντελοποίησης.

Η εργασία αυτή μου προσέφερε πολύτιμες γνώσεις στον τομέα της ανάλυσης στοχαστικών διαδικασιών και μου ανοίγει το δρόμο για μελλοντική έρευνα και ενασχόληση με τον τομέα της στατιστικής σε ακαδημαϊκό καθώς και σε επαγγελματικό επίπεδο.

Ο κώδικας της εργασίας αυτής έχει γραφτεί στη γλώσσα προγραμματισμού Python, και στο περιβάλλον του kaggle όπου ο κώδικας είναι διαθέσιμος για προβολή.

Οι βιβλιοθήκες που χρησιμοποιήθηκαν είναι οι εξής:

```
import numpy as np
import pandas as pd
import matplotlib.pyplot as plt
import statsmodels.api as sm
```

Link Απαλλακτικής Εργασίας στο kaggle:
https://www.kaggle.com/code/markedd/stohastic-data-analysis-semester-project