ΠΑΝΕΠΙΣΤΗΜΙΟ ΠΕΙΡΑΙΩΣ

**ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ**

**ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ**

ΑΝΑΛΥΣΗ ΚΑΤΗΓΟΡΙΚΩΝ ΧΡΟΝΟΣΕΙΡΩΝ

Μανώλης Π. Δρυμώνης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιανουάριος 2005

Π Α Ν Ε Π Ι Σ Τ Η Μ Ι Ο Π Ε Ι Ρ Α Ι Ω Σ

ΤΜΗΜΑ ΣΤΑΤΙΣΤΙΚΗΣ
ΚΑΙ ΑΣΦΑΛΙΣΤΙΚΗΣ ΕΠΙΣΤΗΜΗΣ

ΜΕΤΑΠΤΥΧΙΑΚΟ ΠΡΟΓΡΑΜΜΑ
ΣΠΟΥΔΩΝ
ΣΤΗΝ ΕΦΑΡΜΟΣΜΕΝΗ ΣΤΑΤΙΣΤΙΚΗ

ΑΝΑΛΥΣΗ ΚΑΤΗΓΟΡΙΚΩΝ ΧΡΟΝΟΣΕΡΩΝ

Μανώλης Π. Δρυμώνης

Διπλωματική Εργασία

που υποβλήθηκε στο Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς ως μέρος των απαιτήσεων για την απόκτηση του Μεταπτυχιακού Διπλώματος Ειδίκευσης στην Εφαρμοσμένη Στατιστική

Πειραιάς
Ιανουάριος 2005

Η παρούσα Διπλωματική Εργασία εγκρίθηκε ομόφωνα από την Τριμελή Εξεταστική Επιτροπή που ορίσθηκε από τη ΓΣΕΣ του Τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς στην υπ' αριθμ. …….. συνεδρίασή του σύμφωνα με τον Εσωτερικό Κανονισμό Λειτουργίας του Προγράμματος Μεταπτυχιακών Σπουδών στην Εφαρμοσμένη Στατιστική

Τα μέλη της Επιτροπής ήταν:

- …….. (Επιβλέπων)
- ……..
- ……..

Η έγκριση της Διπλωματική Εργασίας από το Τμήμα Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς δεν υποδηλώνει αποδοχή των γνωμών του συγγραφέα.

UNIVERSITY OF PIRAEUS

DEPARTMENT OF STATISTICS
AND INSURANCE SCIENCE

POSTGRADUATE PROGRAM IN
APPLIED STATISTICS

ANALYSIS OF CATEGORICAL TIME SERIES

By

Manolis P. Drimonis

MSc Dissertation

submitted to the Department of Statistics and Insurance Science of the University of Piraeus in partial fulfilment of the requirements for the degree of Master of Science in Applied Statistics

Piraeus, Greece
January 2005

Στους γονείς μου
Παναγιώτη και Καλλιρρόη
και στα αδέρφια μου
Φωτεινή και Θοδωρή

ΕΥΧΑΡΙΣΤΙΕΣ

Στο σημείο αυτό θα ήθελα να εκφράσω τις θερμές μου ευχαριστιές στην κα. Κατέρη Μαρία, επ. Καθηγήτρια του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για την πολύ σημαντική καθοδήγηση, υποστήριξη και βοήθεια καθ' όλη την διάρκεια της παρούσας εργασίας.

Παράλληλα θέλω να ευχαριστήσω τον κ. Ηλιόπουλο Γεώργιο, επ. Καθηγητή του τμήματος Στατιστικής και Ασφαλιστικής Επιστήμης του Πανεπιστημίου Πειραιώς, για την απλόχερη προσφορά του στην μορφοποίηση της εργασίας, μέσω του *Latex*. Συνάμα ιδιαίτερες ευχαριστίες οφείλω στον κ. Μπαρτζώκα Άρη, αν. Καθηγητή του Πανεπιστημίου Ιωαννίνων για την διάθεση των δεδομένων αλλά και τις εύστοχες «μετεωρολογικές συμβουλές». Επίσης θα ήθελα να ευχαριστίσω και τον κ. Πιττή Νικήτα Καθηγητή του Χρηματοοικονομικού τμήματος του Πανεπιστημίου Πειραιώς, για την συμμετοχή του στην τριμελή εξεταστική επιτροπή.

Τέλος θέλω να ευχαριστήσω όλους τους φίλους που με βοήθησαν, ο καθένας με τον τρόπο του, τους τελευταίους μήνες.

ΔΡΥΜΩΝΗΣ ΜΑΝΩΛΗΣ

Ν.ΙΩΝΙΑ

Ιανουάριος 2005

ΠΕΡΙΛΗΨΗ

Οι κατηγορικές ή ποιοτικές χρονοσειρές αποτελούν ειδική περίπτωση χρονοσειρών που συχνά συναντάμε σε ποίκιλες εφαρμογές. Όπως και στην «κλασική» θεωρία των χρονοσειρών αντιμετωπίζουμε τα ίδια προβλήματα μοντελοποίησης, εκτίμησης, ελέγχων υποθέσεων, ελέγχων καλής προσαρμογής και προβλέψεων.

Στην παρούσα εργασία θα δείξουμε πώς μπορούν να αντιμετωπιστούν τα δεδομένα προβλήματα μέσω της θεωρίας παλινδρόμησης για κατηγορικές χρονοσειρές, η οποία βασίζεται στα γνωστά αποτελέσματα των γενικευμένων γραμμικών μοντέλων καθώς και στην συμπερασματολογία της μεθόδου της μερικής πιθανοφάνειας (*partial likelihood*). Η δεδομένη προσέγγιση κρίνεται ιδιαίτερα ελκυστική αφού δεν απαιτεί την Μαρκοβιανή υπόθεση και την έννοια της στασιμότητας. Παράλληλα τα μοντέλα παλινδρόμησης για κατηγορικές χρονοσειρές παρέχουν την δυνατότητα «οικονομικής» μοντελοποίησης (*parsimonious modelling*) επιτρέποντας την παρουσία τυχαίων χρονοεξαρτώμενων συμμεταβλητών.

Η ανάλυσή μας ξεκινά απο την μελέτη των δίτιμων χρονοσειρών που αποτελούν ειδική περίπτωση των κατηγορικών σειρών. Εν συνεχεία παρουσιάζουμε τις ποιοτικές σειρές δίδοντας αρκετά ασυμπτωτικά αποτελέσματα. Οι ονομαστικές και οι διατάξιμες κατηγορικές χρονοσειρές αναλύονται διεξοδικά σε ξεχωριστό κεφάλαιο. Για την κατανόηση των θεωρητικών αποτελεσμάτων παρουσιάζουμε την προσπάθεια μοντελοποίησης της δίτιμης χρονοσειράς των βροχών του νομού Ιωαννίνων, μέσω του στατιστικού προγράμματος *S-PLUS*.

Τα τελευταία 10 χρόνια έχουν αναπτυχθεί ευέλικτες εναλλακτικές τεχνικές για την στατιστική συμπερασματολογία των ποιοτικών χρονοσειρών, οι οποίες τείνουν να κυριαρχήσουν έναντι των μοντέλων παλινδρόμησης που στηρίζονται στην μερική πιθανοφάνεια. Αν και οι σύγχρονες τάσεις δεν αποτελούν αντικείμενο της δεδομένης εργασίας, παρουσιάζονται κάποιες απο αυτές περιληπτικά σε ξεχωριστή ενότητα.

Τέλος στα παραρτήματα υπάρχουν χρήσιμες πληροφορίες που θα βοηθήσουν τον αναγνώστη να κατανοήσει καλύτερα το θέμα.

ABSTRACT

Categorical-or qualitative-time series are special case of time series which are frequently encountered in diverse applications. As with 'ordinary' time series we faced with the problems of modeling, estimation, model checking, diagnostics and prediction.

At the present project we will show how these problems can faced through the regression theory for categorical time series, whose foundation is based on generalized linear models and partial likelihood inference. This approach can be considered as very attractive since neither the Markov property nor stationarity are assumed. Furthermore, regression methods for categorical time series allow for parsimonious modeling and incorporation of random time-dependent covariates.

Our discussion begins with the analysis of binary time series which consist a special case of categorical time series. After that, we present qualitative time series including some asymptotic results. Nominal and ordinal time series will be considered extensively in a separate chapter. For understanding the theoretical results we present an example in which we attempt modeling the rainfall data from Ioannina.

The last 10 years, a lot of flexible alternative techniques have been evoluted for the statistical inference of categorical time series. These approaches tend to dominate over the regression models which are based on the theory of partial likelihood. Some of the new approaches are mentioned in a seperate chapter.

Finally the appendices include a lot of useful informations which will help the reader to understand the underlying issues.

Περιεχόμενα

Κεφάλαιο 1

Εισαγωγικό κεφάλαιο

1.1 Πρόλογος

Σκοπός αυτής της εργασίας είναι η στατιστική ανάλυση των κατηγορικών χρονοσειρών (*Categorical Time Series*) μέσω της μεθοδολογίας των Γενικευμένων Γραμμικών Μοντέλων (*Generalized Linear Models-GLM*). Οι κατηγορικές χρονοσειρές αποτελούν ένα είδος χρονολογικής σειράς που συχνά συναντάμε στην πράξη. Πρίν προχωρήσουμε στην ανάλυσή τους θα παρουσιάσουμε γενικά στοιχεία της θεωρίας των χρονοσειρών τα οποία θα συντελέσουν στην καλύτερη κατανόηση του δεδομένου στατιστικού πεδίου.

1.2 Γενικά περί Χρονολογικών Σειρών

Ξεκινώντας την ανάλυση των χρονοσειρών (*time series*) θα παραθέσουμε δυο αντιπροσωπευτικούς ορισμούς του συγκεκριμένου κλάδου της Στατιστικής.

Ορισμός 1.2.1 *(Ξενάκης, (1988)) Οι χρονολογικές σειρές αποτελούν την καταγραφή της εξέλιξης ενός φαινομένου, το οποίο επαναλαμβάνεται κατά διαδοχικές χρονικές στιγμές και του οποίου η έκβαση σε κάθε χρόνο διαμορφώνεται από ένα μηχανισμό τύχης.*

Ορισμός 1.2.2 *(Priestley, (1981)) Χρονοσειρά είναι η καταγραφή των τιμών οποιασδήποτε κυμαινόμενης ποσότητας σε διαφορετικά χρονικά σημεία.*

Οι παραπάνω ορισμοί διατυπωμένοι σε αυστηρότερα μαθηματικά πλαίσια μας οδηγούν στους ακόλουθους δυο ορισμούς της χρονοσειράς που δηλώνουν την άμεση σχέση

της με τις αλληλένδετες έννοιες της στοχαστικής διαδικασίας (*stochastic process*) και της δειγματοληπτικής διαδρομής (*sample path*).

Ορισμός 1.2.3 *Μία οικογένεια των τυχαίων μεταβλητών* $\{U_t\}$, $t = 1, 2, \ldots$, *που ορίζεται στον πιθανοθεωρητικό χώρο* $(\Omega, \mathcal{F}, P)$, *όπου* Ω *είναι ο δειγματικός χώρος,* $\mathcal{F}$ *είναι η σ-άλγεβρα και* P *είναι η συνάρτηση πιθανότητας, ονομάζεται στοχαστική διαδικασία. Αν η παράμετρος* t *εκφράζει τον χρόνο τότε η στοχαστική διαδικασία λέγεται χρονολογική σειρά.*

Ορισμός 1.2.4 *(Ξενάκης, (1988)) Χρονοσειρά είναι μια δειγματοληπτική διαδρομή (*sample path*) ή πραγματοποίηση* (realization) *μιας στοχαστικής ανέλιξης* $\{Y_t\}, t = 1, 2, \ldots$ *της οποίας η παράμετρος* t *εκφράζει τον χρόνο. Είναι δηλαδή μια απλή παρατήρηση από μια πολυμεταβλητή κατανομή πιθανότητας.*

1.2.1 Είδη Χρονολογικών Χρονοσειρών

Παραδείγματα χρονοσειρών συναντάμε σε πολλά επιστημονικά πεδία. Το γεγονός αυτό δείχνει την πολυπλοκότητα καθώς και την ευρύτητα του δεδομένου στατιστικού κλάδου. Χαρακτηριστικά είδη λοιπόν σειρών, πέρα των κατηγορικών, είναι

1. *Οικονομικές χρονοσειρές.* Η καταγραφή των τιμών του πληθωρισμού και των επιτοκίων ανά μήνα ή ανά χρόνο αποτελούν παραδείγματα χρονοσειρών που αναφέρονται σε οικονομικά φαινόμενα. Γενικά η ιστορία της οικονομίας μιας χώρας είναι καταγεγραμμένη στην μορφή χρονολογικών σειρών.
2. *Χρονοσειρές φυσικών φαινομένων.* Οι μετρήσεις του επιπέδου του νερού σε μια λίμνη ανά χρόνο, της θερμοκρασίας ή της ατμοσφαιρικής πίεσης ανά ημέρα ή ώρα συνιστούν παραδείγματα τέτοιων σειρών.
3. *Δημογραφικές Χρονοσειρές.* Οι δεδομένες σειρές ασχολούνται με την μελέτη πληθυσμών. Παράδειγμα αυτού του είδους σειράς ήταν η καταγραφή του πληθυσμού της Αγγλίας και της Ουαλλίας ανά χρόνο. Οι δημογράφοι ήθελαν μέσω της χρονοσειράς που παρατήρησαν να εξετάσουν μελλοντικές αλλαγές στον πλυθυσμό καθώς και τους παράγοντες που ευθύνονται για αυτές (*Chatfield* (1996)).
4. *Χρονολογικές σειρές στον ποιοτικό έλεγχο.* Στον ποιοτικό έλεγχο το κύριο ζητούμενο είναι να ανιχνευθούν αλλαγές στην λειτουργία μιας παραγωγικής διαδικασίας μετρώντας την τιμή μιας μεταβλητής που δείχνει την ποιότητα της διεργασίας. Το γράφημα αυτών των μετρήσεων ως προς τον χρόνο συνιστά χρονοσειρά. Όταν πα-

ρατηρήσουμε ότι οι μετρήσεις απομακρύνονται από μια επιθυμητή τιμή (*target value*) απαιτείται να προβούμε στις απαραίτητες διορθωτικές ενέργειες.

Παραδείγματα χρονοσειρών συναντάμε και σε άλλα ευρενητικά πεδία όπως είναι η βιολογία και η κοινωνιολογία. Αξίζει να σημειώσουμε ότι οι μέθοδοι ανάλυσης χρονοσειρών διαφέρουν ανάλογα με την ιδιομορφία του επιστημονικού κλάδου στον οποίο αναφέρονται. Έτσι για παράδειγμα οι μετεωρολογικές χρονοσειρές συνηθίζεται να διερευνώνται μέσω της φασματικής ανάλυσης ενώ οι οικονομικές μέσω μεθόδων του πεδίου του χρόνου (*time domain*).

1.2.2 Συνιστώσες Χρονοσειράς

Οι τιμές μιας χρονοσειράς διαμορφώνονται από επιμέρους παράγοντες, που καλούνται συνιστώσες της σειράς αυτής, και χαρακτηρίζονται από μια συστηματική ή μη συστηματική συμπεριφορά. Ο διαχωρισμός (*decomposition*) μιας πραγματοποίησης μπορεί να γίνει ως προς τις ακόλουθες συνιστώσες.
1. Την *τάση (trend-T_t)*. Είναι η συνιστώσα που εκφράζει την ανοδική ή καθοδική πορεία της σειράς.
2. Την *εποχικότητα (seasonality-S_t)*. Είναι η συνιστώσα που εκφράζει την κυκλική κύμανση μιας χρονολογικής σειράς με περίοδο ανά έτος.
3. Την *κυκλική συνιστώσα (cyclical component-C_t)*. Εκφράζει την κυκλική κύμανση με περίοδο μεγαλύτερη του ενός έτους.
4. Την *άρρυθμη συνιστώσα (irregular component-I_t)*. Η συνιστώσα αυτή ενσωματώνει όλους τους μη συστηματικούς παράγοντες.
Οι τρεις πρώτες συνιστώσες χαραχτηρίζονται ως συστηματικές. Η σύνθεση μιας χρονοσειράς από τις συνιστώσες της μπορεί να γίνει μέσω του προσθετικού υποδείγματος

$$Y_t = T_t + S_t + C_t + I_t,$$

ή μέσω του πολλαπλασιαστικού υποδείγματος

$$Y_t = T_t.S_t.C_t.I_t.$$

1.2.3 Ορολογία.

Μια χρονοσειρά ονομάζεται *συνεχής* όταν οι παρατηρήσεις πραγματοποιούνται συνεχώς στο χρόνο. Ο όρος «συνεχής» χρησιμοποιείται για σειρές αυτού του είδους ακόμη

και αν η μεταβλητή που μετράμε λαμβάνει διακριτές τιμές. Ακόμη μια σειρά λέγεται *διακριτή* όταν οι παρατηρήσεις λαμβάνουν τιμές σε συγκεκριμένες χρονικές στιγμές που συνήθως ισαπέχουν. Ο όρος «διακριτή» χρησιμοποιείται για τις δεδομένες σειρές ακόμη και αν η μεταβλητή ενδιαφέροντος παίρνει τιμές σε διάστημα.

1.2.4 Στόχοι ανάλυσης χρονοσειρών

Υπάρχουν διάφορα αντικείμενα ανάλυσης χρονοσειρών. από αυτά τα χαρακτηριστικότερα είναι

1. *Περιγραφή.* Καταγράφοντας τις τιμές που χαρακτηρίζουν την έκβαση ενός φαινομένου για διαδοχικούς χρόνους και δημιουργώντας το αντίστοιχο γράφημα λαμβάνουμε μια πρώτη εικόνα του συστήματος που μελετάμε. Η γραφική παράσταση της δειγματοληπτικής διαδρομής ως προς τον χρόνο προσφέρει μια άμεση οπτική επιθεώρηση, που βοηθάει να δούμε αν υπάρχουν χαρακτηριστικά όπως η τάση και η περιοδικότητα. Πρέπει να τονίσουμε ότι το γράφημα της σειράς καθώς και οι ενδείξεις που αυτό παρέχει δεν αποτελούν μέρος της στατιστικής συμπερασματολογίας και για αυτό με σκοπό να προβούμε σε συμπεράσματα απαιτείται, ιδιαίτερα σε σύνθετες περιπτώσεις, να είμαστε αρκετά προσεκτικοί.

2. *Ερμηνεία τυχαιότητας.*

Συχνά συναντάμε διαχρονικά φαινόμενα των οποίων η τυχαιότητα είναι δύσκολο να εξηγηθεί. Το γράφημα της δειγματοληπτικής διαδρομής αυτών των συστημάτων παρουσιάζουν την εικόνα της «παραπλανητικής παλινδρόμησης» (*"spurious regression"*). Συγκεκριμένα παρατηρείται η απουσία ενός σταθερού μέσου και οι στοχαστικοί κύκλοι συμβαίνουν παντού καθιστώντας τα δεδομένα μας ακατάλληλα για στατιστική επεξεργασία. Με σκοπό να μειώσουμε την αβεβαιότητά μας για το φαινόμενο λαμβάνουμε για κάθε χρόνο και την τιμή μιας άλλης μεταβλητής που πιστεύουμε ότι σχετίζεται σε μεγάλο βαθμό με την αρχική χρονοσειρά. Αν πράγματι συμβαίνει κάτι τέτοιο τότε οι δειγματοληπτικές διαδρομές των δυο σειρών θα είναι πολύ κοντά η μια στην άλλη και θα είναι παράλληλες με αποτέλεσμα η μεταβλητότητα της νέας σειράς (που προφανώς πρέπει να είναι γνωστή) να βοηθάει στην ερμηνεία της μεταβλητότητας της πρώτης.

3. *Πρόβλεψη (Prediction).*

Κυρίαρχος σκοπός της στατιστικής ανάλυσης χρονοσειρών είναι δοθείσης μιας δειγματοληπτικής διαδρομής ενός φαινομένου, να προβλέψουμε με όσο το δυνατόν μεγαλύτερη ακρίβεια τις μελλοντικές τιμές του. Χαρακτηριστικό παράδειγμα πρόβλεψης είναι το πρόβλημα του ποιοτικού ελέγχου όπου θέλουμε να δούμε αν μελλοντικά η παραγω-

γική διεργασία θα τεθεί «εκτός ελέγχου» ή θα παραμείνει «εντός ελέγχου» ώστε να λάβουμε τα κατάλληλα μέτρα. Το θέμα της πρόβλεψης πρωτομελετήθηκε από τους *Kolmogorov* (1941) και *Wiener* (1949).

1.3 Γιατί και πώς αναπτύχθηκαν οι Χρονοσειρές

Η θεωρία των χρονοσειρών αναπτύχθηκε από την ανάγκη να μελετηθούν διαχρονικά στοχαστικά φαινόμενα των οποίων η εξέλιξη εξαρτάται από τις παρελθούσες καταστάσεις τους. Στο σημείο αυτό τίθενται σημαντικά ερωτήματα όπώς τι είναι στοχαστικό φαινόμενο και τι διαχρονική εξάρτηση, τα οποία προκειμένου να απαντηθούν απαιτείται να ανατρέξουμε αρκετά πίσω στην ιστορία της επιστήμης.
Οι επιστήμονες από πολύ παλιά ασχολούνται με την μελέτη Δυναμικών Συστημάτων (*Dynamics Systems*), δηλαδή φαινομένων που εξελίσσονται μέσα στον χρόνο. Για παράδειγμα, το ηλιακό μας σύστημα αποτελεί χαρακτηριστικό δυναμικό σύστημα αφού η θέση των πλανητών δεν είναι στατική αλλα μεταβάλλεται διαρκώς στο πεδίο του χρόνου. Τα διαχρονικά φαινόμενα διαχωρίζονται με βάση το αν επιδεικνύουν συμπεριφορά «απλής» (*"simplicits"*) ή «πολύπλοκης» (*"complexity"*) δομής. Λέγοντας «απλή» δομή εννοούμε πώς το φαινόμενο έχει στοιχεία προβλεψιμότητας όπώς αρμονικότητα ή περιοδικότητα, ενώ με τον όρο «πολύπλοκη» δομή δηλώνουμε την έλλειψη οποιασδήποτε συστηματικής συμπεριφοράς. Οι επιστήμονες παρατηρώντας τα δυναμικά συστήματα έθεσαν γρήγορα το ερώτημα αν μπορούσαν να προβλέψουν την μελλοντική συμπεριφορά τους. Για παράδειγμα, οι φυσικοί και οι αστρονόμοι ήθελαν να γνωρίζουν ποια θα είναι η θέση του πλανήτη Άρη μετά από τρεις μήνες ενώ οι οικονομολόγοι ενδιαφέρονταν να ξέρουν ποιος θα είναι ο πληθωρισμός του επόμενου έτους. Πότε όμως ένα διαχρονικό φαινόμενο είναι προβλέψιμο; Ένα φαινόμενο που εξελίσσεται μέσα στον χρόνο μπορεί να προβλεφθεί όταν μπορούμε να κατασκευάσουμε ένα μαθηματικό μοντέλο, δηλαδή μια μαθηματική εξίσωση ή ένα σύνολο μαθηματικών εξισώσεων, το οποίο να έχει την ικανότητα να αναπαράγει όσο το δυνατό πιο πιστά την παρατηρούμενη συμπεριφορά.
Με βάση την παραπάνω διαπίστωση μπορούμε να διαχωρίσουμε τα μοντέλα σε ντετερμινιστικά και στοχαστικά. Τα πρώτα είναι εκείνα που περιγράφουν επακριβώς την παρατηρούμενη συμπεριφορά ενώ τα δεύτερα περιγράφουν κατά προσέγγιση την καταγεγραμμένη πορεία του φαινομένου στο χρόνο, με το βαθμό αυτής της προσέγγισης να ποικίλει. Αυτή η ιδιότητα των μοντέλων μας επιτρέπει να ορίσουμε δευτερογενώς και τις έννοιες του ντετερμινιστικού (ή αιτιοκρατικού) φαινομένου και του στοχαστικού φαινομένου. Ντετερμινιστικό φαινόμενο λοιπόν είναι αυτό που περιγράφεται (θα

λέγαμε ότι έχει την καλή τύχη να περιγράφεται) από ένα ντετερμινιστικό μοντέλο ενώ στοχαστικό φαινόμενο είναι αυτό που ερμηνεύεται από ένα στοχαστικό μοντέλο (Πιττής (199)).

1.3.1 Ντετερμινιστικά Μοντέλα

Έχοντας δώσει μια πρώτη εικόνα των μοντέλων, στην ενότητα που θα ακολουθήσει θα παρουσιάσουμε τα αιτιοκρατικά μοντέλα δίνοντας γενικά χαρακτηριστικά τους τα οποία θα μας βοηθήσουν ομαλά να μεταβούμε στην συνέχεια στα στοχαστικά μοντέλα. Τα ντετερμινιστικά μοντέλα λοιπόν διακρίνονται σε δυο βασικές κατηγορίες.

(Α) Ντετερμινιστικά μοντέλα τύπου Ι.
Στην δεδομένη κατηγορία εντάσσονται οι απλές συναρτήσεις του χρόνου. Δηλαδή θα έχουμε μοντέλα της μορφής

$$y_t = f(t)$$

όπου
α) y_t είναι η κατάσταση του συστήματος την χρονική στιγμή t.
β) t είναι ο χρόνος ο οποίος μπορεί να υποτεθεί ότι λαμβάνει είτε φυσικές τιμές (δηλαδή $0, 1, 2, 3, \ldots$) είτε συνεχείς τιμές (δηλαδή $t \in [a, b]$).
γ) f είναι η συνάρτηση που μετασχηματίζει το t σε y_t.
Παράδειγμα. Από την κλασσική μηχανική γνωρίζουμε ότι η θέση ενός κινητού που εκτελεί ελεύθερη πτώση τον χρόνο t δίνεται από την συνάρτηση $S(t) = \frac{1}{2}gt^2$.

(Β) Ντετερμινιστικά μοντέλα τύπου ΙΙ ή Δυναμικά μοντέλα.
Με βάση την αντιμετώπιση του χρόνου ως διακριτού ή συνεχούς διακρίνουμε δυο βασικές κατηγορίες Δυναμικών Μοντέλων.

i) Δυναμικά μοντέλα Διακριτού χρόνου.
Ο εκπρόσωπος αυτής της κατηγορίας μοντέλων είναι οι εξισώσεις διαφορών. Συχνά λοιπόν είναι ευκολότερο να περιγράψουμε την δυναμική συμπεριφορά του συστήματος, όχι άμεσα σε όρους $y_t = f(t)$ αλλά έμμεσα σε όρους του *Νόμου της Κίνησης (Law of Motion)* του y_t. Ο νόμος της κίνησης ενός συστήματος ορίζει τον τρόπο με τον οποίο μεταβάλλονται οι καταστάσεις αυτού μια οποιαδήποτε χρονική στιγμή, ως συνάρτηση των καταστάσεων του σε προγενέστερες χρονικές στιγμές. Κάθε εξίσωση

που περιγράφει τον νόμο της κίνησης ενός συστήματος ονομάζεται *Εξίσωση Διαφορών (Difference Equation).*
Χαρακτηριστικό παράδειγμα εξίσωσης διαφορών αποτελεί το Γραμμικό Δυναμικό Μοντέλο $1^{ου}$ βαθμού που έχει την μορφή

$$y_t = a_0 + a_1 y_{t-1}. \tag{1.1}$$

Προκειμένου η μαθηματική εξίσωση (1.1) να μας δώσει την κίνηση του y_t μέσα στον χρόνο πρέπει να αντιμετωπιστούν τα παρακάτω ζητήματα.
α) Αρχική κατάσταση του συστήματος y_0.
β) Πορεία του συστήματος μέσα στον χρόνο. Δηλαδή καθώς το $t \to \infty$ το σύστημα συγκλίνει σε κάποια τιμή ισσοροπίας $\overline{y}$ ή αποκλίνει; Μήπώς όταν το $t \to \infty$ το σύστημα μετακινείται προς κάποιο είδος ισσοροπίας περιοδικού τύπου; Όπώς γνωρίζουμε η ύπαρξη και το είδος της ισορροπίας του y_t σύμφωνα με την (1.1) εξαρτάται από το y_0 και την τιμή της σταθεράς a_1. Για να επιτευχθούν οι προαναφερθέντες στόχοι απαιτείται να επιλύσουμε την (1.1) δηλαδή να την φέρουμε στην μορφή $y_t = f(t)$.

Σε περιπτώσεις όπου η κατάσταση ενός συστήματος y_t αλληλεπιδρά με την κατάσταση ενός άλλου συστήματος x_t και δεν γνωρίζουμε αν το x_t προκαλεί το y_t ή το αντίστροφο, για την διαχρονική μελέτη της συμπεριφοράς των δύο συστημάτων χρησιμοποιούμε το Σύστημα Εξισώσεων Διαφορών. Ένα παράδειγμα τέτοιου συστήματος που προσπαθεί να περιγράψει τις δυναμικές αλληλεπιδράσεις του y_t και του x_t είναι το ακόλουθο

$$\begin{cases} y_t = a_{01} + a_{11} y_{t-1} + a_{12} x_{t-1} \\ x_t = a_{02} + a_{21} y_{t-1} + a_{22} x_{t-1} \end{cases}. \tag{1.2}$$

Το δεδομένο σύστημα επιτρέπει τις παρελθούσες τιμές της x_t να επηρεάζουν την τρέχουσα τιμή της y_t (εάν $a_{12} \neq 0$) όπώς επίσης επιτρέπει τις παρελθούσες τιμές της y_t να επηρεάζουν την τρέχουσα τιμή της x_t (εάν $a_{21} \neq 0$). Η ύπαρξη και το είδος της ισορροπίας του συστήματος εξαρτώνται από τις ιδιοτιμές του πίνακα $\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} \\ a_{21} & a_{22} \end{bmatrix}$.

ii) Δυναμικά Μοντέλα σε Συνεχή χρόνο.
Σε περίπτωση που μετράμε τον χρόνο σε συνεχές διάστημα, οι βασικές ιδέες της ανάλυσης των ντετερμινιστικών μοντέλων παραμένουν οι ίδιες. Το μόνο που αλλάζει είναι η μεθοδολογική προσέγγιση. Έτσι αντί για Εξισώσεις Διαφορών χρησιμοποιούμε τις *Διαφορικές Εξισώσεις (Differential Equations).* Σημαντική κατηγορία Διαφορικών

Εξισώσεων είναι οι Γραμμικές Διαφορικές Εξισώσεις που έχουν την γενική μορφή

$$a_0(t)\frac{d^n y}{dt^n} + a_1(t)\frac{d^{n-1} y}{dt^{n-1}} + \ldots + a_n(t)y = g(t).$$

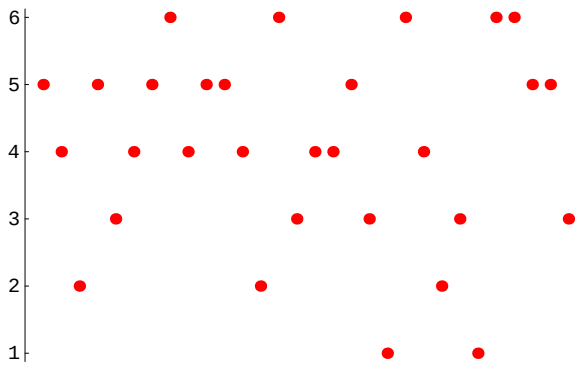
Θεωρώντας $a_0(t) = a_0, a_1(t) = a_1, \ldots, a_n(t) = a_n, g(t) = g$ και θέτοντας $n = 1$ έχουμε την ειδική περίπτωση

$$\frac{dy}{dt} + \beta y = \gamma \tag{1.3}$$

όπου $\beta = \frac{a_n}{a_0}$ και $\gamma = \frac{g}{a_0}$. Η σχέση (1.3) αποτελεί το συνεχές ανάλογο της εξίσωσης (1.1) και εκφράζει τον νόμο της κίνησης του συστήματος. Με σκοπό να βρούμε την πορεία του y στον χρόνο (δηλαδή το $y(t)$) απαιτείται να επιλύσουμε την (1.3). Όπως και στην περίπτωση των Εξισώσεων Διαφορών η ύπαρξη και το είδος ισορροπίας εξαρτάται από την αρχική κατάσταση του συστήματος και την σταθερά β. Τέλος για ντετερμινιστικά φαινόμενα τα οποία συμμεταβάλονται σε συνεχή χρόνο, η μοντελοποίηση τους μπορεί να πραγματοποιηθεί μέσω συστημάτων Διαφορικών Εξισώσεων.
Ολοκληρώνοντας την συζήτηση περι Ντετερμινιστικών Μοντέλων πρέπει να τονίσουμε ότι η αιτιοκρατική θεωρία που χρησιμοποιείται για την κατασκευή τους είναι αναγκαίο να είναι η 'καλύτερη' από τις ήδη υπάρχουσες οδηγώντας σε νόμους με καθολική ισχύ. Ακόμη με τα δεδομένα μοντέλα, η συμπεριφορά των φαινομένων που μελετάμε τον χρόνο $T + k$ είναι ήδη γνωστή την χρονική στιγμή t ακόμη και για $k \to \infty$. Σε περίπτωση που τα παρατηρούμενα δεδομένα δεν συμφωνούν με τα αποτελέσματα που παρέχουν οι μαθηματικές εξισώσεις, εφόσον ορθώς έχουμε θεωρήσει ότι αυτές αναπαράγουν πιστά την εξέλιξη του συστήματος, τότε απαιτείται να αναζητήσουμε καταστάσεις του φαινομένου που δεν έχουν παρατηρηθεί. Όπως θα διαπιστώσουμε στην συζήτηση που θα ακολουθήσει, τα αιτιοκρατικά φαινόμενα αποτέλεσαν την βάση για την κατασκευή υποδειγμάτων που σκοπό είχαν την περιγραφή στοχαστικών φαινομένων, συμβάλλοντας σημαντικά στην ανάπτυξη της θεωρίας των χρονοσειρών.

1.3.2 Μετάβαση στον κόσμο της Τυχαιότητας - Στοχαστικά Μοντέλα

Υπάρχουν διαχρονικά φαινόμενα η παρατηρούμενη συμπεριφορά των οποίων δεν μπορεί να περιγραφεί από τα διαθέσιμα ντετερμινιστικά μοντέλα. Για παράδειγμα η πραγματοποίηση που παριστά 30 διαδοχικές ρίψεις ενός ζαριού στο Σχήμα 1.1 δεν εμφανίζει οποιαδήποτε «κανονικότητα» ή «συστηματικότητα», γεγονός που καθιστά δύσκολη έως και ανέφικτη, την πρόβλεψη μελλοντικών καταστάσεων βάσει παρελθουσών τιμών. Με σκοπό λοιπόν να αναπαραγάγουμε την πολύπλοκη δομή που επιδεικνύουν



Σχήμα 1.1: Τριάντα ζαριές

αρκετά Δυναμικά Συστήματα καταφεύγουμε στα Στοχαστικά Μοντέλα στα οποία όμως χάνεται η απόλυτη ακρίβεια της περιγραφής και της πρόβλεψης. Με αυτήν την απλή διαπίστωση εισερχόμαστε στον κόσμο της Τυχαιότητας ή της Αβεβαιότητας. Στο νέο πλαίσιο προσέγγισης των διαχρονικών φαινομένων ο μηχανισμός που παράγει τις παρατηρούμενες καταστάσεις y_t, ουσιαστικά την δειγματοληπτική διαδρομή της άγνωστης στοχαστικής ανέλιξης, χαρακτηρίζεται από αβεβαιότητά για το αποτέλεσμα του πειράματος, μη ακριβείς προβλέψεις και πολλαπλότητα των δυνατών πραγματοποιήσεων. Ο δεδομένος μηχανισμός που διέπει τα στοχαστικά φαινόμενα ονομάζεται *μηχανισμός τύχης*. Σκοπός των στοχαστικών μοντέλων είναι η όσο το δυνατόν καλύτερη περιγραφή των βασικών χαρακτηριστικών του «αόρατου» μηχανισμού τύχης που γεννά σειρές αριθμών όπως εκείνη του Σχήματος 1.1.
Ο απλούστερος μηχανισμός τύχης ο οποίος χαρακτηρίζει αρκετά μη ντετερμινιστικά φαινόμενα είναι το *Τυχαίο Πείραμα*.

Ορισμός 1.3.1 *Τυχαίο Πείραμα είναι ο μηχανισμός τύχης που πληροί τις ακόλουθες τρείς συνθήκες*
1) Όλα τα δυνατά αποτελέσματα του πειράματος είναι γνωστά εκ των προτέρων.
2) Σε κάθε επανάληψή του πειράματος το αποτέλεσμα δεν είναι γνωστό εκ των προτέρων.
3) Το πείραμα μπορεί να επαναληφθεί κάτω από απόλυτα όμοιες συνθήκες.

Παρατήρηση 1.3.1 Μηχανισμοί τύχης που δεν πληρούν την τρίτη συνθήκη και στους οποίους οι συνθήκες διεξαγωγής του πειράματος αλλάζουν από επανάληψη σε επανάληψη δεν θα ονομάζονται Τυχαίο Πείραμα αλλά απλά Μηχανισμοί Τύχης.

1.3.3 Μαθηματική τυποποίηση του Τυχαίου Πειράματος

Η πρώτη συνθήκη τυποποιείται μέσω του χώρου των αποτελεσμάτων S. Για παράδειγμα στο πείραμα ρίψης ενός νομίσματος δυο φορές για το S θα έχουμε $S = \{\text{ΚΚ}, \text{ΓΓ}, \text{ΚΓ}, \text{ΓΚ}\}$.

Για την τυποποίηση της δεύτερης συνθήκης πρέπει να λάβουμε υπόψη το γεγονός ότι η αβεβαιότητά που μας διεκατέχει σχετικά με την έκβαση του πειράματος κατά την οποιαδήποτε εκτέλεση του, μεταφράζεται σε αβεβαιότητα σχετικά με την εμφάνιση ενός ενδεχομένου ενδιαφέροντος, έστω Α, που εμείς έχουμε προκαθορίσει. Με τον τρόπο αυτό η αβεβαιότητα μας θα σχετίζεται και με κάθε άλλο ενδεχόμενο του πειράματος που περιέχει το Α. Έτσι δημιουργούμε ένα σύνολο $\mathcal{F}$ στο οποίο τοποθετούμε τα ενδεχόμενα ενδιαφέροντος και τα συναφή τους. Το δεδομένο σύνολο είναι κλειστό ως προς τις πράξεις της συμπληρωματικότητας, της ένωσης και της τομής και αποτελεί μια σ-άλγεβρα. Ειδικότερα αν το $\mathcal{F}$ έχει προκύψει από το ενδεχόμενο Α θα το λέμε, σύμφωνα με την θεωρία πιθανοτήτων, χώρο ενδεχομένων που προέκυψε από το Α και άρα μπορούμε να το ονομάσουμε σ-πεδίο που γεννήθηκε από το Α, συμβολίζοντας το με σ(Α). Για να αποκρυσταλλώσουμε όμως την αρχική μας αβεβαιότητα για το αποτέλεσμα του πειράματος σε κάθε στοιχείο της $\mathcal{F}$ προσάπτουμε πιθανότητες μέσω μίας συνάρτησης $P : \mathcal{F} \to [0, 1]$. Με τον τρόπο αυτό μπορούμε να πούμε πως η δεύτερη συνθήκη μεταφράζεται σε μαθηματικούς όρους μέσω της δομής $(S, \mathcal{F}, P(\cdot))$ που είναι ο γνωστός μας *Χώρος Πιθανοτήτων*.

Όπως προαναφέρθηκε για να είναι ένα πείραμα τυχαίο απαιτείται να επαναλαμβάνεται κάτω από απόλυτα όμοιες συνθήκες. Για να συμβαίνει κάτι τέτοιο είναι ανάγκη σε κάθε επανάληψή του να ικανοποιούνται οι συνθήκες της *Ταυτονομίας* και της *Ανεξαρτησίας*. Δηλαδή σε κάθε εκτέλεση του πρέπει ο χώρος πιθανοτήτων να παραμένει ίδιος και το αποτέλεσμα σε οποιοδήποτε χρόνο t να μην εξαρτάται από την έκβαση των προηγούμενων ή επόμενων επαναλήψεων. Για την μαθηματική έκφραση των δυο προαναφερθεισών συνθηκών θεωρούμε n διαδοχικές εκτελέσεις του πειράματος και συμβολίζουμε με $A_1, A_2, \ldots, A_n$ τα ενδεχόμενα ενδιαφέροντος σε κάθε μια από αυτές. Έτσι το πείραμα μας θα εκτελείται υπο απόλυτα όμοια συνθήκες αν ισχύουν

1)Ανεξαρτησία

$$P(A_1 \bigcap A_2 \bigcap \ldots \bigcap A_n) = \prod_{i=1}^{n} P(A_i)$$

2)Ταυτονομία

$$(S, F, P(\cdot))_i = (S, F, P(\cdot)) \quad \forall i = 1, 2, \ldots, n.$$

Συνεπώς το Τυχαίο Πείραμα μοντελοποιείται από δυο οντότητες. Η πρώτη είναι ο χώρος πιθανοτήτων $(S, F, P(\cdot))$ (συνθήκες 1, 2) και η δεύτερη είναι ενα σύνολο ανεξάρτητων και ταυτόνομων δοκιμών που θα το λέμε *Τυχαίο Δείγμα* και θα το συμβολίζουμε με $\sigma_n \equiv (A_1, A_2, \ldots, A_n)$ (συνθήκη 3). Ο πιθανοθεωρητικός χώρος μαζί με το τυχαίο δείγμα συνιστούν μια νέα μαθηματική δομή που ονομάζεται *Απλός Στατιστικός Χώρος* και συμβολίζεται ως $[(S, F, P(\cdot)), \sigma_n]$.
Η κλασική Στατιστική Συμπερασματολογία είχε αναπτύξει από πολύ νωρίς σημαντικές μεθόδους για την μοντελοποίηση φαινομένων τα οποία περιγράφονται από τον Απλό Στατιστικό Χώρο. Τα πράγματα όμως διαφοροποιούνται και δυσκολεύουν αρκετά όταν αναφερόμαστε σε καθημερινά φαινόμενα τα οποία δεν διεξάγονται σε συνθήκες εργαστηρίου, όπως για παράδειγμα τα μετεωρολογικά φαινόμενα, με αποτέλεσμα να μην εξασφαλίζονται οι προυποθέσεις του τυχαίου πειράματος. Για την μοντελοποίηση τέτοιων φαινομένων σημαντική ήταν και εξακολουθεί να είναι η θεωρία των χρονολογικών σειρών που, όπως θα δούμε και στην συνέχεια, οδήγησε σε νέες μεθοδολογικές προσεγγίσεις για την ανάλυση δυναμικών φαινομένων που επιδεικνύουν διαχρονική εξάρτηση και ετερογένεια.

1.4 Ιστορικά στοιχεία της θεωρίας χρονοσειρών

Τα πρώτα μοντέλα στις χρονοσειρές είχαν την γενική μορφή

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

(βλέπε *Priestley* (1981)) όπου
T_t : η 'τάση' που περιγράφει την μακροχρόνια συμπεριφορά της σειράς.
S_t : η περιοδική συνιστώσα.
ϵ_t : το τυχαίο σφάλμα.
Σε πρώτη φάση συνηθιζόταν να εκτιμάται η τάση T_t μέσω μιας πολυωνυμικής συνάρτησης του t χαμηλής τάξης. Με τον τρόπο αυτό οδηγούμασταν στο μοντέλο

$$\widetilde{X}_t = S_t + \epsilon_t, \tag{1.4}$$

όπου πλέον η τάση έχει ενσωματωθεί στην νέα μεταβλητή $\widetilde{X}_t$. Σε περιπτώσεις που η περιοδικότητα του φαινομένου ήταν γνωστή, τότε το μοντέλο (1.4) γραφόταν στην μορφή

$$\widetilde{X}_t = \sum_i (A_i \cdot \cos \omega_i t + B_i \cdot \sin \omega_i t) + \epsilon_t \tag{1.5}$$

με τα ω_i (που εκφράζουν συχνότητες) γνωστά. Σε περιπτώσεις όμως που δεν διαθέταμε αυστηρή γνώση της περιοδικότητας του διαχρονικού φαινομένου ήταν ανάγκη, πέρα από τους συντελεστές $\{A_i\}, \{B_i\}$, να εκτιμήσουμε και τις άγνωστες συχνότητες ω_i. Για το σκοπό αυτό χρησιμοποιήθηκε η συνάρτηση με την ονομασία *περιοδόγραμμα (periodogram)*. Η ανάλυση χρονοσειρών που στηριζόταν στο περιοδόγραμμα ήταν πολύ χρήσιμη τεχνική μόνο όταν εφαρμοζόταν σε σειρές που είχαν συστηματική δομή, αδυνατώντας να περιγράψει φαινόμενα με πολύπλοκες συμπεριφορές. Την επανάσταση στην ανάλυση των χρονολογικών σειρών έφερε ο *Yule* το 1927, που θεωρείται και η χρονιά γέννησης της σύγχρονης θεωρίας τους (βλέπε *Tong* (2001)).

Ο *Yule* λοιπόν ήθελε να μοντελοποιήσει τους στοχαστικούς κύκλους που επιδείκνυε η γνωστή τότε χρονοσειρά των ηλιακών ακτίνων του *Wolfer*. Παρατηρώντας την δεδομένη σειρά διαπίστωσε ότι ταλαντωνόταν με μη «κανονικό» τρόπο και ο κάθε στοχαστικός κύκλος είχε ασύμμετρο σχήμα. Κατά συνέπεια δεν θα ήταν ρεαλιστικό να προσπαθήσει να προσαρμόσει στα δεδομένα το μοντέλο της σχέσης (1.5) το οποίο περιελάμβανε αυστηρά περιοδικούς και συμμετρικούς όρους όπως οι τριγωνομετρικές συναρτήσεις του ημιτόνου και του συνημιτόνου. Ο *Yule* σκέφτηκε πως η παρατηρούμενη συμπεριφορά της σειράς του *Wolfer* μπορούσε να αναπαραχθεί κατά προσέγγιση από την εξαναγκασμένη ταλάντωση που εκτελούσε ένα εκρεμμές, επιτρέποντας όμως στις εξωτερικές δυνάμεις να συμβαίνουν σε τυχαίες χρονικές στιγμές. Με τον τρόπο αυτό το εκρεμμές ταλαντωνόταν παρουσιάζοντας ανωμαλία στις αποκλίσεις του από την θέση ισορροπίας. Οι αυξομοιώσεις αυτών των αποκλίσεων παρουσίαζαν παρόμοια εικόνα με τις κυμάνσεις των ηλιακών ακτινών. Επειδή, όπως γνωρίζουμε, η κίνηση ενός εκρεμμούς (βάσει κατάλληλων υποθέσεων) μπορεί να περιγραφεί από μια διαφορική εξίσωση δεύτερης τάξης (που είναι ντετερμινιστικό μοντέλο) ο *Yule* για την προσεγγιστική περιγραφή της χρονοσειράς του *Wolfer* πρότεινε το στοχαστικό μοντέλο

$$\frac{dX(t)}{dt^2} + a_1 \frac{dX(t)}{dt} + a_2 X(t) = \epsilon(t) \tag{1.6}$$

όπου $X(t)$ η γωνιακή απόκλιση του εκρεμμούς τον χρόνο t από το σημείο ισορροπίας. Το διακριτό ανάλογο της σχέσης (1.6) είναι η δεύτερης τάξης εξίσωση διαφορών

$$X_t + a_1 X_{t-1} + a_2 X_{t-2} = \epsilon_t, \tag{1.7}$$

όπου ϵ_t είναι η διαδικασία λευκού θορύβου (*White Noise*). Η εξίσωση (1.7) που σε σύγχρονη ορολογία ονομάζεται αυτοπαλίνδρομο υπόδειγμα δευτερης τάξης ($AR(2)$) ήταν ουσιαστικά το μοντέλο που πρότεινε ο *Yule*. Η δεδομένη προσέγγιση μαζί με τις

ιδέες που περιέχει σηματοδότησαν μια νέα εποχή στην ανάλυση χρονοσειρών. *Πλέον ήταν ορατό, από τον τρόπο που καταλήξαμε στις εξισώσεις (1.6) και (1.7), ότι η ερμηνεία των στοχαστικών φαινομένων θα στηριζόταν στα ήδη γνωστά αιτιοκρατικά μοντέλα τα οποία όμως θα έπρεπε να προσαρμοστούν κατάλληλα ώστε να λαμβάνουν υπόψη τον παράγοντα τύχη.*

από την στιγμή λοιπόν που οδηγηθήκαμε στο αυτοπαλίνδρομο υπόδειγμα (1.7), οι εξελίξεις στον χώρο των χρονοσειρών ήταν ραγδαίες. Τα πρακτικά προβλήματα που προέκυψαν σε πολλούς διαφορετικούς επιστημονικούς κλάδους, με εξέχοντα εκείνον της μηχανικής των τηλεπικοινωνιών, έφεραν τις χρονοσειρές μπροστά σε νέες προκλήσεις. Η ιδιομορφία των διαχρονικών φαινομένων, αναλόγως με την επιστημονική περιοχή από την οποία προέρχονταν, οδήγησε στην ανάπτυξη αρκετών διαφορετικών προσεγγίσεων Στατιστικής Συμπερασματολογίας των χρονολογικών σειρών.

Έτσι, σε οικονομικά φαινόμενα χρησιμοποιήθηκαν κυρίως οι προσεγγίσεις του πεδίου του χρόνου (*time domain*). Στο συγκεκριμένο πλαίσιο αρχικά η ανάλυση των σειρών στηρίχτηκε στις συναρτήσεις *αυτοσυσχέτισης (autocorrelation)* και *αυτοσυνδιακύμανσης (autocovariance)*. Με την δεδομένη τεχνική οι στατιστικοί μπορούσαν να έχουν μια ένδειξη αν οι χρονοσειρές επεδείκνυαν *στασιμότητα (stationarity)*[1]. Οι πρώτοι που παρουσίασαν την συνάρτηση της αυτοσυσχέτισης ως εργαλείο συμπερασματολογίας στις χρονοσειρές ήταν ο *Bartlett* (1950) και ο *Kendall* (1954).

Άμεση σχέση με τις συναρτήσεις της αυτοσυσχέτισης και της αυτοσυνδιακύμανσης είχε η τεχνική ανάλυσης χρονοσειρών που εισήχθη από τους *Box* και *Jenkins* (1976). Οι δεδομένοι συγγραφείς για στάσιμες χρονοσειρές συνδύασαν τα μοντέλα AR με τα μοντέλα κινητού μέσου (*moving average-MA*) και οδηγήθηκαν σε μια νεα κατηγορία υποδειγμάτων που ονομάστηκαν μικτά αυτοπαλίνδρομα κινητού μέσου (*mixed autoregressive moving average-ARMA*) μοντέλα. Το πλεονέκτημα των μοντέλων $ARMA$ είναι ότι η προσαρμογή τους, σε γενικές γραμμές, απαιτεί πολύ λιγότερες παραμέτρους από ένα 'φτωχό' AR ή ένα 'φτωχό' MA μοντέλο.

Οι προαναφερθείσες τεχνικές ανάλυσης χρονολογικών σειρών δεν εφαρμόσθηκαν για την μοντελοποίηση φυσικών φαινομένων. Για το σκοπό αυτό αξιοποιήθηκε η *φασματική ανάλυση (spectral analysis)*.

Όταν οι χρονοσειρές λοιπόν αναπαριστούν μια φυσική ποσότητα (όπως είναι η τάση

[1]Με την έννοια στασιμότητα εννοούμε πως η χρονοσειρά παρουσιάζει μια συμπεριφορά «σταθερής κατάστασης» (*"steady state"*) και μπορούμε να πούμε πως βρίσκεται σε κατάσταση «στατιστικής ισορροπίας» (*"statistical equilibrium"*).

του ρεύματος, η μετατόπιση, η ταχύτητα κ.τ.λ) συχνά έχει περισσότερο ενδιαφέρον να μελετήσουμε τις φυσικές της ιδιότητες άμεσα παρά να κατασκευάσουμε ένα στατιστικό μοντέλο για τη σειρά. Μια από τις πιο θεμελιακές έννοιες στην φυσική είναι εκείνη της *ενέργειας*, και ίσως η σημαντικότερη μέθοδος που χρησιμοποιείται για την ερμηνεία μια φυσικής διαδικασίας είναι η διάσπασή της σε έναν αριθμό διαφορετικών συνιστωσών συχνότητας (*frequency components*), οι οποίες εξετάζονται ως προς την ποσότητα της συνολικής ενέργειας που απορροφούν. Αυτή η διαδικασία ονομάζεται φασματική ανάλυση (*Subba Rao, Priestley and Lessi (1997)*).

Για την ανάπτυξη των μεθόδων της φασματικής ανάλυσης, που εναλλακτικά ονομάζεται προσέγγιση στο πεδίο των συχνοτήτων (*frequency domain*), σημαντική ήταν η προσφορά του *Bartlett* (1950). Για τις μαθηματικές τεχνικές που χρησιμοποιούνται στην φασματική ανάλυση ο αναγνώστης παραπέμπεται στον *Hamilton* (1994).

Άλλη σημαντική τεχνική της ανάλυσης χρονοσειρών, που συνέβαλε σημαντικά στην διεξαγωγή προβλέψεων, ήταν το λεγόμενο «φιλτράρισμα» (*"filtering"*) των δεδομένων. Η συγκεκριμένη προσέγγιση προέκυψε από την προσπάθεια πρόβλεψης μιας σειράς $\{X_t\}$, την οποία δεν παρατηρούμε άμεσα, μέσω των τιμών που λαμβάνει μια άλλη σειρά $\{Y_t\}$ η οποία διαμορφώνεται από το μοντέλο

$$Y_t = X_t + N_t,$$

όπου $\{N_t\}$ είναι μια διαδικασία θορύβου. Το δεδομένο πρόβλημα εισήχθη από τον *Wiener* (1949). Μια πρωτοποριακή λύση στο συγκεκριμένο πεδίο παρουσιάστηκε από τον *Kalman* (1960) και τοθς *Kalman and Bucy* (1961), οι οποίοι χρησιμοποίησαν την λεγόμενη *«αναπαράσταση κατάστασης-χώρου» ("state-space representation")* των χρονοσειρών. Η πρόταση των *Kalman* και *Bucy* οδήγησε σε έναν ευφυή επαναληπτικό αλγόριθμο για τον υπολογισμό προβλέψεων που είναι γνωστός ως αλγόριθμος *Kalman filter*.

Συνάμα στον αιώνα που πέρασε σημαντικά βήματα σημειώθηκαν στην ανάλυση μη στάσιμων χρονοσειρών. Οι πρώτοι που ασχολήθηκαν με αυτό το αντικείμενο ήταν ο *Fuller* και ο *Dickey* στις αρχές της δεκαετίας του 1980, που κατασκεύασαν το ομώνυμο στατιστικό τέστ για τον έλεγχο μοναδιαίας ρίζας.

Οι πρόσφατες έρευνες στην ανάλυση των χρονοσειρών εστιάζονται στην μελέτη των *μη γραμμικών (non-linear)* μοντέλων. Η κατασκευή τέτοιων μοντέλων μπορεί να παρουσιάζει δυσεπίλυτα προβλήματα από πλευράς υπολογισμών, αλλα ταυτόχρονα οδηγεί στην παρατήρηση συναρπαστικών δομικών ιδιοτήτων των παρατηρούμενων φαινομένων, οι οποίες δεν μπορούν να αναπαραχθούν από τα γραμμικά υποδείγματα. Οι

πιο ενδιαφέρουσες κλάσεις μη γραμμικών μοντέλων είναι οι ακόλουθες
(1) *Bilinear* μοντέλα (*Subba* και *Gabr* (1984)).
(2) *Threshold autoregressive* μοντέλα (*Tong* και *Lim* (1980)).
(3) *Exponential autoregressive* μοντέλα (*Haggan* και *Ozaki* (1979)).
(4) *General state-dependent* μοντέλα (*Priestley* (1980)).

1.5 Ιστορική Αναδρομή στις Κατηγορικές Χρονοσειρές

Ποιοτικές χρονοσειρές συχνά συναντάμε σε ποίκιλες εφαρμογές. Για παράδειγμα, το αν βρέχει ή όχι για ένα διαδοχικό αριθμό ημερών είναι μια δίτιμη χρονοσειρά που αποτελεί την πιο απλή περίπτωση κατηγορικής χρονοσειράς. Ακόμη τα επίπεδα του ύπνου ενός νεογέννητου μωρού, για μια ακολουθία ημερών, με τιμές
(1) ήσυχος ύπνος
(2) ελαφρά ανήσυχος
(3) έντονα ανήσυχος
(4) καθόλου ύπνος
συνιστά ακόμη ένα παράδειγμα κατηγορικής χρονοσειράς. Αναλυτικότερα, στην προκειμένη περίπτωση για τις ημέρες $t = 1, 2, \ldots, N$, η μεταβλητή ενδιαφέροντος Y_t ='κατάσταση ύπνου', λαμβάνει τις ακέραιες τιμές 1,2,3,4. Όπως και στις ποσοτικές χρονοσειρές, οι στατιστικοί ασχολήθηκαν με προβλήματα όπως:
1) Έλεγχος ύπαρξης όρων περιοδικότητας.
2) Αν οι χρονικές υστερήσεις της Y_t επηρεάζουν τις μελλοντικές τιμές της.
3) Ποιος είναι ο καλύτερος τρόπος να προβλέψουμε μελλοντικά επίπεδα της Y_t.
4) Μπορούν άλλες συμμεταβλητές να χρησιμοποιηθούν στις προβλέψεις;
Τα τελευταία 30 χρόνια που το συγκεκριμένο πεδίο απασχολεί τους στατιστικούς έχει προταθεί ένας μεγάλος αριθμός διαφορετικών στρατηγικών για την μοντελοποίηση ποιοτικών χρονοσειρών. Η πρώτη προσέγγιση στο δεδομένο πεδίο πραγματοποιήθηκε μέσω των Μαρκοβιανών Αλυσίδων (*Markov Chains*). Ο συγκεκριμένος στατιστικός κλάδος έχει εξεταστεί διεξοδικά από πολλούς ερευνητές (βλέπε *Karlin and Taylor*, (1975)) και έχει αναπτυχθεί αρκετά η θεωρία Στατιστικής Συμπερασματολογίας του (βλέπε *Basawa and Prakasa Rao* (1980, κεφάλαιο 4)). Η δεδομένη όμως στρατηγική ανάλυσης παρουσιάζει, όπως θα δούμε και στο Κεφάλαιο 2, σημαντικά μειονεκτήματα τα οποία γίνονται εντονότερα όταν οι κατηγορικές αποκρίσεις εξαρτώνται πέρα από το παρελθόν τους και από εξωγενείς μεταβλητές. Στις μέρες μας για την στατιστική ανάλυση των κατηγορικών χρονοσειρών εξέχουσα θέση κατέχουν τα μοντέλα που

στηρίζονται στην θεωρία των *GLM*. Τα δεδομένα μοντέλα εισήχθησαν από τους *Nelder* και *Wedderburn* (1972) και οδήγησαν σε ένα μεγάλο εύρος προσεγγίσεων για συνεχείς όσο και για διακριτές μεταβλητές. Η δεδομένη θεωρία δεν χρησιμοποιήθηκε αμιγώς στις κατηγορικές χρονοσειρές αφού δεν μπορούσε να λάβει υπόψη της την διαχρονική εξάρτηση των δεδομένων.

Η πρώτη απόπειρα μοντελοποίησης των κατηγορικών χρονοσειρών, μέσω των αποτελεσμάτων των *GLM*, οφείλεται στον *Cox* (1970), ο οποίος πρότεινε για δίτιμες μεταβλητές ένα αυτοπαλίνδρομο λογιστικό μοντέλο. Στο δεδομένο υπόδειγμα οι συμμεταβλητές και ένας πεπερασμένος αριθμός παρελθούσων αποκρίσεων είναι μέρος της γραμμικής πρόβλεψης. από τότε σημειώθηκαν αξιόλογα βήματα στο συγκεκριμένο στατιστικό πεδίο και προτάθηκαν αρκετές επεκτάσεις ώστε να εισαχθούν διαφορετικοί τύποι εξάρτησης που να αντιστοιχούν σε ποίκιλες σχέσεις ανάμεσα στις παρατηρήσεις. Οι τροποποιήσεις της θεωρίας των *GLM* ώστε να λαμβάνουν υπόψη την διαχρονική εξάρτηση των δεδομένων, αξιοποιήθηκε στην ανάλυση των Επαναλαμβανόμενων Μετρήσεων(*Repeated Measurements*)[2] (*Liang* και *Zeger* (1986)). Αν και οι δίτιμες χρονοσειρές αναπτύχθηκαν έντονα, μόλις πρόσφατα οι στατιστικοί ασχολήθηκαν με την συμπερασματολογία των διατάξιμων (*ordinal*) κατηγορικών χρονοσειρών με περισσότερες από δύο κατηγορίες (*Lang and Agresti* (1994), *Glonek and McCullagh* (1995)). Στην προσπάθεια αυτή συνετέλεσαν τα γνωστά αποτελέσματα των *McCullagh* και *Nelder* (1983), για την μοντελοποίηση ανεξάρτητων κατηγορικών διατάξιμων αποκρίσεων μέσω *GLM*. Παράλληλα με τους *Lang* και *Agresti* έχουν γίνει και από άλλους συγγραφείς προσπάθειες για την στατιστική ανάλυση των διατάξιμων ποιοτικών σειρών. Οι προσπάθειες αυτές εστιάζονταν στην τροποποίηση δίτιμων μοντέλων εξάρτησης ώστε να μπορούν να επιτρέπουν στις μεταβλητές να διαθέτουν περισσότερες από δυο κατηγορίες. Για παράδειγμα ο *Crouchley* (1995) και ο *Ten Have* (1996) επέκτειναν το δίτιμο μοντέλο του *Conaway* (1990) κατορθώνοντας να μοντελοποιήσουν διατάξιμα εξαρτημένα δεδομένα.

1.6 Βασικές Μέθοδοι Ανάλυσης Κατηγορικών Χρονοσειρών

Για την στατιστική ανάλυση κατηγορικών χρονοσειρών πέρα από τα μοντέλα παλινδρόμησης που στηρίζονται στην θεωρία των γενικευμένων γραμμικών μοντέλων, χρησιμο-

[2]Οι επαναλαμβανόμενες μετρήσεις διαφοροποιούνται από τις χρονοσειρές στο γεγονός ότι αναφέρονται σε δειγματοληπτικές διαδρομές μικρού μήκους.

ποιούνται και άλλες τεχνικές οι οποίες αν και είναι γνωστές από παλιά, μόλις πρόσφατα αξιοποιήθηκαν στα προβλήματα των ποιοτικών σειρών.

Μια σημαντική κατηγορία μοντέλων είναι τα ακέραια αυτοπαλίνδρομα (*integer autoregressive*) και κινητού μέσου (*moving average*) υποδείγματα. Στα πλαίσια αυτών των μοντέλων διακρίνουμε την τεχνική των *Branching Processes with Immigration*. Η δεδομένη προσέγγιση καθορίζεται μέσω της στοχαστικής εξίσωσης

$$X_n = \sum_{i=1}^{X_{n-1}} Y_{n,i} + I_n \quad , n = 1, 2, 3, \ldots \tag{1.8}$$

Οι διαδικασίες $\{Y_{n,i}\}$ και $\{I_n\}$ είναι αμοιβαία ανεξάρτητες, ανεξάρτητες από την X_0 και η κάθε μια συνιστά μια ακολουθία ανεξάρτητων και ταυτόνομων τυχαίων μεταβλητών. Με τον τρόπο αυτό η $\{X_n\}$ είναι μια Μαρκοβιανή Αλυσίδα Διακριτού χρόνου. Η στοχαστική εξίσωση (1.8) μετά από κατάλληλες υποθέσεις λαμβάνει την μορφή

$$X_n = mX_{n-1} + \lambda + \epsilon_n \quad n = 1, 2, 3, \ldots \tag{1.9}$$

όπου $\lambda = E[I_n], m = E[Y_{n,i}], \epsilon_n \equiv X_n - E[X_n \mid F_{n-1}]$ και η διαδικασία $\{\epsilon_n\}$ είναι μια *martingale* διαφορά[3]. Εκτιμήσεις για το m και το λ λαμβάνουμε μέσω της μεθοδολογίας των ελαχίστων τετραγώνων και των σταθμισμένων ελαχίστων τετραγώνων. Ειδικές περιπτώσεις του μοντέλου (1.8) είναι το ακέραιο αυτοπαλίνδρομο υπόδειγμα τάξης p, το ακέραιο μοντέλο κινητού μέσου τάξης q καθώς και το ακέραιο αυτοπαλίνδρομο και κινητού μέσου τάξης p και q αντίστοιχα (*ARMA(p,q)*), (*McKenzie (1985), (1986)*).

Εναλλακτική μεθοδολογική προσέγγιση για την ανάλυση κατηγορικών χρονοσειρών αποτελούν τα μοντέλα μεικτών κατανομών μετάβασης (*mixture transition distribution models-MTD*) που εισήχθηκαν από τον *Raftery* (1985a), επεκτείνοντας προηγούμενη εργασία του *Pegram* (1980), ως μια τεχνική που συντελεί στην «οικονομική» μοντελοποίηση Μαρκοβιανών αλυσίδων μεγάλης τάξης. Σύμφωνα με την δεδομένη ανάλυση ξεπερνιέται το πρόβλημα των εκθετικά αυξανόμενων παραμέτρων για μια Μαρκοβιανή αλυσίδα καθορίζοντας την δεσμευμένη πιθανότητα να παρατηρηθεί $X_t = i_0$, δοθέντος του παρελθόντος, ως ένα γραμμικό συνδυασμό των $X_{t-1}, X_{t-2}, \ldots, X_{t-p}$. Πιο συγκεκριμένα υποθέτουμε ότι

$$P[X_t = i_0 \mid X_{t-1} = i_1, \ldots, X_{t-p} = i_p] =$$

[3] Έστω ότι διαθέτουμε μια στοχαστική ανέλιξη $\{Y_t\}, t = 1, 2, \ldots$ η οποία είναι *martingale*. Τότε η στοχαστική διαδικασία $\{X_t\}, t = 1, 2, \ldots$, με $X_t = Y_t - Y_{t-1}$ θα είναι μια *martingale* διαφορά και θα ισχύει $E[X_t \mid X_{t-1}, X_{t-2}, \ldots] = 0$.

$$= \sum_{j=1}^{p} \lambda_j P[X_t = i_0 \mid X_{t-j} = i_j] = \sum_{j=1}^{p} \lambda_j q_{i_j i_0} \tag{1.10}$$

όπου $i_0, \ldots, i_p$ ανήκουν στο σύνολο $\{1, 2, \ldots, m\}$ και τα $q_{i_j i_0}$ είναι τα στοιχεία του $m \times m$ πίνακα μετάβασης $\mathbf{Q}$. Για το διάνυσμα των υστερήσεων $\boldsymbol{\lambda} = (\lambda_1, \lambda_2, \ldots, \lambda_p)'$ ικανοποιείται η σχέση $\sum_{j=1}^{p} \lambda_j = 1$ με $\lambda_j \geq 0$ ώστε η $P[X_t = i_0 \mid X_{t-1} = i_1, \ldots, X_{t-p} = i_p]$ να παίρνει τιμές ανάμεσα στο 0 και το 1. Σύμφωνα με το μοντέλο (1.10) μειώνεται ο αριθμός των παραμέτρων από $m^p(m-1)$ σε $m(m-1)+(p-1)$. Μπορεί να δειχθεί ότι η ασυμπτωτική συμπεριφορά του μοντέλου *MTD* είναι η ίδια με το πλήρες μοντέλο που χρησιμοποιούμε στις ανώτερης τάξης Μαρκοβιανές αλυσίδες. Η εκτίμηση των παραμέτρων $\boldsymbol{\lambda}$ και q_{ij} του υποδείγματος (1.10) πραγματοποιείται μέσω της μεγιστοποίησης της συνάρτησης του λογαρίθμου της πιθανοφάνειας που είναι

$$\sum_{i_0, \ldots, i_{p-1}}^{m} n_{i_0, \ldots, i_{p-1}} \log(\sum_{j=1}^{p} \lambda_j q_{i_j i_0}) \tag{1.11}$$

βάσει κατάλληλων περιορισμών για το $\boldsymbol{\lambda}$. Στη σχέση (1.11) τα $n_{i_0, \ldots, i_{p-1}}$ μετρούν τον αριθμό των φορών που εμφανίστηκε η πολυμεταβλητή παρατήρηση $\{X_t = i_0, \ldots, X_{t-p} = i_p\}$ στις n επαναλήψεις του πειράματος.

Επειδή υπάρχει σύνδεση ανάμεσα στις κατηγορικές χρονοσειρές και τα *hidden Markov* μοντέλα, το δεδομένο πεδίο έχει αναπτυχθεί αρκετά. Σημαντική ήταν η προσφορά του *MacDonald* και *Zucchini* (1997).

Τέλος για την ανάλυση των ποιοτικών χρονοσειρών αξιοποιούνται και οι μέθοδοι της φασματικής ανάλυσης η οποία είναι ιδιαίτερα σημαντική ειδικά αν ο στόχος μας είναι να ανακαλύψουμε περιοδικές συνιστώσες στα δεδομένα. Θα πρέπει να τονίσουμε ότι για τα κατηγορικά δεδομένα οι συμβατικές μέθοδοι φασματικής ανάλυσης είναι γενικά ακατάλληλες. Ο *Stoffer et al.* (1993) καθώς και οι *Stoffer* και *Tyler* (1998) ξεπέρασαν τις δυσκολίες εισάγοντας έναν ευφυή μετασχηματισμό από τις ποιοτικές σειρές σε ακολουθίες πραγματικών αριθμών, κατορθώνοντας έτσι να χρησιμοποιήσουν τις ήδη γνωστές τεχνικές στα νέα δεδομένα.

Κεφάλαιο 2

Η Μερική Πιθανοφάνεια ως Μέθοδος Στατιστικής Συμπερασματολογίας Χρονοσειρών

2.1 Εισαγωγή

Τα μοντέλα ανάλυσης χρονοσειρών που αναφέρονται σε δεδομένα που ακολουθούν (ακόμη και κατα προσέγγιση) κανονική κατανομή έχουν μακρά παράδοση. Αντίθετα οι μέθοδοι στατιστικής επεξεργασίας μη κανονικών χρονοσειρών, σχετικά πρόσφατα προξένησαν το ενδιαφέρον των στατιστικών. Έτσι και οι τεχνικές ανάλυσης των κατηγορικών χρονοσειρών, που αποτελούν παράδειγμα τέτοιων πραγματοποιήσεων, αναπτύχθηκαν ιδιαίτερα τα τελευταία 30 χρόνια, όπως αναφέραμε και στο εισαγωγικό κεφάλαιο.

Οι ποιοτικές χρονοσειρές καθώς και οι δειγματοληπτικές διαδρομές μεταβλητών που λαμβάνουν διακριτές τιμές, προσεγγίστηκαν αρχικά μέσω των ομογενών Μαρκοβιανών αλυσίδων. Ο δεδομένος όμως τρόπος ανάλυσης μειονεκτούσε διότι χωρίς επιπλέον περιορισμούς ο αριθμός των παραμέτρων αυξανόταν εκθετικά καθώς μεγάλωνε η τάξη της αλυσίδας (*Fahrmeir and Tutz* (2001)). Επιπλέον σε πολλές εφαρμογές οι μη-ομογενείς Μαρκοβιανές αλυσίδες είναι πιο κατάλληλες από την στιγμή που εξωγενείς μεταβλητές συντελούν στην εμφάνιση μη στάσιμων πιθανοτήτων μετάβασης (*Fahrmeir and Kaufmann* (1987)). Τα προβλήματα που επιδεικνύουν τα Μαρκοβιανά μοντέλα αντιμετωπίζονται σε μεγάλο βαθμό από τα μοντέλα παλινδρόμησης για κατηγορικές χρονοσειρές, στα πλαίσια της θεωρίας των γενικευμένων γραμμικών μοντέλων (*Fokianos and Kedem* (2003)). Η ανάπτυξη και θεμελίωση των δεδομένων υποδειγμάτων συντελέστηκε από την θεωρία της μερικής πιθανοφάνειας (*Partial Likelihood-PL*) που αποτελεί σημαντική μέθοδο συμπερασματολογίας για εξαρτημένα δεδομένα (*Cox* (1975)).

Στις ενότητες που θα ακολουθήσουν θα δείξουμε, μέσα από απλά παραδείγματα, τις αναγκαιότητες που μας οδήγησαν στην χρήση της PL προκειμένου να μοντελοποιήσουμε εξαρτημένα κατηγορικά δεδομένα. Πιο συγκεκριμένα, στην Ενότητα 2 θα αναφερθούμε στην έννοια της δεσμευμένης πιθανοφάνειας (*Conditional Likelihood*), που αποτελεί ειδική περίπτωση της PL και θα δείξουμε ότι αποτελεί σημαντική προσέγγιση για την μοντελοποίση στοχαστικών φαινομένων των οποίων ο μηχανισμός τύχης δεν είναι το τυχαίο πείραμα. Εν συνεχεία, στην Ενότητα 3 θα επεκτείνουμε το αναλυτικό μας πλαίσιο ώστε να επιτρέπεται η παρουσία εξωγενών μεταβλητών, δίνοντας μια πρώτη ιδέα της χρηστικότητας της PL. Στην Ενότητα 4 θα δώσουμε αρχικά μια εφαρμογή της δεδομένης μεθόδου όταν ισχύουν κάποιες ευνοικές ιδιότητες, που συντελούν στην μείωση της διάστασης του προβλήματος και αμέσως μετά θα παρουσιάσουμε την γενική μορφή της στην περίπτωση που δεν έχουμε καμία επιπλέον πληροφορία για την φύση της εξάρτησης και της ετερογένειας της σειράς μας. Τέλος, στην Ενότητα 5 θα αναφερθούμε στην PL ως μέθοδο συμπερασματολογίας των κατηγορικών χρονοσειρών, παραθέτοντας παράλληλα ένα απλό παράδειγμα εφαρμογής της για εξαρτημένα δεδομένα, η κατανομή των οποίων ανήκει στην εκθετική οικογένεια.

2.2 Η έννοια της Δεσμευμένης Πιθανοφάνειας

Η αβεβαιότητα ενός διαχρονικού στοχαστικού φαινομένου συνοψίζεται στο *δείγμα*. Επομένως για να έχουμε τον μέγιστο βαθμό στοχαστικής πληροφόρησης θα πρέπει να γνωρίζω την από κοινού κατανομή του δείγματος

$$f(y_1, y_2, \ldots, y_N; \boldsymbol{\theta}),$$

που σημαίνει γνώση τόσο του πιθανοθεωρητικού μοντέλου (φόρμουλα) f όσο και των παραμέτρων του εν λόγω μοντέλου, δηλαδή του $\boldsymbol{\theta}$. Αυτή όμως η επιλογή μοντέλου δεν είναι πάντα ρεαλιστική αφού μπορεί να παρουσιαστεί αδυναμία εκτίμησης των παραμέτρων λόγω του μικρού αριθμού παρατηρήσεων. Συνάμα, επειδή συχνά ο μηχανισμός τύχης ο οποίος καθορίζει την εξέλιξη του φαινομένου, δεν εμπίπτει στο πλαίσιο του τυχαίου πειράματος, δεν είναι εφικτή η αναγωγή της πιθανοφάνειας στην εύχρηστη σχέση

$$f(y_1, y_2, \ldots, y_N; \boldsymbol{\theta}) \overset{i.i.d}{=} \prod_{t=1}^{N} f(y_t; \boldsymbol{\theta}).$$

Σε πρώτη φάση για την αντιμετώπιση του προβλήματος μοντελοποίησης διαχρονικών μη ντετερμινιστικών φαινομένων τα οποία επιδεικνύουν *εξάρτηση* ή *ετερογένεια* χρη-

σιμοποιήθηκε η προσέγγιση της *Conditional Likelihood* (*Andersen* (1973)). *Η δεδομένη μέθοδος πιθανοφάνειας στηρίχτηκε αποκλειστικά στην δειγματοληπτική διαδρομή της μεταβλητής ενδιαφέροντος* Y_t. Η αξία της δεσμευμένης πιθανοφάνειας καθίσταται ιδιαίτερη σημαντική μόνον εφόσον η πραγματοποίηση του δείγματος

$$y_1, y_2, \ldots, y_N$$

παρέχει *επαρκή στοιχεία για την φύση της εξάρτησης και της ετερογένειας* του αόρατου μηχανισμού, δηλαδή της στοχαστικής ανέλιξης $\{Y_t\}, t = 1, 2, \ldots$, η οποία γεννά τα παρατηρούμενα δεδομένα. Πράγματι, για διαχρονικά φαινόμενα τα οποία παρουσιάζουν εξάρτηση ή ετερογένεια, για την από κοινού κατανομή του δείγματος χρησιμοποιείται η ακόλουθη διάσπαση

$$\begin{aligned} f(y_1, y_2, \ldots, y_N; \boldsymbol{\theta}) = & f_N(y_N \mid y_{N-1}, \ldots, y_1; \boldsymbol{\phi_N}) \cdot \\ & \cdot f_{N-1}(y_{N-1} \mid y_{N-2}, \ldots, y_1; \boldsymbol{\phi_{N-1}}) \ldots f_2(y_2 \mid y_1; \boldsymbol{\phi_2}) \cdot f_1(y_1; \boldsymbol{\phi_1}) \end{aligned} \tag{2.1}$$

άρα

$$f(y_1, y_2, \ldots, y_N; \boldsymbol{\theta}) \overset{non-i.i.d}{=} f_1(y_1; \boldsymbol{\phi_1}) \cdot \prod_{t=2}^{N} f(y_t \mid y_{t-1}, \ldots, y_1; \boldsymbol{\phi_t}),$$

όπου $\boldsymbol{\phi_1}$ είναι το διάνυσμα των παραμέτρων που αντιστοιχεί στην περιθώρια κατανομή f_1 της τυχαίας μεταβλητής y_1, ενώ $\boldsymbol{\phi_t}$ για $t = 2, 3, \ldots, N$, είναι τα παραμετρικά διανύσματα που αντιστοιχούν στις δεσμευμένη πυκνότητα πιθανότητας της τυχαίας μεταβλητής y_t δοθέντος του παρελθόντος $\{y_{t-1}, y_{t-2}, \ldots, y_1\}$. Η παραπάνω όμως διάσπαση της N-διάστατης από κοινού κατανομής σε γινόμενο N-1 δεσμευμένων επί μια οριακή εγκυμονεί προβλήματα. Το σπουδαιότερο είναι ότι ο αριθμός των δεσμευμένων μεταβλητών δεν είναι σταθερός για όλες τις δεσμευμένες κατανομές. Επιπλέον οι άγνωστες παράμετροι διαφέρουν για κάθε κατανομή. Εάν όμως γνωρίζαμε ότι η στοχαστική διαδικασία που γεννά το φαινόμενο ήταν *Markov* τάξης 1 (M1) και αυστηρά στάσιμη τότε για την πιθανοφάνεια του δείγματος θα είχαμε

$$f(y_1, y_2, \ldots, y_N; \boldsymbol{\theta}) \overset{M-St}{=} f_1(y_1; \boldsymbol{\phi_1}) \cdot \prod_{t=2}^{N} f(y_t \mid y_{t-1}; \boldsymbol{\phi}).$$

Αγνοώντας την περιθώρια κατανομή $f_1(y_1; \boldsymbol{\phi_1})$ για την μοντελοποίηση του φαινομένου μπορούμε να στηριχτούμε στην δεσμευμένη κατανομή $f(y_t \mid y_{t-1}; \boldsymbol{\phi})$ η οποία αποτελεί πλέον τον φορέα της εξάρτησης άρα και της στοχαστικής πληροφόρησης για την μεταβλητή ενδιαφέροντος Y_t.

Σκοπός της μοντελοποίησης είναι για κάθε χρονική στιγμή t να προβλέψουμε το αποτέλεσμα του μηχανισμού τύχης, δηλαδή την τιμή της τυχαίας μεταβλητής Y_t, πρίν

το φαινόμενο ολοκληρωθεί. Η πρόβλεψη μας επιθυμούμε να γίνεται με όσο το δυνατόν μεγαλύτερη ακρίβεια αξιοποιώντας την πληροφορία που παρέχει η φύση της διαχρονικής εξάρτησης των δεδομένων. Έτσι για το δεδομένο παράδειγμα που υποθέσαμε ότι η χρονοσειρά μας παρουσιάζει την ιδιότητα (M1) και είναι αυστηρά στάσιμη, η αναμενόμενη τιμή της μεταβλητής ενδιαφέροντος κάποια χρονική στιγμή υπολογίζεται βάσει της πληροφορίας που παρέχει η τιμή της την αμέσως προηγούμενη χρονική στιγμή. Επομένως η μοντελοποίηση του φαινομένου μας θα στηριχτεί στις ακόλουθες δεσμευμένες ροπές[1]

$$E(Y_t \mid \wp), \quad Var(Y_t \mid \wp),$$

όπου $\wp$ αποτελεί το Σύνολο Πληροφορίας (*Information Set*) του φαινομένου και το οποίο στην προκειμένη περίπτωση, επειδή η ανέλιξη είναι στάσιμη και M1, ισούται με $\wp = \{Y_{t-1} = y_{t-1}\}$. Αναλόγως λοιπόν με την επιλογή των δεσμευμένων ροπών που είναι οι φορείς της πληροφορίας ή της εξάρτησης για την στοχαστική διαδικασία $\{Y_t\}, t = 1, 2, \ldots$, θα οδηγηθούμε σε μια σειρά μοντέλων τα οποία ονομάζονται *Μοντέλα Παλινδρόμησης*. Ο ενδιαφερόμενος αναγνώστης μπορεί να βρεί παραδείγματα συμπερασματολογίας βάσει της *Conditional Likelihood* στους *Kalbfleisch and Sprott* (1970).

2.3 Συμπερασματολογία στην περίπτωση ύπαρξης εξωγενών μεταβλητών

Μέχρι τώρα αναφερθήκαμε σε διαχρονικά φαινόμενα τα οποία μοντελοποιούνται μέσω μιας μονοδιάστατης στοχαστικής ανέλιξης. Στα μοντέλα παλινδρόμησης λοιπόν η πληροφορία για την εξέλιξη του φαινομένου παρεχόταν αποκλειστικά από την ιστορία-παρελθόν της μεταβλητής ενδιαφέροντος Y_t. Είναι λογικό να αναρωτηθούμε, μήπως υπάρχουν και άλλες μεταβλητές σχετικές με το φαινόμενο που θα αυξήσουν το πεδίο πληροφοριών για την Y_t, μειώνοντας περισσότερο την αβεβαιότητα μας για την μελλοντική εξέλιξη του δυναμικού συστήματος. Διευρύνουμε λοιπόν το αναλυτικό μας πλαίσιο ώστε να μπορεί να συμπεριλάβει όχι μόνο τις πληροφορίες που προέρχονται από την ιστορία του ίδιου του φαινομένου αλλά και πληροφορίες από άλλες μεταβλητές σχετικές με το φαινόμενο. Αυτό επιτυγχάνεται επεκτείνοντας την έννοια της στοχαστικής ανέλιξης σε αυτή της διανυσματικής στοχαστικής ανέλιξης. Έστω το τυχαίο

[1] Λαμβάνοντας υπόψη ότι $Var(Y) = E(X^2) - [E(X)]^2$ και κατ' επέκταση $Var(Y_t \mid \wp) = E(X^2 \mid \wp) - [E(X \mid \wp)]^2$.

διάνυσμα

$$\mathbf{U_t} = \begin{pmatrix} Y_t \\ X_t \end{pmatrix}.$$

Μια διανυσματική στοχαστική ανέλιξη είναι μια ακολουθία τέτοιων διανυσμάτων, δηλαδή $\mathbf{U}_1, \mathbf{U}_2, \ldots$. Πλέον η κατανομή του $\mathbf{U}$ σε μια χρονική στιγμή t θα είναι η $f_t(\boldsymbol{u}_t; \boldsymbol{\theta_t})$ η οποία δεν είναι μονοδιάστατη αλλά είναι η από κοινού διδιάστατη συνάρτηση πυκνότητας πιθανότητας των y_t και x_t δηλαδή η $f_t(y_t, x_t; \boldsymbol{\theta_t})$. Έστω η δειγματοληπτική διαδρομή $\mathbf{U}_1, \mathbf{U}_2, \ldots, \mathbf{U}_N$ της διανυσματικής στοχαστικής ανέλιξης $\{\mathbf{U}_t\}, t = 1, 2, \ldots$. Αν η μελέτη των δεδομένων μας υποδείξει ότι ο μηχανισμός τύχης που παράγει το φαινόμενο, επιδεικνύει ταυτονομία και ανεξαρτησία, τότε η από κοινού κατανομή του δείγματος γράφεται

$$f(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N; \boldsymbol{\theta}) \overset{i.i.d}{=} \prod_{t=1}^{N} f(\boldsymbol{u}_t; \boldsymbol{\phi}),$$

όπου $\boldsymbol{\phi}$ είναι το διάνυσμα των παραμέτρων της κατανομής των ανεξάρτητων και ταυτόνομων τυχαίων μεταβλητών $y_t, t = 1, 2, \ldots, N$. Συχνά όμως, όπως προαναφέρθηκε και στην περίπτωση της μονοδιάστατης στοχαστικής ανέλιξης $\{Y_t\}$, παρουσιάζεται διαχρονική εξάρτηση ή ετερογένεια με αποτέλεσμα η διαδικασία μοντελοποίησης να είναι επίπονη. Την λύση στην αναζήτηση στατιστικού μοντέλου το οποίο με οικονομικό τρόπο να περιγράφει τα δεδομένα, μας την παρέχει σε πρώτη φάση ο προσδιορισμός της φύσης της εξάρτησης και της ετερογένειας. Αξίζει να σημειώσουμε ότι οι ιδιότητες που συναντήσαμε για την μονοδιάστατη στοχαστική διαδικασία (π.χ *Markov*, στασιμότητα κ.τ.λ) μπορούν εύκολα να μεταφερθούν στο διευρυμένο πλαίσιο των διανυσματικών στοχαστικών ανελίξεων. Έτσι για παράδειγμα αν με κάποιο τρόπο διαπιστώσουμε ότι η διανυσματική στοχαστική ανέλιξη $\{\mathbf{U}_t\}, t = 1, 2, \ldots$, παρουσιάζει τις ιδιότητες M1 και είναι αυστηρά στάσιμη, τότε η από κοινού κατανομή του δείγματος που αρχικά γράφεται

$$f(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N; \boldsymbol{\theta}) \overset{non-i.i.d}{=} f_1(\boldsymbol{u}_1; \boldsymbol{\phi_1}) \cdot \prod_{t=2}^{N} f_t(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}, \ldots, \boldsymbol{u}_1; \boldsymbol{\phi_t})$$

τώρα μετατρέπεται στην

$$f(\boldsymbol{u}_1, \boldsymbol{u}_2, \ldots, \boldsymbol{u}_N; \boldsymbol{\theta}) \overset{M-St}{=} f_1(\boldsymbol{u}_1; \boldsymbol{\theta_1}) \cdot \prod_{t=2}^{N} f(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}; \boldsymbol{\phi}).$$

Αγνοώντας την $f_1(\boldsymbol{u}_1; \boldsymbol{\theta_1})$ με καταλληλη επιλογή αρχικών συνθηκών γίνεται φανερό πως θα χτίσουμε το μοντέλο μας πάνω στην δεσμευμένη κατανομή

$$f(\boldsymbol{u}_t \mid \boldsymbol{u}_{t-1}; \boldsymbol{\phi}) \equiv f(y_t, x_t \mid y_{t-1}, x_{t-1}; \boldsymbol{\phi}).$$

Στο σημείο αυτό αν θεωρήσουμε ότι η μεταβλητή ενδιαφέροντος είναι η Y_t, μπορούμε να σταματήσουμε την συμμετρική αντιμετώπιση των Y_t και X_t. Επομένως μπορούμε να γράψουμε

$$f(y_t, x_t \mid y_{t-1}, x_{t-1}; \boldsymbol{\phi}) = f(y_t \mid x_t, y_{t-1}, x_{t-1}; \boldsymbol{\lambda_1}) \cdot f(x_t \mid y_{t-1}, x_{t-1}; \boldsymbol{\lambda_2}),$$

όπου $\boldsymbol{\lambda_1}$ και $\boldsymbol{\lambda_2}$ είναι αντιστοίχως τα παραμετρικά διανύσματα των δεσμευμένων κατανομών της y_t δοθέντων των x_t, y_{t-1}, x_{t-1} και της x_t δοθέντων των y_{t-1}, x_{t-1}. Έτσι το τελικό πιθανοτικό μοντέλο για την Y_t θα είναι το $f(y_t \mid x_t, y_{t-1}, x_{t-1}; \boldsymbol{\lambda_1})$ και θα χτιστεί βάση των δεσμευμένων ροπών

$$E(Y_t \mid X_t, Y_{t-1}, X_{t-1}), \quad Var(Y_t \mid X_t, Y_{t-1}, X_{t-1}). \tag{2.2}$$

Γίνεται λοιπόν φανερό ότι αναλόγως με την επιλογή των δεσμευμένων ροπών, που είναι οι φορείς της πληροφορίας ή της εξάρτησης για την στοχαστική διαδικασία $\{Y_t\}, t = 1, 2, \ldots$, θα οδηγηθούμε σε μια σειρά μοντέλων τα οποία ονομάζονται *Δυναμικά Μοντέλα Παλινδρόμησης*.

Σύμφωνα με τα προαναφερθέντα παρατηρούμε ότι αγνοήσαμε την πληροφορία της $f(x_t \mid y_{t-1}, x_{t-1}; \boldsymbol{\lambda_2})$. Η αγνόηση του δεδομένου όρου σε πρώτη φάση μπορεί να δικαιολογηθεί, χωρίς να συμβαίνει πάντα, από το ότι η μεταβλητή X_t είναι εξωγενής. Αυτό σημαίνει ότι η X_t γεννάται έξω από το σύστημα ανάλυσης της Y_t και κατα συνέπεια ο γεννεσιουργός μηχανισμός της μας είναι αδιάφορος.

Στο παράδειγμα που μόλις αναφέραμε, στο οποίο η από κοινού κατανομή του δείγματος γράφτηκε στην μορφή

$$f(y_1, x_1, \ldots, y_N, x_N; \boldsymbol{\theta}) \overset{M=St}{=}$$

$$f_1(y_1, x_1; \boldsymbol{\phi_1}) \cdot \prod_{t=2}^{N} f(x_t \mid y_{t-1}; \boldsymbol{\lambda_2}) \cdot \prod_{t=1}^{N} f(y_t \mid x_t, y_{t-1}, x_{t-1}; \boldsymbol{\lambda_1}),$$

θεωρήσαμε ως μεταβλητή ενδιαφέροντος την Y_t και για την συμπερασματολογία βασιστήκαμε στον όρο

$$\prod_{t=1}^{N} f(y_t \mid x_t, y_{t-1}, x_{t-1}; \boldsymbol{\lambda_1}),$$

αγνοώντας την χρονοεξαρτώμενη X_t. Η διαδικασία που μόλις περιγράψαμε αποτελεί μια ειδική περίπτωση χρήσης της *Μερικής Πιθανοφάνειας* (*Partial Likelihood*), η οποία αποτελεί σημαντική μεθοδολογική προσέγγιση στην στατιστική αναλυση χρονοσειρών. Η συγκεκριμένη περιοχή έχει μελετηθεί διεξοδικά από πολλούς στατιστικούς, όπως ο

Bhat (1974) και *Crowder* (1976). Η αξία της δεδομένης μεθόδου έγκειται στο γεγονός ότι δεν απαιτεί την ισχύ υποθέσεων, που χαρακτηρίζουν τις στοχαστικές ανελίξεις όπως π.χ η στασιμότητα και η ιδιότητα *Markov*. Εξάλλου στην πράξη η φύση της διαχρονικής εξάρτησης ή της ετερογένειας είναι πολύπλοκη και δύσκολα προσδιορίζεται. Έτσι, για το προηγούμενο παράδειγμα ακόμη και όταν δεν γνωρίζαμε ότι ισχύει η στασιμότητα και η ιδιότητα M1, η από κοινού κατανομή του δείγματος μπορεί να γραφεί

$$f(y_1, x_1, \ldots, y_N, x_N; \boldsymbol{\theta}) = f_1(x_1; \boldsymbol{\theta_1}) \cdot \prod_{t=2}^{N} f(x_t \mid d_t; \boldsymbol{\kappa_t}) \cdot \prod_{t=1}^{N} f(y_t \mid c_t; \boldsymbol{\mu_t}),$$

όπου $d_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1})$ και $c_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1}, x_t)$ ενώ $\boldsymbol{\kappa_t}$ και $\boldsymbol{\mu_t}$ είναι αντιστοίχως τα διανύσματα των παραμέτρων των δεσμευμένων κατανομών των x_t και y_t. Το δεύτερο γινόμενο της προηγούμενης σχέσης συνιστά την μερική πιθανοφάνεια του δείγματος η οποία με κατάλληλες τροποποιήσεις, που θα δούμε στη συνέχεια, καθίσταται κατάλληλη για συμπερασματολογία. Ακόμη, σε αναλογία με τα προαναφερθέντα, το στατιστικό μοντέλο θα χτιστεί σε όρους των δεσμευμένων ροπών $E(Y_t \mid \wp)$, $Var(Y_t \mid \wp)$ με $\wp \equiv c_t$. Σίγουρα πληροφορία για το $\boldsymbol{\theta}$ υπάρχει και στο πρώτο γινόμενο και επομένως γεννάται το ερώτημα τι θα συμβεί αν παραληφθεί αυτός ο παράγοντας. Αποδεικνύεται ότι κάτω από κάποιες λογικές συνθήκες η απώλεια πληροφορίας για το $\boldsymbol{\theta}$ είναι μικρή και το αντάλλαγμα αυτής της αγνόησης (ακόμη και αν η X_t είναι εξωγενής) είναι ότι ο εναπομείναντας παράγοντας είναι μια χρήσιμη μορφή συνάρτησης πιθανοφάνειας(*Wong* (1986)).

2.4 Εφαρμογή της PL στην μοντελοποίηση

Μέχρι τώρα έγινε ένα πρώτο βήμα στην εξήγηση της συνάρτησης μερικής πιθανοφάνειας χώρις όμως να αναδειχθεί ξεκάθαρα η λειτουργικότητά της στην μοντελοποίηση χρονολογικών σειρών στις οποίες τα δεδομένα είναι εξαρτημένα και οι συμμεταβλητές και ίσως τα βοηθητικά δεδομένα είναι τυχαία και χρονοεξαρτώμενα. Θα πρέπει να τονίσουμε ότι η μέθοδος της μερικής πιθανοφάνειας οφείλει την χρηστικότητα της σε μία «έξυπνη δέσμευση» (το σημείο αυτό θα γίνει κατανοητό στην ανάλυση που θα ακολουθήσει). Για να διευκρινιστούν τα προαναφερθέντα δίνεται το ακόλουθο παράδειγμα.

Στην περίπτωση της διανυσματικής στοχαστικής ανέλιξης $\{\boldsymbol{U}_t\}, t = 1, 2, \ldots$, που παρουσιάσαμε και για την οποία η μεταβλητή ενδιαφέροντος Y_t έχει τις ιδιότητες της στασιμότητας και της M1, καταλήξαμε ότι η συμπερασματολογία θα βασιστεί στο γινόμενο $\prod_{t=1}^{N} f(y_t \mid x_t, y_{t-1}, x_{t-1}; \boldsymbol{\lambda_1})$ και επομένως το αντίστοιχο δυναμικό μοντέλο

θα χτιστεί βάσει των δεσμευμένων ροπών της σχέσης (2.2). Υποθέτοντας ότι ισχύει

$$E(Y_t \mid \wp) = \beta_0 + \beta_1 \cdot X_t + \beta_2 \cdot Y_{t-1} + \beta_3 \cdot X_{t-1}$$

και

$$Var(Y_t \mid \wp) = \sigma^2 \quad \text{(ομοσκεδαστικότητα)}$$

καταλήγουμε στο ομοσκεδαστικό δυναμικό μοντέλο παλινδρόμησης

$$\begin{cases} Y_t = \beta_0 + \beta_1 \cdot X_t + \beta_2 \cdot Y_{t-1} + \beta_3 \cdot X_{t-1} \\ Var(Y_t \mid \wp) = \sigma^2 \end{cases} \tag{2.3}$$

από το οποίο γίνεται φανερό ότι για το σταθερό διάνυσμα $\boldsymbol{\lambda_1}$ ισχύει $\boldsymbol{\lambda_1} = (\beta_0, \beta_1, \beta_2, \beta_3, \sigma^2)$. Το παραμετρικό διάνυσμα του μοντέλου θα εκτιμηθεί μέσω της συνάρτησης μερικής πιθανοφάνειας η οποία γράφεται

$$PL(\boldsymbol{\lambda_1}; y_1, y_2, ..., y_N) = \prod_{t=1}^{N} f(y_t \mid x_t, y_{t-1}, x_{t-1}; \boldsymbol{\lambda_1})$$

ώστε να τονιστεί ότι για την συμπερασματολογία στηριζόμαστε στις δεσμευμένες κατανομές της μεταβλητής ενδιαφέροντος Y_t για κάθε t.

Στο σημείο αυτό γεννάται ένα σημαντικό ερώτημα. Τι θα γινόταν αν δεν γνωρίζαμε ότι για την αρχική μας διανυσματική στοχαστική ανέλιξη ίσχυε η Μαρκοβιανή ιδιότητα και η στασιμότητα; Σε αυτή την περίπτωση η μερική πιθανοφάνεια θα έχει την μορφή

$$PL(\boldsymbol{\mu_t}; y_1, y_2, ..., y_N) = \prod_{t=1}^{N} f(y_t \mid c_t; \boldsymbol{\mu_t}),$$

όπου $c_t = (y_1, x_1, \ldots, y_{t-1}, x_{t-1}, x_t)$. Στο σημείο αυτό φαίνεται πως η προσέγγιση της μερικής πιθανοφάνειας, χωρίς την πληροφορία για τα χαρακτηριστικά της εξάρτησης και της ετερογένειας, καθίσταται προβληματική για την συμπερασματολογία. Συγκεκριμένα για κάθε χρονική στιγμή η παραπάνω δεσμευμένη κατανομή της μεταβλητής ενδιαφέροντος Y_t, εξαρτάται από ένα διαφορετικό σύνολο πληροφοριών το οποίο όσο εξελίσσεται το φαινόμενο διευρύνεται. Αυτό έχει σαν αποτέλεσμα όσο αυξάνεται το μέγεθος N της δειγματοληπτικής διαδρομής να αυξάνεται και το πλήθος των παραμέτρων. Κατά συνέπεια αντί να λαμβάνουμε περισσότερη πληροφορία για ένα σύνολο σταθερών παραμέτρων δεχόμαστε πληροφορία για ένα αυξανόμενο αριθμό παραμέτρων (η διάσταση του παραμετρικού διανύσματος $\boldsymbol{\mu_t}$ αυξάνει καθώς περνάει ο χρόνος), γεγονός που συντελεί στην δυσκολία μοντελοποίησης. Για να ξεπεραστεί το πρόβλημα η

προαναφερθείσα πιθανοφάνεια τροποποιήθηκε και η δεσμευμένη κατανομή $f(y_t \mid c_t; \boldsymbol{\mu_t})$ για κάθε χρονική στιγμή αντικαταστήθηκε από την δεσμευμένη κατανομή

$$f(y_t \mid \mathcal{F}_{t-1}).$$

Το $\mathcal{F}_{t-1}$ είναι η σ-άλγεβρα η οποία γεννάται από παρελθούσες τιμές της μεταβλητής ενδιαφέροντος ή και ακόμη παροντικές τιμές (όταν είναι γνωστές) των επεξηγηματικών μεταβλητών (*covariates*). *Αναλυτικότερα θα λέγαμε πως η ιστορία* $\mathcal{F}_{t-1}$ *εμπεριέχει οτιδήποτε είναι γνωστό στον παρατηρητή του φαινομένου την χρονική στιγμή* $t-1$*, και επομένως είναι δυνατόν να περιέχει και τιμές της* X_t *ή κάποιας βοηθητικής μεταβλητής* W_t *εφόσον αυτές είναι εκ των προτέρων γνωστές.* Με σκοπό να τυποποιήσουμε τα προαναφερθέντα, για κάθε χρονική στιγμή t θεωρούμε πως η τιμή της μεταβλητής ενδιαφέροντος Y_t διαμορφώνεται από το p-διάστατο διάνυσμα

$$\mathbf{Z}_{t-1} = (Z_{(t-1)1}, Z_{(t-1)2}, \ldots, Z_{(t-1)p})', \quad t = 1, 2, \ldots, N, \tag{2.4}$$

το οποίο περιέχει παρελθούσες ή και παροντικές ερμηνευτικές μεταβλητές. Ουσιαστικά το $\mathbf{Z}_{t-1}$ για κάθε χρόνο t συνοψίζει την προγενέστερη εξέλιξη του συστήματος όπως αυτή διαμορφώνεται μέχρι τον χρόνο $t-1$. Ενδεικτικά, το $\mathbf{Z}_{t-1}$ μπορεί να έχει την μορφή

$$\mathbf{Z}_{t-1} = (Y_{t-1}, Y_{t-2}, X_t, W_t)'$$

όπου Y_{t-1}, Y_{t-2} είναι οι χρονικές υστερήσεις $1^{\eta\varsigma}$ και $2^{\eta\varsigma}$ τάξης της Y_t ενώ X_t και W_t είναι δυο ντετερμινιστικές μεταβλητές των οποίων οι τιμές για τον χρόνο t είναι ήδη γνωστές από τον χρόνο $t-1$. Σύμφωνα λοιπόν με την συζήτηση που προηγήθηκε για κάθε χρονική στιγμή εξέλιξης του συστήματος πέρα από την χρονοσειρά $\{Y_t\}$ που ονομάζεται *απόκριση* (*response*), θεωρούμε και την διανυσματική στοχαστική ανέλιξη $\{\mathbf{Z}_{t-1}\}$ που καλείται *συμμεταβλητή διαδικασία* (*covariate process*). Έτσι για την σ-άλγεβρα $\mathcal{F}_{t-1}$ θα ισχύει

$$\mathcal{F}_{t-1} = \sigma\{Y_{t-1}, Y_{t-2}, \ldots, X_t, W_t, \ldots, \mathbf{Z}_{t-1}, \mathbf{Z}_{t-2}, \ldots, \}.$$

Στον παραπάνω συμβολισμό πρέπει να σημειώσουμε ότι το $\mathbf{Z}_{t-1}$ εμπεριέχει ήδη παρελθούσες τιμές της αποκριτικής μεταβλητής. Αξίζει να υπενθυμίσουμε ότι η ιστορία του φαινομένου κάθε χρονική στιγμή παρατήρησης είναι ένα υπερσύνολο των αντίστοιχων ιστοριών όπως αυτές διαμορφώνονται σε προηγούμενες χρονικές στιγμές μελέτης του δυναμικού συστήματος. Έτσι για το διαχρονικό φαινόμενο θα έχουμε μία αύξουσα ακολουθία σ-αλγεβρών $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2 \ldots$ για τις αντίστοιχες στιγμές παρατήρησης

$t = 0, 1, 2, \ldots$. Επομένως αν την χρονική στιγμή $t-1$ έχουμε παρατηρήσει το φαινόμενο και έχουμε καταγράψει την τιμή Y_{t-1}, τότε για να προβλέψουμε την μελλοντική τιμή Y_t θα λάβουμε υπόψη την σ-άλγεβρα $\mathcal{F}_{t-1}$ η οποία εμπεριέχει την ιστορία του φαινομένου από την στιγμή που ξεκινήσαμε να το καταγράφουμε. Γίνεται φανερό πως πλέον το *Information Set* είναι το $\wp \equiv \mathcal{F}_{t-1}$ και σε αντιστοιχία με τα μοντέλα παλινδρόμησης και τα δυναμικά μοντέλα παλινδρόμησης το στατιστικό μοντέλο που θα δημιουργήσουμε για την μεταβλητή ενδιαφέροντος Y_t θα στηριχτεί στις δεσμευμένες ροπές

$$\mu_t = E[Y_t \mid \mathcal{F}_{t-1}] \quad \text{και} \quad \sigma_t^2 = Var[Y_t \mid \mathcal{F}_{t-1}].$$

Αφού κάθε χρονική στιγμή το *Information Set* είναι το $\mathcal{F}_{t-1}$ ο παρατηρητής μπορεί να επιλέξει τις δεσμευμένες ροπές να εξαρτώνται από ένα διάνυσμα *συγκεκριμένων* μεταβλητών $\boldsymbol{Z}_{t-1}$ (των οποίων οι τιμές θα διαφοροποιούνται ανα χρονική περίοδο), και το οποίο κατα την κρίση του θα εμπεριέχει την απαιτούμενη πληροφορία για την περιγραφή του φαινομένου, επιτρέποντας *οικονομική μοντελοποίηση* (*parsimonious modeling*) και όσο το δυνατόν ακριβέστερες προβλέψεις. Έτσι η δεσμευμένη κατανομή της Y_t για κάθε χρονική στιγμή ως προς $\mathcal{F}_{t-1}$, δηλαδή η $f(y_t \mid \mathcal{F}_{t-1})$ μπορεί να γραφεί ως $f(y_t; \boldsymbol{\beta})$, οπού $\boldsymbol{\beta}$ είναι το *σταθερό* διάνυσμα των παραμέτρων σύμφωνα με το οποίο για κάθε χρονική στιγμή οι δεσμευμένες ροπές συνδέονται με το διάνυσμα $\boldsymbol{Z}_{t-1}$. Η PL τα τελευταία χρόνια διαδραματίζει σημαντικό ρόλο και σε άλλα επιστημονικά πεδία που σχετίζονται με τον χώρο της Στατιστικής και των Πιθανοτήτων. Χαρακτηριστικά αναφέρουμε ότι αποτελεί βασικό εργαλείο μοντελοποίησης διαδικασιών μετάδοσης σήματος (*signal processing*) (*Adali* και *Ni* (2003)).

2.5 Αξιοποίηση της θεωρίας της PL για την Συμπερασματολογία των Κατηγορικών Χρονοσειρών

Η θεωρία της PL διαδραματίζει καθοριστικό ρόλο στην μοντελοποίηση των κατηγορικών χρονοσειρών. Αυτό συμβαίνει διότι όπως είδαμε στις προηγούμενες ενότητες η δεδομένη μέθοδος δεν προυποθέτει ούτε στασιμότητα ούτε την Μαρκοβιανή ιδιότητα. Με τον τρόπο αυτό οι στατιστικοί κατόρθωσαν να ξεπεράσουν τα προβλήματα που παρουσίαζε η ανάλυση των ποιοτικών σειρών μέσω των Μαρκοβιανών Μοντέλων, που αναφέραμε στην ενότητα 2.1 . Στα πλαίσια της θεωρίας των GLM και της εκθετικής οικογένειας κατανομών η τεχνική της PL είναι αρκετά ελκυστική για την στατιστική ανάλυση των κατηγορικών χρονολογικών σειρών, επιτρέποντας μια συνεχή συμπερα-

σματολογία με βάση ένα «φίλτρο» το οποίο γεννάται από οτιδήποτε είναι γνωστό κατά τον χρόνο της παρατήρησης (*Fokianos* και *Kedem* (2001)).

Στα κεφάλαια που θα ακολουθήσουν θα εφαρμόσουμε την θεωρία της PL για διάφορους τύπους ποιοτικών σειρών. Αναλυτικότερα στο Κεφάλαιο 3 θα παρουσιάσουμε την μέθοδο της PL για δίτιμες σειρές ενώ στο Κεφάλαιο 5 θα αναφερθούμε διεξοδικά στην συμπερασματολογία για ονοματικές και διατάξιμες κατηγορικές χρονοσειρές.

Στο σημείο αυτό παραθέτουμε ένα απλό παράδειγμα εφαρμογής της PL στα πλαίσια της εκθετικής οικογένειας κατανομών. Έστω η δίτιμη χρονοσειρά $\{Y_t\}, t = 1, 2, \ldots, N$. Θεωρώντας ότι για κάθε στιγμή καταγραφής του φαινομένου η τιμή της απόκρισης εξαρτάται από δυο υστερήσεις της, από έναν βοηθητικό παράγοντα X, του οποίου η τιμή κατά τον χρόνο t είναι γνωστή από τον $t-1$, καθώς και από ένα όρο περιοδικότητας 12 μονάδων, το διάνυσμα των συμμεταβλητών $\mathbf{Z}_{t-1}$ θα πάρει την μορφή

$$\boldsymbol{Z}_{t-1} = (1, Y_{t-1}, Y_{t-2}, X_t, \cos\big(\frac{2\pi t}{12}\big))'.$$

Έτσι για το εν λόγω παράδειγμα η δεσμευμένη μέση τιμή $E[Y_t \mid \mathcal{F}_{t-1}]$ θα συνδέεται μέσω κάποιας γνωστής συνάρτησης g με τον γραμμικό συνδυασμό

$$\beta_1 + \beta_2 Y_{t-1} + \beta_3 Y_{t-2} + \beta_4 X_t + \beta_5 \cos\big(\frac{2\pi t}{12}\big).$$

Επομένως το στατιστικό μοντέλο που θα χρησιμοποιήσουμε για συμπερασματολογία θα έχει την μορφή

$$g(\mu_t(\boldsymbol{\beta})) = \beta_1 + \beta_2 Y_{t-1} + \beta_3 Y_{t-2} + \beta_4 X_t + \beta_5 \cos\big(\frac{2\pi t}{12}\big)$$

όπου $\boldsymbol{\beta} = (\beta_1, \beta_2, \beta_3, \beta_4, \beta_5)'$. Πλέον, σύμφωνα με το δεδομένο στατιστικό μοντέλο, η εκτίμηση της αναμενόμενης τιμής της μεταβλητής ενδιαφέροντος την χρονική στιγμή t ανάγεται στην εκτίμηση του σταθερού *διανύσματος* $\boldsymbol{\beta}$, δηλαδή θα ισχύει

$$\hat{\mu_t} = \mu_t(\hat{\boldsymbol{\beta}}).$$

Για την εκτίμηση του σταθερού διανύσματος $\boldsymbol{\beta}$ θα χρησιμοποιηθεί η μερική πιθανοφάνεια της δειγματοληπτικής διαδρομής, που αφορά την μεταβλητή ενδιαφέροντος Y_t, και η οποία στην τελική της μορφή δίνεται από

$$PL(\boldsymbol{\beta}; y_1, y_2, \ldots, y_N) = \prod_{t=1}^{N} f(y_t; \boldsymbol{\beta}),$$

όπου υπενθυμίζουμε ότι $f(y_t; \boldsymbol{\beta}) \equiv f(y_t \mid \mathcal{F}_{t-1})$. Στο σημείο αυτό, λαμβάνοντας υπόψη την συζήτηση που προηγήθηκε, θα δώσουμε τον γενικό ορισμό της PL (βλέπε *Fokianos and Kedem* (2001)).

Ορισμός 2.5.1 Έστω $\mathcal{F}_{t-1}$, $t = 1, 2, \ldots$, μια αύξουσα ακολουθία σ-αλγεβρών, $\mathcal{F}_0 \subset \mathcal{F}_1 \subset \mathcal{F}_2, \ldots$, και έστω $Y_1, Y_2, \ldots$ μια ακολουθία τυχαίων μεταβλητών ορισμένων σε ένα κοινό χώρο πιθανότητας έτσι ώστε η Y_t να είναι $\mathcal{F}_t$ μετρήσιμη. Συμβολίζοντας την πυκνότητα του Y_t δοθέντος $\mathcal{F}_{t-1}$, με $f_t(y_t; \boldsymbol{\theta})$, όπου $\boldsymbol{\theta} \in R^p$ είναι ένα σταθερό διάνυσμα. Τότε η συνάρτηση PL που σχετίζεται με το $\boldsymbol{\theta}, \mathcal{F}_t$ και τα δεδομένα $Y_1, Y_2, \ldots, Y_N$, δίνεται από το γινόμενο

$$PL(\boldsymbol{\theta}; y_1, y_2, \ldots, y_N) = \prod_{t=1}^{N} f(y_t; \boldsymbol{\theta}). \tag{2.5}$$

Το διάνυσμα το οποίο μεγιστοποιεί την σχέση (2.5) ονομάζεται εκτιμητής μέγιστης μερικής πιθανοφάνειας (*maximum partial likelihood estimator-MPLE*). Ο δεδομένος εκτιμητής υπό συνθήκες ομαλότητας ικανοποιεί τις ιδιότητες των συνηθισμένων εκτιμητών μέγιστης πιθανοφάνειας, για ανεξάρτητα δεδομένα, δηλαδή είναι συνεπής (*consistent*) και ασυμπτωτικά κανονικός (*asymptotically normal*). Για την δεδομένη κλάση εκτιμητών οι προαναφερθείσες ασυμπτωτικές ιδιότητες μαζί με την αποδοτικότητα (*efficiency*) έχουν μελετηθεί διεξοδικά κυρίως από τον *Wong* (1986), αλλά και από άλλους συγγραφείς όπως ο *Slud* (1982). Είναι σημαντικό να τονίσουμε ότι οι εκτιμητές $MPLE$ ικανοποιούν τις παραπάνω ασυμπτωτικές ιδιότητες εξαιτίας της άμεσης σχέσης της θεωρίας της PL με τις στοχαστικές ανελίξεις *martingale* (*Brown* (1971)). Περισσότερες πληροφορίες για τα ασυμπτωτικά αποτελέσματα της PL θα αναφέρουμε στα επόμενα κεφάλαια.

Κεφάλαιο 3

Μοντέλα Παλινδρόμησης για Δίτιμες Χρονοσειρές

3.1 Εισαγωγή

Στο δεδομένο κεφάλαιο θα ασχοληθούμε με την στατιστική ανάλυση των δίτιμων χρονοσειρών (*binary time series*). Για την συμπερασματολογία στις συγκεκριμένες σειρές καθώς και στις ποιοτικές σειρές με περισσότερα από δυο επίπεδα, θα αξιοποιηθεί η θεωρία των GLM (*McCullagh and Nelder* (1989)). Στην Ενότητα 2 του κεφαλαίου θα δώσουμε ορισμένα γενικά στοιχεία, πάνω στα οποία θα στηριχτεί η συμπερασματολογία μας. Στην ίδια παράγραφο θα ξεκινήσει και η ανάλυση των δίτιμων χρονοσειρών δίδοντας θεωρητικά αποτελέσματα και παραθέτοντας σημαντικά πεδία εφαρμογών τους. Εν συνεχεία στην 3^{η} ενότητα θα παρουσιάσουμε την λογιστική παλινδρόμηση ως μια από τις βασικότερες τεχνικές μοντελοποίησης δίτιμων πραγματοποιήσεων. Εναλλακτικά υποδείγματα θα παρουσιαστούν επιγραμματικά στην Ενότητα 4. Αμέσως μετά, στην παράγραφο 5, θα μιλήσουμε για την διαδικασία εύρεσης του $MPLE$ για δίτιμα εξαρτημένα δεδομένα. Στην Ενότητα 6 θα ορίσουμε κάποιους σημαντικούς πίνακες που συναντάμε στις κατηγορικές χρονοσειρές, τους οποίους όμως θα τους προσαρμόσουμε στις δίτιμες μεταβλητές. Η συμπερασματολογία για τις συγκεκριμένες σειρές θα ολοκληρωθεί στην Ενότητα 7 όπου θα παρουσιάσουμε τις ασυμπτωτικές ιδιότητες του $MPLE$. Τέλος, στην Ενότητα 8 θα δώσουμε μια θεωρητική εφαρμογή των προαναφερθέντων στην περίπτωση της λογιστικής παλινδρόμησης.

3.2 GLM και Εξαρτημένα Δίτιμα Δεδομένα

Για διαχρονικά στοχαστικά φαινόμενα των οποίων η εξέλιξη, για κάθε χρονική στιγμή παρατήρησης, δηλώνεται από μία ποιοτική μεταβλητή ενδιαφέροντος Y_t, διαπιστώσαμε

(βλέπε παράδειγμα, σελίδα 38 του κεφαλαίου 2) ότι για την στατιστική τους ανάλυση θα στηριχτούμε στα ακόλουθα:

(i) Στατιστικό Μοντέλο

$$g(\mu_t(\boldsymbol{\beta})) = \boldsymbol{\beta}'\mathbf{Z}_{t-1}, \quad t = 1, 2, \ldots, N,$$

το οποίο είναι GLM. Στο συγκεκριμένο υπόδειγμα η συνάρτηση g όπως είδαμε συνδέει κατά γραμμικό τρόπο την δεσμευμένη ροπή $\mu_t = E(Y_t \mid \mathcal{F}_{t-1})$ με συγκεκριμένες συμμεταβλητές $\boldsymbol{Z}_{t-1}$ μέσω ενός σταθερού διανύσματος παραμέτρων $\boldsymbol{\beta}$. Δηλαδή αν $\mathbf{Z}_{t-1} = (Z_{(t-1)1}, Z_{(t-1)2}, \ldots, Z_{(t-1)p})'$ και $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$ καταλήξαμε ότι θα ισχύει

$$g(\mu_t(\boldsymbol{\beta})) = \boldsymbol{\beta}'\boldsymbol{Z}_{t-1} = \beta_1 Z_{(t-1)1} + \beta_2 Z_{(t-1)2} + \ldots + \beta_p Z_{(t-1)p} \equiv \eta_t. \qquad (3.1)$$

(ii) Μοντέλο Πιθανότητας

$$f(y_t; \boldsymbol{\beta}) \equiv f(y_t \mid \mathcal{F}_{t-1})$$

(βλέπε ενότητα 2.5, σελίδα 39).

(iii) Δειγματικό Μοντέλο

$$\boldsymbol{Y} = (Y_1, Y_2, \ldots, Y_N)$$

που δεν είναι τυχαίο δείγμα.

Για την δεδομένη μεθοδολογική προσέγγιση, το κλασικό γραμμικό μοντέλο που αναφέρεται στην συνεχή μεταβλητή Y_t υπό την υπόθεση της κανονικότητας και της ανεξαρτησίας, αποτελεί ειδική περίπτωση με την g να ταυτίζεται με την ταυτοτική συνάρτηση. Η επιλογή της δεδομένης οικογένειας στατιστικών μοντέλων καθίσταται καθοριστική στην στατιστική ανάλυση κατηγορικών χρονοσειρών (*Categorical Time Series*) αφού εξασφαλίζει ότι τόσο οι προβλέψεις όσο και οι προσαρμοσμένες τιμές (*fitted values*), που ουσιαστικά αποτελούν και οι δύο εκτιμήσεις πιθανοτήτων «μετάβασης», θα ανήκουν στο διάστημα [0,1]. Η απλούστερη περίπτωση των κατηγορικών χρονοσειρών είναι οι δίτιμες χρονοσειρές, τις οποίες θα εξετάσουμε διεξοδικά.

Γενικά η μεθοδολογία των GLM για ανεξάρτητες παρατηρήσεις, $Y_i, i = 1, 2, \ldots, N$, οι οποίες προέρχονται από την κατανομή *Bernoulli*, έχει αναπτυχθεί έντονα (*Cox* και

Snell (1989)). Το γεγονός αυτό οφείλεται στο μεγάλο πεδίο εφαρμογών των δίτιμων δεδομένων σε πολλούς κλάδους της επιστήμης. Στην Ιατρική και την Ψυχολογία λοιπόν οι δίτιμες παρατηρήσεις, (π.χ. μια εξέταση είναι «θετική» ή «αρνητική» ως προς κάποια νόσο), είναι αρκετά διαδεδομένες. Παράλληλα η ανάγκη χρησιμοποίησης δίτιμων μεταβλητών εμφανίζεται και σε άλλες επιστημονικές περιοχές όπως είναι η Μετεωρολογία. Συχνά λοιπόν τα κλιματολογικά δεδομένα αποκτούν διαφορετικό ενδιαφέρον όταν από μια συνεχή κλίμακα μέτρησης μετατρέπονται σε δίτιμες αποκρίσεις. Παράδειγμα τέτοιας μετατροπής είναι η χρήση της δίτιμης μεταβλητής Y_t στην οποία ανατίθεται είτε η τιμή '1' αν η ημερήσια βροχόπτωση είναι μεγαλύτερη των '$5mm$' είτε η τιμή '0' αν η ποσότητα βροχής ανα ημέρα είναι μικρότερη ή ίση των '$5mm$'(*Coe* και *Stern*, (1984)).

Σκοπός της συζήτησης που θα ακολουθήσει είναι να παρουσιαστούν τα γνωστά αποτελέσματα των *GLM* για δίτιμα δεδομένα, όπως τα μοντέλα *logit* και *probit*, υπό το πρίσμα μιας δίτιμης χρονοσειράς $\{Y_t\}, t = 1, 2, \dots$, στην οποία η μεταβλητή ενδιαφέροντος, Y_t, παρουσιάζει διαχρονική εξάρτηση με χρονικές υστερήσεις της όσο και με τυχαίες χρονοεξαρτώμενες μεταβλητές ή βοηθητικές αιτιοκρατικές μεταβλητές (*Keenan*, (1982)). Σίγουρα πρέπει να ληφθεί υπόψη η δεδομένη διαχρονική εξάρτηση ως πηγή στοχαστικής πληροφόρησης και επομένως ο κλασικός συμβολισμός και ορολογία των δίτιμων *GLM* απαιτείται να τροποποιηθεί (*Fahrmeir and Tutz*, (2001), κεφάλαιο 6).

Έστω λοιπόν ένα διαχρονικό φαινόμενο στο οποίο σε κάθε επανάληψη είναι δυνατόν να συμβεί είτε «επιτυχία» είτε «αποτυχία». Ο παρατηρητής του φαινομένου μέσω των παρατηρήσεων που καταγράφει, που στην προκειμένη περίπτωση είναι μια ακολουθία με '1' και '0', προσπαθεί να κατασκευάσει ένα στατιστικό μοντέλο. Το στατιστικό μοντέλο δεν είναι τίποτε περισσότερο παρά ένα εργαλείο το οποίο με οικονομικό τρόπο επιχειρεί να περιγράψει την υποκείμενη στοχαστική διαδικασία που γέννησε την δειγματοληπτική διαδρομή (*sample path*). Αναμφισβήτητα η καλύτερη προσέγγιση στον μηχανισμό τύχης που παράγει το φαινόμενο είναι για κάθε χρονική στιγμή t η ακριβής εκτίμηση, μέσω του στατιστικού μοντέλου, της επικρατέστερης τιμής για την μεταβλητή ενδιαφέροντος, Y_t, λαμβάνοντας όμως υπόψη την ιστορία του συστήματος από την στιγμή που ξεκινήσαμε να το παρατηρούμε μέχρι και τον χρόνο $t-1$. Γίνεται λοιπόν φανερό πως για κάθε χρονική στιγμή t η κατανομή της τυχαίας μεταβλητής Y_t δοθείσης της ιστορίας $\mathcal{F}_{t-1}$ είναι *Bernoulli*. Δηλαδή για κάθε t ισχύει

$$Y_t \mid \mathcal{F}_{t-1} \sim Bernoulli(1, \pi_t(\boldsymbol{\beta})), \tag{3.2}$$

όπου $\pi_t(\boldsymbol{\beta})$ είναι η δεσμευμένη πιθανότητα επιτυχίας η οποία στην προκειμένη περίπτωση ταυτίζεται και με την δεσμευμένη μέση τιμή. Δηλαδή

$$\pi_t(\boldsymbol{\beta}) = P_{\boldsymbol{\beta}}(Y_t = 1 \mid \mathcal{F}_{t-1}) \equiv \mu_t(\boldsymbol{\beta}). \tag{3.3}$$

Τόσο στην (3.2) όσο και στην (3.3) το $\mathcal{F}_{t-1}$, όπως έχουμε ήδη αναφέρει, παριστά οτιδήποτε είναι γνωστό στον παρατηρητή την χρονική στιγμή $t-1$ τόσο για την μεταβλητή Y_t όσο και για τις χρονοεξαρτώμενες τυχαίες συμμεταβλητές. Ουσιαστικός σκοπός της δεδομένης ανάλυσης είναι να μοντελοποιήσουμε την δεσμευμένη πιθανότητα επιτυχίας (3.3) μέσω ενός γενικευμένου μοντέλου παλινδρόμησης της μορφής (3.1), το οποίο εξαρτάται από το $\boldsymbol{\beta}$ και εν συνεχεία να εκτιμήσουμε το $\boldsymbol{\beta}$ δοθείσης της δίτιμης χρονοσειράς $\{Y_t\}, t = 1, 2, \ldots, N$ και μιας συγκεκριμένης χρονοεξαρτώμενης διανυσματικής στοχαστικής ανέλιξης $\{\boldsymbol{Z}_{t-1}\}, t = 1, 2, \ldots, N$.

Πριν προχωρήσουμε την ανάλυση μας στις δίτιμες χρονοσειρές αναφέρουμε ότι τα αποτελέσματα που θα παρουσιάσουμε στηρίζονται στην θεωρία των GLM και στην Εκθετική Οικογένεια Κατανομών (E.O.K) για εξαρτημένες παρατηρήσεις (*Fokianos and Kedem* (2002)). Αρκετά γενικά αποτελέσματα της δεδομένης μεθοδολογικής προσέγγισης βρίσκονται στο Παράρτημα A1, στο οποίο συχνά θα παραπέμπεται ο ενδιαφερόμενος αναγνώστης.

Στην περίπτωση της ακολουθίας των δίτιμων εξαρτημένων παρατηρήσεων Y_t μπορούμε να πούμε πως για κάθε χρονική στιγμή t η δεσμευμένη κατανομή τους δοθέντος του παρελθόντος ανήκει στην E.O.K με $\alpha_t(\phi) = 1$ (βλέπε Παράρτημα A1). Πράγματι από την (3.2) για κάθε $t = 1, 2 \ldots, N$ έχουμε

$$f(y_t \mid \mathcal{F}_{t-1}) = P_{\boldsymbol{\beta}}(Y_t = y_t \mid \mathcal{F}_{t-1}) = [\pi_t(\boldsymbol{\beta})]^{y_t}[1 - \pi_t(\boldsymbol{\beta})]^{1-y_t}, \qquad y_t = 0, 1,$$

η οποία μετα από κατάλληλες τροποποιήσεις παίρνει την μορφή της σχέσης (A.1). Πράγματι έχουμε

$$f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}) = \exp\Big\{y_t \log\Big(\frac{\pi_t(\boldsymbol{\beta})}{1 - \pi_t(\boldsymbol{\beta})}\Big) + \log(1 - \pi_t(\boldsymbol{\beta}))\Big\} \tag{3.4}$$

με $\mu_t(\boldsymbol{\beta}) = \pi_t(\boldsymbol{\beta}), \theta_t = \log\Big(\frac{\pi_t}{1-\pi_t}\Big), b(\theta_t) = -\log(1 - \pi_t), V(\pi_t) = \pi_t(1 - \pi_t), \phi = 1, \omega_t = 1$. Επομένως για κάθε χρονική στιγμή θα ισχύει η εξίσωση (3.1) που σημαίνει ότι η συνάρτηση g της δεσμευμένης πιθανότητας επιτυχίας παλινδρομεί γραμμικά στο διάνυσμα των συμμεταβλητών $\boldsymbol{Z}_{t-1}$. Η (3.1) γράφεται ισοδύναμα ως

$$\pi_t(\boldsymbol{\beta}) = h(\eta_t) \tag{3.5}$$

με $h \equiv g^{-1}$ (η συνάρτηση h υπάρχει διότι έχουμε υποθέσει ότι η g είναι γνησίως μονότονη). Από την (3.5) είναι φανερό ότι για να λάβουμε εκτιμήσεις της π_t που ανήκουν στο διάστημα $[0,1]$ θα πρέπει για την αντίστροφη συνάρτηση της συνάρτησης σύνδεσης h (*inverse link function*) να ισχύει $h : R \to [0,1]$, λαμβάνοντας υπόψη ότι $\eta_t = \boldsymbol{\beta}' \boldsymbol{Z}_{t-1} \in R$.

Σύμφωνα με τα προαναφερθέντα για την στατιστική ανάλυση δίτιμων χρονοσειρών θα στηριχτούμε στην (3.5). Αναλόγως με την επιλογή της συνάρτησης h θα έχουμε μια σειρά μοντέλων παλινδρόμησης για την συμπερασματολογία της δεσμευμένης πιθανότητας επιτυχίας $\pi_t(\boldsymbol{\beta})$. Επειδή γνωρίζουμε ότι για την αθροιστική συνάρτηση κατανομής (*cumulative distribution function-cdf*) F_X της τυχαίας μεταβλητής X, με $F_X(x) = P(X \leq x)$ ισχύει $F_X : R \to [0,1]$, συμπεραίνουμε ότι θα επιλέξουμε την h ανάμεσα από κάποιες γνωστές *cdf*. Έτσι το μοντέλο συμπερασματολογίας για την $\pi_t(\boldsymbol{\beta})$ θα έχει την μορφή

$$P_{\boldsymbol{\beta}}(Y_t = 1 \mid \mathcal{F}_{t-1}) = F(\boldsymbol{\beta}' \boldsymbol{Z}_{t-1}) = h(\boldsymbol{\beta}' \mathbf{Z}_{t-1}). \tag{3.6}$$

3.3 Λογιστικό Μοντέλο Παλινδρόμησης

Όταν επιλέξουμε στην (3.6) την F να είναι η *cdf* της τυπικής λογιστικής κατανομής (*standard logistic distribution*), τότε προκύπτει το γενικό μοντέλο λογιστικής παλινδρόμησης (*logistic regression model*) για την π_t (*Bonney* (1987)). Αναλυτικότερα, αν F_l είναι η *cdf* της λογιστικής κατανομής με

$$h(x) \equiv F_l(x) = \frac{e^x}{1+e^x} = \frac{1}{1+e^{-x}}, \quad -\infty < x < \infty,$$

τότε για $x = \boldsymbol{\beta}' \boldsymbol{Z}_{t-1} \in R$ θα έχουμε το γενικό μοντέλο λογιστικής παλινδρόμησης

$$P_{\boldsymbol{\beta}}(Y_t = 1 \mid \mathcal{F}_{t-1}) = F_l(\boldsymbol{\beta}' \boldsymbol{Z}_{t-1}) = \frac{1}{1+\exp[-\boldsymbol{\beta}' \boldsymbol{Z}_{t-1}]}. \tag{3.7}$$

Παρατήρηση 3.3.1 *Για δίτιμες χρονοσειρές η αντίστροφη συνάρτηση της $F_l(x)$ δηλαδή η $g(x) = F_l^{-1}(x) = \log\left(\frac{x}{1-x}\right)$ ισούται με τον κανονικό σύνδεσμο. Πράγματι, σύμφωνα με τις σχέσεις (3.4) και (3.5) έχουμε*

$$b'(\theta_t) = \frac{\exp(\theta_t)}{1+\exp(\theta_t)}$$

και επομένως για τον κανονικό σύνδεσμο g σύμφωνα με την σχέση (A.6) θα έχουμε

$$g(\pi_t) = (b')^{-1}(\pi_t) = \log\left(\frac{\pi_t}{1-\pi_t}\right) = \theta_t.$$

Η προαναφερθείσα συνάρτηση σύνδεσης, που είναι και η συνηθέστερη επιλογή για δίτιμα δεδομένα, ονομάζεται *logit*. Έτσι το μοντέλο (3.7) γράφεται στην ισοδύναμη έκφραση

$$\text{logit}(\pi_t(\boldsymbol{\beta})) = \log\left(\frac{\pi_t}{1-\pi_t}\right) = \boldsymbol{\beta}'\boldsymbol{Z}_{t-1}. \tag{3.8}$$

3.4 Εναλλακτικά Μοντέλα

Άλλη επιλογή για την συνάρτηση h αποτελεί η *extreme value distribution* με $F(x) = 1 - \exp(-\exp(x))$ όπου $x \in R$. Έτσι για $x = \boldsymbol{\beta}'\boldsymbol{Z}_{t-1}$ το αντίστοιχο μοντέλο παλινδρόμησης για την δεσμευμένη πιθανότητα επιτυχίας θα είναι

$$P_{\boldsymbol{\beta}}(Y_t = 1 \mid \mathcal{F}_{t-1}) = 1 - \exp(-\exp(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})) \tag{3.9}$$

με αντίστοιχη συνάρτηση σύνδεσης την

$$\log\{-\log(1-\pi_t(\boldsymbol{\beta}))\} = \boldsymbol{\beta}'\boldsymbol{Z}_{t-1}, \tag{3.10}$$

γνωστή ως *complementary log-log*.

Ακόμη, μια συχνή επιλογή της συνάρτησης h αποτελεί η *cdf* της τυπικής κανονικής κατανομής Φ. Το αντίστοιχο μοντέλο παλινδρόμησης που προκύπτει για την $\pi_t(\boldsymbol{\beta})$ ονομάζεται *probit* και δίνεται από την σχέση

$$P_{\boldsymbol{\beta}}(Y_t = 1 \mid \mathcal{F}_{t-1}) = \Phi(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}) \tag{3.11}$$

και η αντίστοιχη συνάρτηση σύνδεσης είναι η $g = \Phi^{-1}$ (*Finney*, (1971)).

3.5 PL εκτίμηση στις δίτιμες χρονοσειρές

Η μερική πιθανοφάνεια μιας δειγματοληπτικής διαδρομής με μεταβλητή ενδιαφέροντος Y_t της οποίας η δεσμευμένη κατανομή $Y_t \mid \mathcal{F}_{t-1}$ ανήκει στην E.O.K, γράφεται (βλέπε Παράρτημα A2) ως

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^{N} f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1})$$

όπου $f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}) \equiv f_t(y_t; \boldsymbol{\beta})$. Στην προκειμένη περίπτωση όμως, επειδή $Y_t \mid \mathcal{F}_{t-1} \sim Bernoulli(1, \pi_t(\boldsymbol{\beta}))$, θα ισχύει $f_t(y_t; \boldsymbol{\beta}) = P_{\boldsymbol{\beta}}(Y_t = y_t \mid \mathcal{F}_{t-1}) = [\pi_t(\boldsymbol{\beta})]^{y_t}[1 -$

$\pi_t(\boldsymbol{\beta})]^{1-y_t}, y_t = 0, 1$. Επομένως η συνάρτηση μερικής πιθανοφάνειας για το $\boldsymbol{\beta}$ θα πάρει την μορφή

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^{N} [\pi_t(\boldsymbol{\beta})]^{y_t} [1 - \pi_t(\boldsymbol{\beta})]^{1-y_t} \overset{(3.3),(3.6)}{=}$$

$$= \prod_{t=1}^{N} [F(\boldsymbol{\beta}' \boldsymbol{Z}_{t-1})]^{y_t} [1 - F(\boldsymbol{\beta}' \boldsymbol{Z}_{t-1})]^{1-y_t}. \tag{3.12}$$

Για να προκύψει ο εκτιμητής μέγιστης μερικής πιθανοφάνειας του $\boldsymbol{\beta}$ απαιτείται η μεγιστοποίηση του λογαρίθμου της συνάρτησης μερικής πιθανοφάνειας (βλέπε Παράρτημα A.2, σχέση A.8), που στην προκειμένη περίπτωση ισούται με

$$\ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \{y_t log F(\boldsymbol{\beta}^{'} \boldsymbol{Z}_{t-1}) + (1 - y_t) log(1 - F(\boldsymbol{\beta}^{'} \boldsymbol{Z}_{t-1}))\} \equiv \sum_{t=1}^{N} \ell_t. \tag{3.13}$$

Παρατηρώντας την (3.13) διαπιστώνουμε ότι στις δίτιμες χρονοσειρές η εξάρτηση των ℓ_t άρα και του $\ell(\boldsymbol{\beta})$ από το παραμετρικό διάνυσμα $\boldsymbol{\beta}$ είναι άμεση αφού η αθροιστική συνάρτηση κατανομής F εκφράζεται συναρτήσει αυτού. Το γεγονός αυτό δείχνει την ευκολία των υπολογισμών όταν διαθέτουμε δίτιμα δεδομένα.

Υποθέτοντας ότι η F είναι παραγωγίσιμη και έχει συνάρτηση πυκνότητας πιθανότητας (*pdf*) $f = F^{'}$ και σε περίπτωση που ο εκτιμητής μέγιστης μερικής πιθανοφάνειας (*Maximum Partial Likelihood Estimator-MPLE*) $\hat{\boldsymbol{\beta}}$ υπάρχει, τότε προκύπτει από τις εξίσωσεις μερικής πιθανοφάνειας (*partial likelihood equation*)

$$\mathbf{S}_N(\boldsymbol{\beta}) = \nabla \ell(\boldsymbol{\beta}) = \mathbf{0} \tag{3.14}$$

ή πιο αναλυτικά

$$\begin{pmatrix} \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1} \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_2} \\ \vdots \\ \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p} \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ \vdots \\ 0 \end{pmatrix}.$$

Για τον υπολογισμό της j συνιστώσας του $\mathbf{S}_N(\boldsymbol{\beta})$ θα βασιστούμε στο γεγονός ότι

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{t=1}^{N} \frac{\partial \ell_t}{\partial \beta_j}, \quad j = 1, 2, \ldots, p \tag{3.15}$$

Έτσι έχουμε

$$\frac{\partial \ell_t}{\partial \beta_j} = y_t \frac{1}{F(\boldsymbol{\beta}^{'} \boldsymbol{Z}_{t-1})} f(\boldsymbol{\beta}^{'} \boldsymbol{Z}_{t-1}) z_{(t-1)j} - (1 - y_t) \frac{1}{1 - F(\boldsymbol{\beta}^{'} \boldsymbol{Z}_{t-1})} f(\boldsymbol{\beta}^{'} \boldsymbol{Z}_{t-1}) z_{(t-1)j} \Rightarrow$$

$$\frac{\partial \ell_t}{\partial \beta_j} = f(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1}) z_{(t-1)j} \Big\{ \frac{y_t}{F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})} - \frac{1-y_t}{1-F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})} \Big\} \Rightarrow$$

$$\frac{\partial \ell_t}{\partial \beta_j} = f(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1}) z_{(t-1)j} \Big\{ \frac{(1-F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1}))y_t - (1-y_t)F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})}{F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})(1-F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})} \Big\} \Rightarrow$$

$$\frac{\partial \ell_t}{\partial \beta_j} = \frac{f(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1}) z_{(t-1)j} (y_t - F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1}))}{F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})(1-F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})} \overset{(3.3),(3.6)}{\Longrightarrow}$$

$$\frac{\partial \ell_t}{\partial \beta_j} = \frac{f(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1}) z_{(t-1)j}}{F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})(1-F(\boldsymbol{\beta}^{'}\boldsymbol{Z}_{t-1})} (Y_t - \pi_t(\boldsymbol{\beta})) \tag{3.16}$$

Έτσι, από τις (3.15), (3.16) και το γεγονός ότι $\eta_t = \boldsymbol{\beta}'\boldsymbol{Z}_{t-1}$, για την j συνιστώσα του διανύσματος των μερικών σκόρ προκύπτει,

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{t=1}^{N} Z_{(t-1)j} \frac{f(\eta_t)}{F(\eta_t)(1-F(\eta_t))} (Y_t - \pi_t(\beta)), j = 1, 2, \ldots, p. \tag{3.17}$$

Επιπλέον, θέτοντας $D(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1}) = D(\eta_t) = \frac{f(\eta_t)}{F(\eta_t)(1-F(\eta_t))}$, για το *partial score vector* θα έχουμε

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} D(\eta_t)(Y_t - \pi_t(\boldsymbol{\beta})) \tag{3.18}$$

ή πιο αναλυτικά

$$\mathbf{S}_N(\boldsymbol{\beta}) = \Big(\sum_{t=1}^{N} Z_{(t-1)1} D(\eta_t)(Y_t - \pi_t(\boldsymbol{\beta})), \ldots, \sum_{t=1}^{N} Z_{(t-1)p} D(\eta_t)(Y_t - \pi_t(\boldsymbol{\beta})) \Big)'.$$

Υπενθυμίζοντας ότι $\pi_t(\boldsymbol{\beta}) = \mu_t(\boldsymbol{\beta})$ και $\sigma_t^2(\boldsymbol{\beta}) = \mu_t(\boldsymbol{\beta})(1-\mu_t(\boldsymbol{\beta}))$ η (3.18) προκύπτει και από την γενική σχέση ορισμού του $\mathbf{S}_N(\boldsymbol{\beta})$ (A.14) που αναφέρεται στο Παράρτημα Α2. Παρατηρώντας την (3.18) γίνεται φανερό ότι το *partial score vector* εξαρτάται από το μήκος της δειγματοληπτικής διαδρομής N. Έτσι μια πρώτη παρατήρηση που μπορούμε να κάνουμε είναι ότι η ακρίβεια των εκτιμητών του παραμετρικού διανύσματος $\boldsymbol{\beta}$, οι οποίοι όπως αναφέραμε προκύπτουν από την λύση της εξίσωσης $\nabla \mathbf{S}_N(\boldsymbol{\beta}) = \mathbf{0}$, και κατ᾽ επέκταση η καταλληλότητα του στατιστικού μοντέλου

$$g(\pi_t(\boldsymbol{\beta})) = \boldsymbol{\beta}'\boldsymbol{Z}_{t-1}, \qquad t = 1, 2, \ldots, N,$$

θα επηρεάζεται (πέρα από την επιλογή του $\boldsymbol{Z}_{t-1}$) και από το μέγεθος της πραγματοποίησης N. Γενικά στην μελέτη διαχρονικών φαινομένων μέσω χρονολογικών σειρών, αποφεύγονται οι σύντομες δειγματοληπτικές διαδρομές. Ο ερευνητής του διαχρονικού φαινομένου απαιτείται να διαθέτει ικανοποιητικά μεγάλες πραγματοποιήσεις αφού με

αυτόν τον τρόπο θα έχει καλύτερη εποπτεία κατορθώνοντας να καταγράψει χαρακτηριστικά της χρονοσειράς τα οποία κρίνονται ιδιαίτερα σημαντικά για την ορθή στατιστική συμπερασματολογία (π.χ εποχικότητα, *patterns*, τάσεις κ.τ.λ). Συνάμα, όπως θα δούμε και στην συνέχεια, η κανονικότητα του εκτιμητή $\hat{\boldsymbol{\beta}}$ $(MPLE)$ εξασφαλίζεται ασυμπτωτικά, δηλαδή καθώς το $N \to \infty$.

Για οποιαδήποτε χρονική στιγμή παρατήρησης του φαινομένου μπορούμε να ορίσουμε την ποσότητα

$$\mathbf{S}_t(\boldsymbol{\beta}) = \sum_{s=1}^{t} \mathbf{Z}_{s-1} D(\eta_s)(Y_s - \pi_s(\boldsymbol{\beta})), t = 1, 2, \ldots, N. \tag{3.19}$$

Η παραπάνω ποσότητα είναι ένα τυχαίο διάνυσμα αφού η τιμή της την χρονική στιγμή $t-1$ είναι άγνωστη και εξαρτάται από την τιμή που θα πάρει η μεταβλητή ενδιαφέροντος την χρονική στιγμή t, έστω $Y_t = y_t$.

Το $S_t(\boldsymbol{\beta})$ είναι το *μερικό άθροισμα* (*partial sum*) το οποίο αντιστοιχεί στον χρόνο t. Έτσι θεωρώντας τον χρόνο $t = 0$ ως αφετηρία παρακολούθησης του φαινομένου, παραθέτουμε για τις πρώτες 3 χρονικές στιγμές που καταγράφεται η τιμή της μεταβλητής ενδιαφέροντος τα αντίστοιχα μερικά αθροίσματα

$$\begin{aligned} t = 1 &: S_1(\boldsymbol{\beta}) = \boldsymbol{Z}_0 D(\eta_1)(Y_1 - \pi_1(\boldsymbol{\beta})) \\ t = 2 &: S_2(\boldsymbol{\beta}) = S_1(\boldsymbol{\beta}) + \boldsymbol{Z}_1 D(\eta_2)(Y_2 - \pi_2(\boldsymbol{\beta})) \\ t = 3 &: S_3(\boldsymbol{\beta}) = S_2(\boldsymbol{\beta}) + \boldsymbol{Z}_2 D(\eta_3)(Y_3 - \pi_3(\boldsymbol{\beta})) \end{aligned}$$

Σύμφωνα λοιπόν με τα προαναφερθέντα μπορούμε να ορίσουμε την *στοχαστική ανέλιξη των μερικών σκορ* (*partial score process*) $\{\mathbf{S}_t(\boldsymbol{\beta})\}, t = 1, 2, \ldots$, που είναι μια διανυσματική στοχαστική ανέλιξη . Η δεδομένη στοχαστική ανέλιξη επειδή είναι *martingale* (*Hall* και *Heyde* (1980))[1] στοχαστικών ανελίξεων θα συμβάλει όπως θα δούμε στην συνέχεια στην συμπερασματολογία των κατηγορικών χρονοσειρών. Πράγματι για την *partial score process* ισχύει η ακόλουθη πρόταση.

Πρόταση 3.5.1 *Για την διανυσματική στοχαστική ανέλιξη* $\{\mathbf{S}_t(\boldsymbol{\beta})\}, t = 1, 2, \ldots$ *ισχύει*

$$E[\mathbf{S}_{t+1}(\boldsymbol{\beta}) \mid \mathcal{F}_t] = \mathbf{S}_t(\boldsymbol{\beta}). \tag{3.20}$$

[1] Μια στοχαστική ανέλιξη $\{X_t\}, t = 1, 2, \ldots$ είναι *martingale* αν ισχύει $E|X_t| < +\infty$ και $E[X_t \mid X_{t-1}, X_{t-2}, \ldots] = X_{t-1}$ σχεδόν βεβαίως. Δηλαδή ένα στοχαστικό φαινόμενο θα χαρακτηρίζεται από την ιδιότητα *martingale* αν η αναμενόμενη τιμή της X_t δοθέντος του παρελθόντος ισούται με την τιμή της την αμέσως προηγούμενη χρονική στιγμή, δηλαδή με την X_{t-1}.

Απόδειξη 3.5.1

$$E[\mathbf{S}_{t+1}(\boldsymbol{\beta}) \mid \mathcal{F}_t] = E[\sum_{s=1}^{t+1} \boldsymbol{Z}_{s-1} D(\eta_s)(Y_s - \pi_s(\boldsymbol{\beta})) \mid \mathcal{F}_t]$$

$$= E[\sum_{s=1}^{t} \boldsymbol{Z}_{s-1} D(\eta_s)(Y_s - \pi_s(\boldsymbol{\beta})) + \boldsymbol{Z}_t D(\eta_{t+1})(Y_{t+1} - \pi_{t+1}(\boldsymbol{\beta})) \mid \mathcal{F}_t]$$

$$\overset{(3.19)}{=} S_t(\boldsymbol{\beta}) + \boldsymbol{Z}_t D(\eta_{t+1})\Big\{E[Y_{t+1} \mid \mathcal{F}_{t+1}] - \mu_{t+1}\Big\}$$

$$= S_t(\boldsymbol{\beta}) + \boldsymbol{Z}_t D(\eta_{t+1})\Big\{\mu_{t+1} - \mu_{t+1}\Big\} = S_t(\boldsymbol{\beta}). \tag{3.21}$$

Ακόμη για το partial score vector ισχύει

$$E[\mathbf{S}_N(\boldsymbol{\beta})] = \mathbf{0}, \tag{3.22}$$

καθώς

$$E[\mathbf{S}_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}] = E[\sum_{t=1}^{N} \boldsymbol{Z}_{t-1} D(\eta_t)(Y_t - \pi_t(\boldsymbol{\beta})) \mid \mathcal{F}_{t-1}] = \mathbf{0}$$

και τελικά $E[\mathbf{S}_N(\boldsymbol{\beta})] = E[(E[\mathbf{S}_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}])] = E[(\mathbf{0})] = \mathbf{0}$.

Η λύση των εξισώσεων των μερικών σκορ είναι ο $MPLE$ του διανύσματος $\boldsymbol{\beta}$. Το σύστημα των εξισώσεων (3.18) είναι μη γραμμικό και συνήθως λύνεται με την μέθοδο *Fisher scoring*, η οποία τροποποιείται κατάλληλα ώστε να λαμβάνει υπόψη την εξάρτηση. Για περισσότερες πληροφορίες για την δεδομένη επαναληπτική διαδικασία ο ενδιαφερόμενος παραπέμπεται στο Παράρτημα Β.

3.6 Σημαντικοί Πίνακες για την Συμπερασματολογία των Δίτιμων Χρονοσειρών

Σε αυτή την ενότητα ορίζουμε κάποιους σημαντικούς πίνακες για την στατιστική συμπερασματολογία των δίτιμων χρονοσειρών (και γενικά των κατηγορικών χρονοσειρών όπως θα δούμε στο επόμενο κεφάλαιο) μέσω της μεθοδολογίας των GLM. Στην ανάλυση που θα ακολουθήσει οι νέες έννοιες θα συγκριθούν με τα ήδη γνωστα αποτελέσματα των GLM με σκοπό την καλύτερη δυνατή ερμηνεία τους.

3.6.1 Αθροιστικός κατά συνθήκη πίνακας πληροφορίας

Στα κλασικά *GLM* όπου τα $Y_i, i = 1, 2, \ldots, N$, είναι ανεξάρτητα ορίζεται ο πίνακας πληροφορίας του *Fisher* ο οποίος δίνεται από την σχέση

$$I(\boldsymbol{\beta}) = Cov(\boldsymbol{U}(\boldsymbol{\beta})) = E[\boldsymbol{U}(\boldsymbol{\beta}) \cdot \boldsymbol{U}(\boldsymbol{\beta})'] \tag{3.23}$$

(αφού $E[\boldsymbol{U}(\boldsymbol{\beta})] = \mathbf{0}$).

Εναλλακτικά για το $I(\boldsymbol{\beta})$ έχουμε

$$I(\boldsymbol{\beta}) = -E\Big[\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}\Big] = E\Big[\frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}\Big]^2.$$

Παρατήρηση 3.6.1 Το $I(\boldsymbol{\beta})$ είναι ένας πίνακας διάστασης $p \times p$ ο οποίος, εφόσον γνωρίζουμε το παραμετρικό διάνυσμα $\boldsymbol{\beta}$, παίρνει μια διαφορετική τιμή για κάθε δυνατό δείγμα. Κατα κάποιο τρόπο ο δεδομένος πραγματικός πίνακας εκφράζει 'το ποσό της πληροφορίας' που περιέχεται στα δεδομένα $\boldsymbol{Y}$ για την άγνωστη παράμετρο $\boldsymbol{\beta}$. Αξίζει να αναφέρουμε ότι ο $I^{-1}(\boldsymbol{\beta})$ ταυτίζεται με τον ασυμπτωτικό πίνακα διακυμάνσεων-συνδιακυμάνσεων του εκτιμητή μέγιστης πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ (*Rao*, (1973), σελίδα 364).

Κατα ανάλογο τρόπο για την περίπτωση των δίτιμων χρονοσειρών $\{Y_t\}, t = 1, 2, \ldots, N$ *μπορούμε να ορίσουμε ένα αντίστοιχο πίνακα του* $I(\boldsymbol{\beta})$ *ο οποίος υπο μια έννοια θα εμπεριέχει την πληροφορία που παρέχει η δειγματοληπτική διαδρομή για το υπο μελέτη φαινόμενο.* Λαμβάνοντας υπόψη την σχέση (3.23) και το γεγονός ότι η κατάσταση του φαινομένου την χρονική στιγμή t,όπως αυτή εκφράζεται από την μεταβλητή ενδιαφέροντος Y_t, επηρεάζεται από την ιστορία $\mathcal{F}_{t-1}$, εισάγουμε τον πίνακα $\mathbf{G}_N(\boldsymbol{\beta})$, διάστασης $p \times p$, που ονομάζεται *αθροιστικός κατα συνθήκη πίνακας πληροφορίας* (*cumulative conditional information matrix*) και δίνεται από την σχέση

$$\mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} \boldsymbol{Z}'_{t-1} D(\eta_t) f(\eta_t) \tag{3.24}$$

Απόδειξη της (3.24):

$$\mathbf{G}_N(\boldsymbol{\beta}) = Cov[S_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}] = E\Big\{(S_N(\boldsymbol{\beta}) - E(S_N(\boldsymbol{\beta}))] \cdot [S_N(\boldsymbol{\beta}) - E[S_N(\boldsymbol{\beta}))]' \mid \mathcal{F}_{t-1}\Big\}$$

$$\overset{(3.22)}{\Longrightarrow} \mathbf{G}_N(\boldsymbol{\beta}) = E\Big\{[S_N(\boldsymbol{\beta}][S_N(\boldsymbol{\beta})]' \mid \mathcal{F}_{t-1}\Big\}.$$

Αρχικά υπολογίζουμε τον πίνακα $\mathbf{X} = [\mathbf{S}_N(\boldsymbol{\beta}][S_N(\boldsymbol{\beta})]'$ βάση του ορισμού του διανύσματος $\mathbf{S}_N(\boldsymbol{\beta})$. Έτσι για τον δεδομένο πίνακα προέκυψε ότι τα στοιχεία του θα έχουν

την ακόλουθη μορφή

$$X_{ij} = \begin{cases} \sum_{t=1}^{N} D^2(\eta_t) Z_{(t-1)j}^2 (Y_t - \pi_t(\boldsymbol{\beta}))^2 & , i = j \\ & \mu\varepsilon \quad i,j = 1,2,\ldots,p \\ \sum_{t=1}^{N} D^2(\eta_t) Z_{(t-1)i} Z_{(t-1)j} (Y_t - \pi_t(\boldsymbol{\beta}))^2 & , i \neq j \end{cases}$$

Εν συνεχεία υπολογίζουμε τις δεσμευμένες μέσες τιμές των στοιχείων του πίνακα $\mathbf{X}$. Για το σκοπό αυτό λαμβάνουμε υπόψη ότι $E[(Y_t - \pi_t(\boldsymbol{\beta}))^2 \mid \mathcal{F}_{t-1}] = Var[Y_t \mid \mathcal{F}_{t-1}] \equiv \sigma_t^2(\boldsymbol{\beta})$. Επειδή όμως στην περίπτωση που εξετάζουμε ισχύει ότι $Y_t \mid \mathcal{F}_{t-1} \sim Bernoulli(\pi_t(\boldsymbol{\beta}))$ θα έχουμε

$$\sigma_t^2(\boldsymbol{\beta}) = \pi_t(\boldsymbol{\beta})(1 - \pi_t(\boldsymbol{\beta})) \overset{(3.3),(3.6)}{=} F(\eta_t)(1 - F(\eta_t)) \tag{3.25}$$

και επομένως για τα στοιχεία του πίνακα $\mathbf{G}_N(\boldsymbol{\beta})$ θα ισχύει

$$G_{N(ij)}(\boldsymbol{\beta}) = \begin{cases} \sum_{t=1}^{N} D^2(\eta_t) Z_{(t-1)j}^2 F(\eta_t)(1 - F(\eta_t)) & , i = j \\ & \mu\varepsilon \quad i,j = 1,2,\ldots,p \\ \sum_{t=1}^{N} D^2(\eta_t) Z_{(t-1)i} Z_{(t-1)j} F(\eta_t)(1 - F(\eta_t)) & , i \neq j \end{cases}$$

Επειδή όμως $D(\eta_t) = \frac{f(\eta_t)}{F(\eta_t)(1-F(\eta_t))}$ αποδεικνύεται άμεσα η σχέση (3.24). Στην πράξη όμως για τα στοιχεία $G_{N(ij)}(\boldsymbol{\beta})$ συνηθίζουμε να χρησιμοποιούμε τις ακόλουθες εκφράσεις

$$G_{N(ij)}(\boldsymbol{\beta}) = \begin{cases} \sum_{t=1}^{N} Z_{(t-1)j}^2 \frac{f^2(\eta_t)}{F(\eta_t)(1-F(\eta_t))} & , i = j \\ & \mu\varepsilon \quad i,j = 1,2,\ldots,p \\ \sum_{t=1}^{N} Z_{(t-1)i} Z_{(t-1)j} \frac{f^2(\eta_t)}{F(\eta_t)(1-F(\eta_t))} & , i \neq j \end{cases}$$

Έτσι η τελική μορφή του *αθροιστικού κατά συνθήκη πίνακα πληροφορίας* θα είναι

$$\mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} \boldsymbol{Z}_{t-1}' \frac{f^2(\boldsymbol{\beta}' \boldsymbol{Z}_{t-1})}{F(\boldsymbol{\beta}' \boldsymbol{Z}_{t-1})(1 - F(\boldsymbol{\beta}' \boldsymbol{Z}_{t-1}))}. \tag{3.26}$$

Η (3.26) προκύπτει εναλλακτικά από τον γενικό ορισμό του $\mathbf{G}_N(\boldsymbol{\beta})$ σύμφωνα με την σχέση (A.16) του Παραρτήματος A2, λαμβάνοντας υπόψη ότι $\mu_t(\boldsymbol{\beta}) = \pi_t(\boldsymbol{\beta}) = F(\boldsymbol{\beta}' \mathbf{Z}_{t-1})$ και $\frac{\partial \mu_t}{\partial \eta_t} = f(\eta_t)$ με $\eta_t = \boldsymbol{\beta}' \mathbf{Z}_{t-1}$) καθώς και την σχέση (3.25).

3.6.2 Παρατηρούμενος πίνακας πληροφορίας

Ο παρατηρούμενος πίνακας πληροφορίας (*observed information matrix*) ορίζεται ως

$$\mathbf{H}_N(\boldsymbol{\beta}) \equiv \nabla \nabla' (-\log PL(\boldsymbol{\beta})) \tag{3.27}$$

ή πιο αναλυτικά

$$\mathbf{H}_N(\boldsymbol{\beta}) = \begin{pmatrix} -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_1^2} & -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_2} & \cdots & -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_1 \partial \beta_p} \\ -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_1} & -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_2^2} & \cdots & -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_2 \partial \beta_p} \\ \vdots & \vdots & \ddots & \vdots \\ -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_1} & -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_p \partial \beta_2} & \cdots & -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_p^2} \end{pmatrix}$$

Παρατηρήσεις. Ο δεδομένος πίνακας ταυτίζεται με τον παρατηρούμενο πίνακα πληροφορίας του *Fisher*

$$\mathbf{I}_o(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}$$

που τον συναντάμε στην κλασική περίπτωση που έχουμε ανεξάρτητες παρατηρήσεις. Ο $\mathbf{I}_o(\boldsymbol{\beta})$ είναι συνεπής εκτιμητής του $\mathbf{I}(\boldsymbol{\beta})$ και συνηθίζεται στην πράξη αφού παρακάμπτεται η δυσκολία υπολογισμού των αναμενόμενων τιμών των ποσοτήτων $-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j}$.

Ακόμη παρατηρούμε ότι ισχύει

$$\mathbf{H}_N(\boldsymbol{\beta}) = (-1) \times \mathbf{H}(\boldsymbol{\beta})$$

όπου $\mathbf{H}(\boldsymbol{\beta}) = \frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}^2}$ είναι ο *Εσσιανός* (*Hessian*) πίνακας που επίσης χρησιμοποιείται στην περίπτωση των ανεξάρτητων παρατηρήσεων για την εύρεση των εκτιμητών μέγιστης πιθανοφάνειας μέσω της μεθόδου *Newton-Raphson* (βλέπε *Agresti*, (2002), κεφάλαιο 4).

Η παραπάνω συγκριτική παρουσίαση του πίνακα $\mathbf{H}_N(\boldsymbol{\beta})$ με ήδη γνωστούς πίνακες από τα κλασσικά *GLM* με μια πρώτη ματιά φαίνεται χωρίς νόημα αλλά όπως θα δούμε στην συνέχεια βοηθά στην κατανόηση του ρόλου που διαδραματίζει ο δεδομένος πίνακας για τις κατηγορικές χρονοσειρές.

Ο *παρατηρούμενος πίνακας πληροφορίας* ικανοποιεί την σχέση (*Fokianos* και *Kedem*, (2002), σελίδα 12)

$$\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) - \mathbf{R}_N(\boldsymbol{\beta}) \tag{3.28}$$

όπου $\mathbf{R}_N(\boldsymbol{\beta})$ είναι ο πίνακας που όταν προστεθεί στον $\mathbf{H}_N(\boldsymbol{\beta})$ δίνει τον *αθροιστικό υπό συνθήκη πίνακα πληροφορίας*. Ο $\mathbf{R}_N(\boldsymbol{\beta})$ στην περίπτωση μας είναι

$$\mathbf{R}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t(\phi)} \sum_{t=1}^{N} \boldsymbol{Z}_{t-1} d_t(\boldsymbol{\beta}) \boldsymbol{Z}'_{t-1} (Y_t - \mu_t(\boldsymbol{\beta})), \tag{3.29}$$

όπου $d_t(\boldsymbol{\beta}) = [\partial^2 u(\eta_t)/\partial \eta_t^2]$. Ο γενικός τύπος ορισμού του $\mathbf{R}_N(\boldsymbol{\beta})$ βρίσκεται στο Παράρτημα Α2, στην σχέση (Α.20). Σύμφωνα όμως με τη σχέση (Α.7) και το γεγονός

ότι $u(\cdot) = (g \circ b')^{-1}(\cdot)$ θα ισχύει $u(\eta_t) = \theta_t$. Συνάμα λόγω της (3.4) για την φυσική παράμετρο θα έχουμε

$$\theta_t = \log\big(\frac{\pi_t(\boldsymbol{\beta})}{1-\pi_t(\boldsymbol{\beta})}\big)$$

και επειδή αναφερόμαστε σε δίτιμες χρονοσειρές

$$\overset{(3.3),(3.6)}{\Longrightarrow} \theta_t(\eta_t) = \log\big(\frac{F(\eta_t)}{1-F(\eta_t)}\big).$$

Επομένως

$$\frac{d\theta_t(\eta_t)}{d\eta_t} = \frac{1}{\frac{F}{1-F}} \cdot \frac{f(1-F)-F(1-F)}{(1-F)^2} = \frac{1-F}{F} \cdot \frac{f(1-F)+fF}{(1-F)^2} \Rightarrow$$

$$\frac{d\theta_t(\eta_t)}{d\eta_t} = \frac{f}{F(1-F)} \equiv D(\eta_t).$$

Έτσι θέτοντας στην σχέση (3.29) $d_t(\boldsymbol{\beta}) = \frac{d}{d\eta_t}(D(\eta_t)) \equiv W(\eta_t)$ και επειδή $\alpha_t(\phi) = 1$, (βλέπε σελίδα 45) ο πίνακας $\mathbf{R}_N(\boldsymbol{\beta})$ για την περίπτωση των δίτιμων χρονοσειρών γράφεται στην μορφή

$$\mathbf{R}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1}\boldsymbol{Z}'_{t-1} W(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})(Y_t - \mu_t(\boldsymbol{\beta})). \tag{3.30}$$

3.7 Ασυμπτωτικά Αποτελέσματα στις Δίτιμες Χρονοσειρές

Έχοντας ολοκληρώσει την παρουσίαση των βασικών πινάκων που θα μας βοηθήσουν να βρούμε τους εκτιμητές μέγιστης μερικής πιθανοφάνειας (*MPLE*) θα αναφέρουμε μερικά σημαντικά ασυμπτωτικά αποτελέσματα, στην περίπτωση που διαθέτουμε δίτιμες εξαρτημένες παρατηρήσεις. Τα αποτελέσματα που θα δώσουμε προκύπτουν από την άμεση εφαρμογή της ασυμπτωτικής θεωρίας που ισχύει γενικά στις κατηγορικές χρονοσειρές και παρουσιάζεται αναλυτικά στο Παράρτημα Α3.

Έτσι ο πίνακας $G(\boldsymbol{\beta})$ που ορίζεται γενικά σύμφωνα με τις σχέσεις (Α.21) και (Α.22) στην ειδική περίπτωση των δίτιμων χρονοσειρών γράφεται ως

$$\frac{\mathbf{G}_N(\boldsymbol{\beta})}{N} \to \mathbf{G}(\boldsymbol{\beta}) = \int_{R^p} \boldsymbol{z}\boldsymbol{z}' \frac{f^2(\boldsymbol{\beta}'\boldsymbol{z})}{F(\boldsymbol{\beta}'\boldsymbol{z})(1-F(\boldsymbol{\beta}'\boldsymbol{z}))} \nu(dz). \tag{3.31}$$

Βάσει των ασυμπτωτικών αποτελεσμάτων μπορούμε να προχωρήσουμε στην διατύπωση ενός θεωρήματος το οποίο εξασφαλίζει την ύπαρξη, την μοναδικότητα και προσδιορίζει την ασυμπτωτική κατανομή των εκτιμητών μέγιστης μερικής πιθανοφάνειας τόσο για τις διτιμες, όσο και τις κατηγορικές σειρές όπως θα δούμε στο Κεφάλαιο 4 (*Wong* (1986)).

Θεώρημα 3.7.1 *Ο εκτιμητής* MPLE $\hat{\boldsymbol{\beta}}$ *είναι σχεδόν βεβαίως μοναδικός για επαρκώς μεγάλο* N *και καθώς* $N \to \infty$ *ισχύουν τα ακόλουθα*

(i)

$$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$$

(ii)

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_p(\mathbf{0}, \mathbf{G}^{-1}(\boldsymbol{\beta}))$$

(iii)

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) - \frac{1}{\sqrt{N}}\mathbf{G}^{-1}(\boldsymbol{\beta})\mathbf{S}_N(\boldsymbol{\beta}) \xrightarrow{p} \mathbf{0}$$

Παρατηρούμε λοιπόν πως ο εκτιμητής MPLE $\hat{\boldsymbol{\beta}}$ στην περίπτωση των δίτιμων σειρών, ικανοποιεί τις ίδιες ασυμπτωτικές ιδιότητες με τους συνήθεις εκτιμητές μέγιστης πιθανοφάνειας όταν έχουμε ανεξάρτητα δίτιμα δεδομένα (*Silvapulle* (1981)), (*Wedderburn* (1976)). Δηλαδή είναι συνεπής και ασυμπτωτικά κανονικός (CAN).

3.8 Συμπερασματολογία στην Λογιστική Παλινδρόμηση

Με σκοπό να γίνουν κατανοητά τα θεωρητικά αποτελέσματα που παρουσιάστηκαν στις προγούμενες παραγράφους θα τα εφαρμόσουμε στην περίπτωση της λογιστικής παλινδρόμησης.

Όπως αναφέραμε λοιπόν στην αρχή του κεφαλαίου τα μοντέλα που θα χρησιμοποιήσουμε για την στατιστική συμπερασματολογία σχετικά με την δεσμευμένη πιθανότητα επιτυχίας, στην περίπτωση των δίτιμων χρονολογικών σειρών, θα έχουν την γενική μορφή της (3.6), όπου η συνάρτηση F είναι η αθροιστική συνάρτηση κατανομής και όπως είδαμε ταυτίζεται με την αντίστροφη της συνάρτηση σύνδεσης $h : R \to [0, 1]$. Αν επιλέξουμε ως g^{-1} την *cdf* της τυπικής λογιστικής κατανομής τότε καταλήγουμε στο μοντέλο λογιστικής παλινδρόμησης το οποίο είναι το πιο συνηθισμένο στις δίτιμες κατηγορικές χρονοσειρες. Στην προκειμένη περίπτωση για την *cdf* F_l και την συνάρτηση σύνδεσης g ισχύουν οι σχέσεις των σελίδων 46 και 47. Στο σημείο αυτό είμαστε έτοιμοι να δώσουμε τις μορφές των πινάκων που παρουσιάσμε στην προηγούμενη παράγραφο. Αρχικά καλούμαστε να υπολογίσουμε την ποσότητα $D(x) \equiv \frac{f(x)}{F(x)(1-F(x))}$ όπου $x = \eta_t$ και $f(x) = \frac{\partial F(x)}{\partial x} = \frac{e^x}{(1+e^x)^2}$. Έτσι θα έχουμε

$$D(x) = \frac{\frac{e^x}{(1+e^x)^2}}{\frac{e^x}{1+e^x}(1 - \frac{e^x}{1+e^x})} = \frac{e^x(1+e^x)^2}{(1+e^x)^2 e^x} = 1 \tag{3.32}$$

επιπλέον

$$W(x) = \frac{d}{dx}D(x) = 0. \tag{3.33}$$

Για τον υπολογισμό του πίνακα $\mathbf{S}_N(\boldsymbol{\beta})$ έχουμε

$$(3.18) \overset{(3.32)}{\Rightarrow} \mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1}(Y_t - \pi_t(\boldsymbol{\beta})) \tag{3.34}$$

Για τον πίνακα $\mathbf{G}_N(\boldsymbol{\beta})$ προκύπτει

$$(3.24) \overset{(3.32)}{\Rightarrow} \mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1}\boldsymbol{Z}'_{t-1} f(\eta_t), \tag{3.35}$$

όμως για την f έχουμε

$$f(\eta_t) = \frac{e^{\eta_t}}{[1+e^{\eta_t}]^2} = \frac{e^{\eta_t}}{1+e^{\eta_t}} \cdot \frac{1}{1+e^{\eta_t}}. \tag{3.36}$$

από την (3.36) προκύπτει $\frac{e^{\eta_t}}{1+e^{\eta_t}} = \pi_t(\boldsymbol{\beta})$ και $1 - \pi_t(\boldsymbol{\beta}) = \frac{1}{1+e^{\eta_t}}$. Έτσι μια δεύτερη μορφή για την f θα είναι

$$f(\eta_t) = \pi_t(\boldsymbol{\beta})(1 - \pi_t(\boldsymbol{\beta})). \tag{3.37}$$

Σύμφωνα με τις δύο παραπάνω μορφές για την f προκύπτουν αντιστοίχως οι ακόλουθες δύο μορφές του πίνακα $\mathbf{G}_N(\boldsymbol{\beta})$ για την περίπτωση της λογιστικής παλινδρόμησης,

$$\mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1}\boldsymbol{Z}'_{t-1} \frac{\exp(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})}{[1+\exp(\boldsymbol{\beta}'\boldsymbol{Z}_{t-1})]^2} \tag{3.38}$$

και

$$\mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \boldsymbol{Z}_{t-1}\boldsymbol{Z}'_{t-1}\pi_t(\boldsymbol{\beta})(1 - \pi_t(\boldsymbol{\beta})). \tag{3.39}$$

Ο πίνακας $\mathbf{R}_N(\boldsymbol{\beta})$ γίνεται

$$(3.30) \overset{(3.33)}{\Rightarrow} \mathbf{R}_N(\boldsymbol{\beta}) = \mathbf{0}. \tag{3.40}$$

Έτσι σύμφωνα με την σχέση (3.28) θα ισχύει ότι $\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta})$.

Παρατήρηση 3.8.1 Το αποτέλεσμα ότι $\mathbf{R}_N(\boldsymbol{\beta}) = \mathbf{0}$ ήταν αναμενόμενο. Αυτό συμβαίνει διότι στην περίπτωση της λογιστικής παλινδρόμησης η συνάρτηση g ταυτίζεται με τον κανονικό σύνδεσμο και επομένως θα ισχύει άμεσα ότι $d_t = 0$ (βλέπε *Fokianos-Kedem*, 2002, σελ. 14).

Ολοκληρώνοντας την δεδομένη θεωρητική εφαρμογή δίνουμε και την μορφή του οριακού πίνακα πληροφορίας ανα παρατήρηση. Έτσι σύμφωνα με τις σχέσεις (3.24), (A.21), (A.22) και (3.36) έχουμε

$$\frac{\mathbf{G}_N(\boldsymbol{\beta})}{N} \to \mathbf{G}(\boldsymbol{\beta}) = \int_{R^p} \boldsymbol{z}\boldsymbol{z}' \frac{e^{\boldsymbol{\beta}'\boldsymbol{z}}}{(1+e^{\boldsymbol{\beta}'\boldsymbol{z}})^2} \nu(dz) \tag{3.41}$$

κατά πιθανότητα καθώς $N \to \infty$. Χαρακτηριστικό παράδειγμα δίτιμης χρονοσειράς θα παρουσιάσουμε στο κεφάλαιο 6 και το οποίο θα αναφέρεται σε δεδομένα βροχοπτώσεων.

Παρατήρηση 3.8.2 Οι διαδικασίες που ακολουθούμε για τους ελέγχους υποθέσεων και την αξιολόγηση των υποδειγμάτων παρουσιάζονται αναλυτικά στο $3^{ο}$ κεφάλαιο που θα μιλήσουμε για τις κατηγορικές χρονοσειρές. Τα αποτελέσματα που ισχύουν στις ποιοτικές σειρές εύκολα εφαρμόζονται και στις δίτιμες σειρές, αφού οι δεύτερες αποτελούν ειδική κατηγορία των τελευταίων, όπως θα δούμε αναλυτικά.

Κεφάλαιο 4

Ανάλυση Κατηγορικών Χρονοσειρών

4.1 Εισαγωγή

Οι δίτιμες χρονοσειρές που παρουσιάστηκαν στο προηγούμενο κεφάλαιο αποτελούν την απλούστερη περίπτωση κατηγορικών χρονολογικών σειρών. Στη πράξη όμως συναντάμε διαχρονικά φαινόμενα στα οποία η ποιοτική μεταβλητή ενδιαφέροντος, Y_t, παρουσιάζει περισσότερα από δυο δυνατά αποτελέσματα. Έτσι αν η κατηγορική αποκριτική μεταβλητή, που χαρακτηρίζει την έκβαση του φαινομένου, για κάθε χρονική στιγμή έχει m δυνατά επίπεδα, τότε η δειγματοληπτική διαδρομή που θα προκύψει θα είναι μια ακολουθία από αριθμούς που ανήκουν στο σύνολο $\{1, 2, \ldots, m\}$ (*Fokianos and Kedem* (2003)).

Όπως και στην περίπτωση των δίτιμων χρονοσειρών τα δεδομένα μας παρουσιάζουν διαχρονική εξάρτηση και επομένως δεν μπορούμε να χρησιμοποιήσουμε την εύχρηστη διάσπαση της από κοινού κατανομής του δείγματος σε γινόμενο N περιθωρίων πυκνοτήτων πιθανότητας, εφόσον διαθέτουμε μια δειγματοληπτική διαδρομή του φαινομένου μήκους N. Για να ξεπεραστεί το δεδομένο πρόβλημα θα χρησιμοποιήσουμε ξανά την μερική πιθανοφάνεια, την οποία εισαγάγαμε στο 2^o κεφάλαιο. Μέσω της δεδομένης μεθοδολογικής προσέγγισης όπως θα δούμε στην συνέχεια είναι δυνατή η διερεύνηση του διαχρονικού φαινομένου με ικανοποιητική ακρίβεια αλλά συγχρόνως, που είναι και το πιο σημαντικό, με σχετική ευκολία. Συγκεκριμένα για κάθε χρονική στιγμή t θα μπορέσουμε να εκτιμήσουμε την πιθανότητα εμφάνισης του κάθε επιπέδου της Y_t λαμβάνοντας όμως υπόψη την ιστορία του φαινομένου μέχρι τον χρόνο $t-1$, που είναι η $\mathcal{F}_{t-1}$. Το $\mathcal{F}_{t-1}$ είναι η σ-άλγεβρα η οποία εμπεριέχει οτιδήποτε γνωρίζουμε για το φαινόμενο πριν από τον χρόνο t, ακόμη και τιμές μεταβλητών για τον χρόνο t που είναι ήδη γνωστές από τον χρόνο $t-1$. Θα πρέπει να σημειώσουμε ότι η σ-άλγεβρα $\mathcal{F}_{t-1}$

ουσιαστικά παράγεται από το διάνυσμα των συμμεταβλητών $\mathbf{Z}_{t-1}$. Αυτό συμβαίνει διότι κάθε χρονική στιγμή ο παρατηρητής του φαινομένου αξιοποιεί την πληροφορία του παρελθόντος μέσω του δεδομένου χρονοεξαρτώμενου διανύσματος το οποίο έχει σταθερή μορφή. Έτσι θα αντιμετωπίζουμε με τον ίδιο τρόπο τις δεσμευμένες ροπές ως προς $\mathbf{Z}_{t-1}$ και $\mathcal{F}_{t-1}$. Ακόμη θα πρέπει να τονίσουμε ότι το $\mathbf{Z}_{t-1}$ κατα την διαδικασία στατιστικής μοντελοποίησης αλλάζει συνεχώς διάσταση μέχρι να καταλήξουμε στο καλύτερο δυνατό υπόδειγμα, σύμφωνα με διαγνωστικούς ελέγχους και κριτήρια καλής προσαρμογής.

Στην προσπάθεια μοντελοποίησης των κατηγορικών χρονοσειρών θα στηριχτούμε και πάλι στην θεωρία των Γενικευμένων Γραμμικών Μοντέλων(GLM). Για να γίνει όμως κάτι τέτοιο όπως θα δούμε στην συνέχεια απαιτείται να επεκτείνουμε τις έννοιες της Εκθετικής Οικογένειας Κατανομών (Ε.Ο.Κ) και των GLM αντιστοίχως στις έννοιες της πολυμεταβλητής Ε.Ο.Κ (π.Ε.Ο.Κ) και των πολυμεταβλητών γενικευμένων γραμμικών μοντέλων ($mGLM$) (*Fahrmeir and Tutz* (2001)).

Στην Ενότητα 2 του δεδομένου κεφαλαίου θα δώσουμε κάποιους χρήσιμους συμβολισμούς και ορολογίες που συναντάμε στις ποιοτικές σειρές. Πιο αναλυτικά θα δείξουμε πως από το μονομεταβλητό Y_t οδηγούμαστε στην περιγραφή των καταστάσεων του συστήματος, για κάθε χρονική στιγμή, μέσω του διανύσματος $(Y_{t1}, Y_{t2}, \ldots, Y_{tq})'$ όπου $q = m-1$. Εν συνεχεία στην Ενότητα 3 θα αναφερθούμε στην π.Ε.Ο.Κ και στα $mGLM$ καταλήγοντας στο γενικό μοντέλο παλινδρόμησης για κατηγορικές σειρές με m επίπεδα. Στις παραγράφους 4 και 5 θα παρουσιάσουμε την διαδικασία συμπερασματολογίας για ποιοτικές χρονοσειρές που έχουν αντιστοίχως $m = 3$ και $m > 3$ επίπεδα. Αμέσως μετά, στην Ενότητα 6 θα δώσουμε ορισμένα σημαντικά ασυμπτωτικά αποτελέσματα, ενώ στην Παράγραφο 7 θα αναφερθούμε στους ελέγχους υποθέσεων. Τέλος στην Ενότητα 8 θα εξηγήσουμε πώς πραγματοποιούνται οι διαγνωστικοί ελέγχοι στις κατηγορικές χρονοσειρές.

4.2 Συμβολισμοί-Ορολογία

Έστω ότι παρατηρούμε μια κατηγορική χρονοσειρά $\{Y_t\}, t = 1, 2, \ldots, N$ και έστω m ο αριθμός των κατηγοριών της. Σε κάθε ένα από τα δυνατά αποτελέσματα του φαινομένου για οποιαδήποτε χρονική στιγμή αντιστοιχεί ένας ακέραιος στο σύνολο $\{1, 2, \ldots, m\}$. Έτσι για κάθε χρονική στιγμή παρατήρησης του φαινομένου οι δυνατές τιμές της τυχαίας μεταβλητής Y_t θα είναι οι $1, 2, \ldots, m-1, m$. Η ανάθεση ακεραίων

αριθμών στα επίπεδα της μεταβλητής ενδιαφέροντος δεν είναι μοναδική. Για παράδειγμα, έστω ότι παρακολουθούμε την καθημερινή προτίμηση ενός ατόμου ως προς το μεταφορικό μέσο. Έστω ότι τα δυνατά μεταφορικά μέσα που μπορεί να επιλέξει κάποιο άτομο είναι το αυτοκίνητο, το λεωφορείο και το τρένο. Κατα συνέπεια οι δυνατές τιμές της Y_t που δηλώνουν για κάθε μέρα την επιλογή του τρόπου μετακίνησης είναι οι ακέραιοι 1, 2, 3. Δυο από τους δυνατούς τρόπους που μπορούν οι δεδομένοι ακέραιοι να αντιστοιχιστούν με τους τρείς τρόπους μετακίνησης είναι οι ακόλουθοι:

$$\text{αυτοκίνητο} \longleftrightarrow 1$$
$$\text{λεωφορείο} \longleftrightarrow 2$$
$$\text{τρένο} \longleftrightarrow 3$$

και

$$\text{λεωφορείο} \longleftrightarrow 1$$
$$\text{τρένο} \longleftrightarrow 2$$
$$\text{αυτοκίνητο} \longleftrightarrow 3.$$

Προφανώς ακολουθώντας διαφορετικές κωδικοποιήσεις διατρέχουμε τον κίνδυνο να οδηγηθούμε σε διαφορετικά αποτελέσματα μετά την ολοκλήρωση της στατιστικής ανάλυσης. Με σκοπό να μειώσουμε την αυθαιρεσία που παρουσιάζεται με την ανάθεση ακεραίων στις κατηγορίες της Y_t παρατηρούμε ότι η κατάσταση του φαινομένου που μελετάμε οποιαδήποτε χρονική στιγμή t μπορεί να περιγραφεί μέσω του τυχαίου διανύσματος $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tq})'$ διαστασης $q = (m-1) \times 1$. Τα στοιχεία του διανύσματος $\mathbf{Y}_t$ ορίζονται ως εξής

$$Y_{tj} = \begin{cases} 1, & \text{αν η } j\text{-οστή κατηγορία παρατηρείται τον χρόνο } t \\ & \text{για} \qquad j = 1, 2 \ldots, q, \quad t = 1, 2, \ldots N \\ 0, & \text{αλλιώς} \end{cases} \tag{4.1}$$

Για παράδειγμα, αν την χρονική στιγμή t για το $\mathbf{Y}_t$ ισχύει $\mathbf{Y}_t = (1, 0, 0, \ldots, 0)'$, που σημαίνει ότι $Y_{t1} = 1$ και $Y_{tj} = 0$ για $j = 2, 3, \ldots, q$, τότε την δεδομένη χρονική στιγμή για την μεταβλητή ενδιαφέροντος θα ισχύει ότι $Y_t = 1$. Ακόμη αν την χρονική στιγμή t παρατηρήσουμε για το διάνυσμα $\mathbf{Y}_t$ την τιμή $\mathbf{Y}_t = (0, 0, 0, \ldots, 0)'$ δηλαδή $Y_{tj} = 0$ για $j = 1, 2, \ldots, q$ τότε θα ισχύει $Y_t = m$.

Παρατήρηση 4.2.1 Με βάση το παραπάνω παράδειγμα μπορούμε να πούμε ότι επειδή για κάθε χρονική στιγμή η Y_t έχει m δυνατά αποτελέσματα για την συγκεκριμένη

χρονική στιγμή θα υπάρχουν m δυνατές τιμές του διανύσματος $\mathbf{Y}_t$ οι οποίες θα αντιστοιχούν στις κατηγορίες του φαινομένου. Κάθε ένα εξ' αυτών των διανυσμάτων δεν μπορεί να έχει σε περισσότερες από μια θέσεις τον αριθμό 1.

Επανερχόμενοι στο παράδειγμα με την καθημερινή προτίμηση μεταφορικού μέσου ενός ατόμου και προσθέτοντας την επιλογή του ποδηλάτου τότε η τυχαία μεταβλητή Y_t έχει $m = 4$ δυνατές τιμές που θα είναι οι ακέραιοι 1, 2, 3, 4. Επομένως τον χρόνο t το αντίστοιχο διάνυσμα $\mathbf{Y}_t$ διάστασης $q = 3 \times 1$ θα έχει επίσης 4 δυνατές τιμές. Πράγματι σύμφωνα με την ανάθεση

$$\begin{gathered}\text{ποδήλατο} \longleftarrow 1 \\ \text{αυτοκίνητο} \longleftrightarrow 2 \\ \text{λεωφορείο} \longleftrightarrow 3 \\ \text{τρένο} \longleftrightarrow 4\end{gathered}$$

θα προκύψουν για την χρονική στιγμή t τα διανύσματα

$$\begin{aligned}\mathbf{Y}_t &= (1,0,0)' \text{ (αντιστοιχεί στην τιμή } Y_t = 1) \\ \mathbf{Y}_t &= (0,1,0)' \text{ (αντιστοιχεί στην τιμή } Y_t = 2) \\ \mathbf{Y}_t &= (0,0,1)' \text{ (αντιστοιχεί στην τιμή } Y_t = 3) \\ \mathbf{Y}_t &= (0,0,0)' \text{ (αντιστοιχεί στην τιμή } Y_t = 4)\end{aligned}$$

Επιλέγοντας μια διαφορετική ανάθεση εύκολα μπορούμε να διαπιστώσουμε ότι οι παραπάνω τέσσερις τιμές που προέκυψαν για το $\mathbf{Y}_t$ παραμένουν αναλλοίωτες. Βλέπουμε λοιπόν πως η περιγραφή του φαινομένου για κάθε χρονική στιγμή μέσω του διανύσματος $\mathbf{Y}_t$, που από εδώ και στο εξής θα το καλούμε «διάνυσμα κατάστασης» τον χρόνο t, μειώνει την αυθαιρεσία που παρατηρήθηκε από την ανάθεση ακεραίων στις κατηγορίες της μεταβλητής Y_t.

4.3 Θεωρητικό Πλαίσιο

Είναι φανερό ότι η δεσμευμένη κατανομή της τυχαίας μεταβλητής Y_{tj} (που παίρνει τις τιμές 1 αν εμφανιστεί η κατηγορία j και 0 αν δεν εμφανιστεί) δοθείσης της σ-άλγεβρας $\mathcal{F}_{t-1}$ ακολουθεί κατανομή $Bernoulli(1, \pi_{tj})$ όπου

$$\pi_{tj} = E(Y_{tj} \mid \mathcal{F}_{t-1}) = P(Y_{tj} = 1 \mid \mathcal{F}_{t-1})$$

για $j = 1, 2, \ldots, m$ και $t = 1, 2, \ldots, N$. Επομένως για κάθε χρονική στιγμή που εξελίσσεται το φαινόμενο στο διάνυσμα κατάστασης $\mathbf{Y}_t$ θα αντιστοιχεί το διάνυσμα $\pi_t = (\pi_{t1}, \pi_{t2}, \ldots, \pi_{tq})'$. Οι συνιστώσες του τελευταίου διανύσματος είναι οι δεσμευμένες πιθανότητες εμφάνισης της κάθε μιας από τις $1, 2, \ldots, q$ κατηγορίες την χρονική στιγμή t δοθείσης της ιστορίας $\mathcal{F}_{t-1}$ και ονομάζονται «*πιθανότητες μετάβασης*» (*"transition probabilities"*).

Ο καλύτερος τρόπος για να μειώσουμε την αβεβαιότητα που διέπει την έκβαση του φαινομένου την χρονική στιγμή t είναι να εκτιμήσουμε με όσο το δυνατό μεγαλύτερη ακρίβεια τις «*πιθανότητες μετάβασης*» καθώς και την π_{tm} που αντιστοιχούν στην δεδομένη χρονική στιγμή. Με σκοπό να συμπερασματολογήσουμε για τις παραμέτρους $\pi_{tj}, j = 1, 2, \ldots, m$, θα στηριχτούμε στην από κοινού κατανομή του δείγματος. Πλέον όμως αντί της δειγματοληπτικής διαδρομής των εξαρτημένων μεταβλητών

$$Y_1, Y_2, \ldots, Y_N$$

η στοχαστική πληροφόρηση συγκενρώνεται στο δείγμα των εξαρτημένων διανυσμάτων

$$\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{tm})', \quad t = 1, 2, \ldots, N.$$

Τα δεδομένα διανύσματα παίρνουν τιμές στο σύνολο των m διάστατων διανυσμάτων τα οποία έχουν ακριβώς σε μια θέση 1 και στις υπόλοιπες 0. Βλέπουμε λοιπόν ότι διαθέτουμε ένα μη τυχαίο δείγμα, μεγέθους N, από την m-διάστατη πολυωνυμική κατανομή

$$M_m(1; \pi_{t1}, \pi_{t2}, \ldots, \pi_{tm}), \sum_{j=1}^{m} \pi_{tj} = 1, \sum_{j=1}^{m} Y_{tj} = 1 \quad \mu\epsilon \quad \pi_{tj} \in (0, 1)$$

για $j = 1, 2, \ldots, m$, $t = 1, 2, \ldots, N$. Στο σημείο αυτό μπορούμε να εξηγήσουμε αναλυτικότερα γιατι χρησιμοποιήθηκε η πολυωνυμική κατανομή $M_m(1; \pi_{t1}, \ldots, \pi_{tm})$.

4.3.1 Η πολυωνυμική κατανομή στις κατηγορικές χρονοσειρές

Πρίν αναφερθούμε στην πολυωνυμική κατανομή όταν έχουμε ποιοτικές χρονοσειρές (δηλαδή όταν διαθέτουμε εξαρτημένα δεδομένα) θα υπενθυμίσουμε πώς ορίζεται η συγκεκριμένη κατανομή όταν έχουμε τυχαίο δείγμα. Έστω λοιπόν ότι εκτελούμε ένα πείραμα n ανεξάρτητες φορές. Σε κάθε επανάληψη μπορεί να συμβεί ένα από τα m ενδεχόμενα $A_1, A_2, \ldots, A_m$ με αντίστοιχες πιθανότητες πραγματοποίησης $\pi_1, \pi_2, \ldots, \pi_m$ που είναι ίδιες σε κάθε επανάληψη και για τις οποίες ισχύει η σχέση $\sum_{j=1}^{J} \pi_{tj} = 1$.

Έτσι αν n_j είναι ο αριθμός των φορών που πραγματοποιήθηκε το ενδεχόμενο A_j τότε η από κοινού κατανομή των τυχαίων μεταβλητών n_j για $j = 1, 2, \ldots, m$, γράφεται

$$f(n_1, n_2, \ldots, n_J; \pi_1, \pi_2, \ldots, \pi_J) = \frac{n!}{n_1! n_2! \ldots n_J!} \prod_{j=1}^{J} \pi_j^{n_j}. \tag{4.2}$$

Στην περίπτωση χρονολογικών σειρών με κατηγορικές μεταβλητές Y_t για $t = 1, 2, \ldots, N$ τα δεδομένα μας σχετικά με την μεταβλητή ενδιαφέροντος δεν συνιστούν τυχαίο δείγμα, αφού οι επαναλήψεις δεν είναι μεταξύ τους ανεξάρτητες. Κάθε χρονική στιγμή t που πραγματοποιείται λοιπόν το μη τυχαίο πείραμα διαθέτουμε ένα πληθυσμό μεγέθους $n = 1$ του οποίου η τελική τιμή(στην ουσία η τιμή της μεταβλητής ενδιαφέροντος Y_t) ανάμεσα στις τιμές $1, 2, \ldots, m$ δείχνει την έκβαση του φαινομένου τον δεδομένο χρόνο. Οι συνιστώσες $Y_{tj}, j = 1, 2, \ldots, m$ (σε αναλογία με τα n_j) είναι οι συχνότητες εμφάνισης των επιπέδων της μεταβλητής ενδιαφέροντος Y_t και προφανώς μόνο μια εξ' αυτών μπορεί να πάρει την τιμή 1. Έτσι θα ισχύει

$$\frac{1!}{y_{t1}!, y_{t2}!, \ldots, y_{tm}!} = 1 \tag{4.3}$$

Σύμφωνα με τις σχέσεις (2), (3) η δεσμευμένη από κοινού κατανομή των $y_{t1}, y_{t2}, \ldots, y_{tm}$ δοθείσης της ιστορίας $\mathcal{F}_{t-1}$ δίνεται από την σχέση

$$f(y_{t1}, y_{t2}, \ldots, y_{tm}; \pi_{t1}, \pi_{t2}, \ldots, \pi_{tm} \mid \mathcal{F}_{t-1}) = \prod_{j=1}^{m} \pi_{tj}^{y_{tj}}, \quad t = 1, 2, \ldots, N. \tag{4.4}$$

Επομένως σε αναλογία με το γεγονός ότι $Y_{tj} \mid \mathcal{F}_{t-1} \sim Bernoulli(1, \pi_{tj})$ θα ισχύει

$$\mathbf{Y}_t = (Y_{t1}, \ldots, Y_{tm})' \mid \mathcal{F}_{t-1} \sim M_m(1; \pi_{t1}, \ldots \pi_{tm}) \tag{4.5}$$

Για την κατασκευή στατιστικού μοντέλου μέσω του οποίου θα συμπερασματολογήσουμε για τις παραμέτρους $\pi_{t1}, \ldots, \pi_{tm}$ θα στηριχτούμε στην θεωρία των πολυδιάστατων γενικευμένων γραμμικών μοντέλων, αφού πρώτα δείξουμε ότι η πολυωνυμική κατανομή στην περίπτωση εξαρτημένων παρατηρήσεων ανήκει στην πολυδιάστατη εκθετική οικογένεια κατανομών την οποία επίσης θα επεκτείνουμε για την περίπτωση των κατηγορικών χρονοσειρών.

Για πολυδιάστατες τυχαίες μεταβλητές, δηλαδή για τυχαία διανύσματα η εκθετική οικογένεια στην περίπτωση που αυτά είναι ανεξάρτητα μεταξύ τους ορίζεται ως εξής

Ορισμός 4.3.1 *Έστω* $\mathbf{Y} = (Y_1, Y_2, \ldots, Y_m)' \in A \subseteq R^m$ *μια* m*-διάστατη τυχαία μεταβλητή της οποίας η συνάρτηση πυκνότητας πιθανότητας εξαρτάται από την* m*-διάστατη παράμετρο* $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)' \subseteq R^m$ *με* $m \geq 1$. *Η* $\mathbf{Y}$ *ανήκει στην π.Ε.Ο.Κ*

αν

$$f(\mathbf{y};\boldsymbol{\theta}) = c(\boldsymbol{\theta}) \exp\Big\{ \sum_{j=1}^{m} Q_j(\boldsymbol{\theta}) T_j(\mathbf{y}) \Big\} h(\mathbf{y}) I_A(\mathbf{y}) \tag{4.6}$$

όπου $\mathbf{y} = (y_1, y_2, \ldots, y_m)'$ *και* $y_i, \quad i = 1, 2, \ldots, m$ *είναι η παρατηρηθείσα τιμή της τυχαίας μεταβλητής* Y_i. *Οι* $Q_j(\cdot), T_j(\cdot), h(\cdot), c(\cdot)$ *είναι γνωστές συναρτήσεις. Για να ανήκει όμως η* $\mathbf{Y}$ *στην πολυμεταβλητή E.O.K εκτός του ότι απαιτείται η συνάρτηση πυκνότητας πιθανότητας της να γράφεται στην μορφή (6), συγχρόνως πρέπει για το πεδίο ορισμού της να ισχύει ότι* $A = \Big\{ \mathbf{y} \in R^m : f(\mathbf{y};\boldsymbol{\theta}) > 0 \Big\}$ *και να είναι ανεξάρτητο του αγνώστου παραμετρικού διανύσματος* $\boldsymbol{\theta}$. *Ακόμη η συνάρτηση* $h(\cdot)$ *είναι θετική επί του συνόλου A, και η σταθερά (υπο την έννοια ότι δεν εξαρτάται από το* $\mathbf{y}$*)* $c(\boldsymbol{\theta})$ *είναι επίσης θετική για κάθε* $\boldsymbol{\theta} \in R^m$.

Ο παραπάνω ορισμός της πολυμεταβλητής E.O.K ισχύει και στην περίπτωση που δεν έχουμε τυχαίο πείραμα. Έτσι όταν διαθέτουμε κατηγορικές χρονοσειρές θα λέμε ότι η $\mathbf{Y} \mid \mathcal{F}_{t-1}$ θα ανήκει στην π.E.O.K αν ισχύουν
(i)

$$f(\mathbf{y};\boldsymbol{\theta} \mid \mathcal{F}_{t-1}) = c(\boldsymbol{\theta}) \exp\Big\{ \sum_{j=1}^{m} Q_j(\boldsymbol{\theta}) T_j(\mathbf{y}) \Big\} h(\mathbf{y}) I_A(\mathbf{y}) \tag{4.7}$$

(ii)

$$A = \Big\{ \mathbf{y} \in R^m : f(\mathbf{y};\boldsymbol{\theta} \mid \mathcal{F}_{t-1}) > 0 \Big\}.$$

Η πολυωνυμική κατανομή στην περίπτωση των κατηγορικών χρονολογικών σειρών ανήκει στην πολυδιάστατη εκθετική οικογένεια κατανομών αφού γράφεται σύμφωνα με την σχέση (7) και το σύνολο A είναι ανεξάρτητο του $\boldsymbol{\theta}$. Αναλυτικότερα θα ισχύει $c(\theta) = 1$, $h(\mathbf{y}) = 1$, $T_j(y_j) = y_j$ και $Q_j(\theta) = \log \pi_j$.

Έχοντας παρουσιάσει την π.E.O.K θα εισαγάγουμε το πολυμεταβλητό γενικευμένο γραμμικό μοντέλο πάνω στο οποίο θα στηρίξουμε την συμπερασματολογία μας σχετικά με τις πιθανότητες $\pi_{tj}, j = 1, 2, \ldots, m$. Με σκοπό να μειώσουμε την διάσταση του προβλήματος ορίζουμε

$$Y_{tm} = 1 - \sum_{j=1}^{q} Y_{tj} \tag{4.8}$$

και

$$\pi_{tm} = 1 - \sum_{j=1}^{q} \pi_{tj}. \tag{4.9}$$

Παρατήρηση 4.3.1 Η μείωση της διάστασης του προβλήματος από m σε $q = m-1$ μέσω των σχέσεων (4.8) και (4.9) είναι αναγκαία από θεωρητική σκοπιά. Διευκρινίζοντας, σύμφωνα με την διαδικασία που θα παρουσιάσουμε επιδιώκεται αρχικά η εύρεση του $MPLE$ του $\boldsymbol{\beta}$ και εν συνεχεία η εκτίμηση των πιθανοτήτων $\pi_{tj}(\boldsymbol{\beta}), j = 1, 2, \ldots, m$, ώστε να αποκτήσουμε μια εικόνα για την εξέλιξη του φαινομένου. Αν δεν μειώσουμε την διάσταση του προβλήματος θα συναντήσαμε πρόβλημα στον πίνακα διακυμάνσεων-συνδιακυμάνσεων $\boldsymbol{\Sigma}_t(\boldsymbol{\beta})$, ο οποίος σε αυτή την περίπτωση θα ήταν διάστασης $m \times m$ και θα δινόταν από την σχέση

$$\sigma_t^{(ij)}(\boldsymbol{\beta}) = \begin{cases} -\pi_{ti}(\boldsymbol{\beta})\pi_{tj}(\boldsymbol{\beta}) & , \text{ αν } i \neq j \\ \pi_{ti}(\boldsymbol{\beta})(1 - \pi_{ti}(\boldsymbol{\beta})) & , \text{ αν } i = j \end{cases},$$

όπου $j = 1, 2, \ldots, m$. Αυτό συμβαίνει διότι ο $MPLE$ του δεδομένου πίνακα, που είναι ο τυχαίος πίνακας $S_t(\hat{\boldsymbol{\beta}})$ είναι ιδιάζων. Το πρόβλημα αυτό λύνεται αν εξ᾽ αρχής αγνοήσουμε την τελευταία συντεταγμένη Y_{tm} του διανύσματος $(Y_{t1}, Y_{t2}, \ldots, Y_{tm})'$, που στον χρόνο t δηλώνει την αντίστοιχη πολυωνυμική δοκιμή. Τότε ο $MPLE$ του πίνακα $\boldsymbol{\Sigma}_t$ είναι ο $\mathbf{S}_t$ αν του αφαιρέσουμε την m^{η} γραμμή και την m^{η} στήλη. Για ευκολία τον τελευταίο πίνακα εξακολουθούμε να τον συμβολίζουμε με $\mathbf{S}_t$. Στο σημείο αυτό διακρίνεται και ο ουσιαστικός λόγος που θεωρήσαμε τις σχέσεις (4.8) και (4.9).

Το μοντέλο λοιπόν που θα παρουσιάσουμε στον χρόνο t θα εκτιμά τις πιθανότητες $\pi_{t1}, \ldots, \pi_{tq}$ και εν συνεχεία μέσω της σχέσης (4.9) θα λαμβάνουμε και την εκτίμηση της π_{tm}. Στις δίτιμες κατηγορικές χρονοσειρές είχαμε καταλήξει στο γενικό μοντέλο (3.6), όπου η $h : R \to [0, 1]$ ήταν η αντίστροφη της συνάρτησης σύνδεσης, $\boldsymbol{\beta}$ το p-διάστατο σταθερό διάνυσμα των παραμέτρων και $\mathbf{Z}_{t-1}$ το p- διάστατο διάνυσμα των τυχαίων χρονοεξαρτώμενων συμμεταβλητών.

Στην περίπτωση των κατηγορικών χρονοσειρών επιθυμούμε για κάθε χρόνο t να συμπερασματολογήσουμε σχετικά με τις δεσμευμένες πιθανότητες εμφάνισης της κάθε μιας από τις $1, 2, \ldots, m$ δυνατές κατηγορίες της Y_t που αντιστοίχως είναι οι $\pi_{t1}, \pi_{t2}, \ldots, \pi_{tm}$. *Για την κατασκευή κατάλληλου μοντέλου για αυτό το σκοπό απαιτείται να λάβουμε υπόψη ότι οι μεταβλητές απόκρισης $Y_{tj}, j = 1, 2, \ldots, m$ είναι συσχετισμένες μεταξύ τους. Κατ᾽ επέκταση και οι πιθανότητες $\pi_{tj}, j = 1, 2, \ldots, m$, αλληλοεπηρεάζονται.* Είναι πλέον ορατό ότι για την συμπερασματολογία των δεσμευμένων πιθανοτήτων επιτυχίας για το κάθε επίπεδο του μη τυχαίου πειράματος, δεν είναι δυνατό να στηριχτούμε σε m ξεχωριστά μοντέλα της μορφής (3.6), αλλα απαιτείται να προχωρήσουμε την ανάλυση μας αξιοποιώντας τις μεθόδους της πολυμεταβλητής ανάλυσης.

Το πολυδιάστατο μοντέλο που θα παρουσιάσουμε εξετάζει από κοινού τις πιθανότητες $\pi_{t1}, \ldots, \pi_{tq}$, ενώ για την συμπερασματολογία σχετικά με την π_{tm} στηρίζεται στην σχέση (4.9). Αναλυτικότερα, σε κάθε αποκριτική μεταβλητή Y_{tj} για τις πρώτες q κατηγορίες αντιστοιχεί ένα p-διάστατο διάνυσμα τυχαίων χρονοεξαρτώμενων συμμεταβλητών (σταθεράς μορφής για κάθε χρόνο t) που ορίζεται ως εξής:

$$\mathbf{Z}_{(t-1)j} = (Z_{(t-1)j1}, Z_{(t-1)j2}, \ldots, Z_{(t-1)jp})', j = 1, 2, \ldots, q.$$

Τα διανύσματα $\mathbf{Z}_{(t-1)j}, j = 1, 2, \ldots, q$, αποτελούν τις στήλες του $p \times q$ πίνακα

$$\mathbf{Z}_{t-1} = \begin{pmatrix} Z_{(t-1)11} & Z_{(t-1)21} & \ldots & Z_{(t-1)q1} \\ Z_{(t-1)12} & Z_{(t-1)22} & \ldots & Z_{(t-1)q2} \\ \vdots & \vdots & \ddots & \vdots \\ Z_{(t-1)1p} & Z_{(t-1)2p} & \ldots & Z_{(t-1)qp} \end{pmatrix},$$

ο οποίος για διαδοχικές χρονικές στιγμές συνιστά την διανυσματική συμμεταβλητή διαδικασία $\{\mathbf{Z}_{t-1}\}, t = 1, 2, \ldots, N$. Η πιθανότητα εμφάνισης της $j = 1, 2, \ldots, q$ κατηγορίας την χρονική στιγμή t δεν θα επηρεάζεται μόνο από το διάνυσμα $\mathbf{Z}_{(t-1)j}$. Επειδή όπως έχουμε προαναφέρει, οι μεταβλητές απόκρισης $Y_{tj}, j = 1, 2, \ldots, m$ είναι εξαρτημένες η π_{tj} θα επηρεάζεται και από τα διανύσματα συμμεταβλητών των υπολοίπων κατηγοριών

$$\mathbf{Z}_{(t-1)1}, \mathbf{Z}_{(t-1)2}, \ldots, \mathbf{Z}_{(t-1),j-1}, \mathbf{Z}_{(t-1),j+1}, \ldots, \mathbf{Z}_{(t-1)q},$$

μέσω ενός σταθερού p-διάστατου διανύσματος $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots, \beta_p)'$. Αναλυτικότερα θα έχουμε

$$\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta} = \begin{pmatrix} \sum_{k=1}^{p} Z_{(t-1)1k}\beta_k \\ \sum_{k=1}^{p} Z_{(t-1)2k}\beta_k \\ \vdots \\ \sum_{k=1}^{p} Z_{(t-1)qk}\beta_k \end{pmatrix} = \begin{pmatrix} \eta_{t1} \\ \eta_{t2} \\ \vdots \\ \eta_{tq} \end{pmatrix} = \boldsymbol{\eta}_t \quad .$$

Με σκοπό να γίνουν κατανοητά τα προαναφερθέντα αναφέρουμε ένα παράδειγμα. Έστω λοιπόν η κατηγορική χρονοσειρά $\{Y_t\}, t = 1, 2, \ldots, N$, με $m = 4$ δυνατές τιμές για την Y_t, τις 1, 2, 3, 4. Θεωρώντας ότι η Y_t επηρεάζεται απο την Y_{t-1}, Y_{t-2} την X_{t-1} και την W_t τότε το διάνυσμα $\mathbf{Z}_{(t-1)j}$ θα είναι διάστασης $p = 4$ και θα έχει την μορφή

$$\mathbf{Z}_{(t-1)j} = (Y_{(t-1)j}, Y_{(t-2)j}, X_{t-1}, W_t)', j = 1, 2, 3(= q).$$

Επομένως ο πίνακας $\mathbf{Z}_{t-1}$ θα είναι διάστασης 4×3 και θα έχει την ακόλουθη μορφή

$$\mathbf{Z}_{t-1} = \begin{pmatrix} Y_{(t-1)1} & Y_{(t-1)2} & Y_{(t-1)3} \\ Y_{(t-2)1} & Y_{(t-2)2} & Y_{(t-2)3} \\ X_{t-1} & X_{t-1} & X_{t-1} \\ W_t & W_t & W_t \end{pmatrix}.$$

Παρατηρούμε ότι στην 1^{η} και 2^{η} γραμμή του παραπάνω πίνακα βρίσκονται αντίστοιχα τα διανύσματα κατάστασης της Y_t για τις χρονικές στιγμές $t-1$ και $t-2$. Αν για τον χρόνο $t-1$ ισχύει $Y_{t-1} = 1$ δηλαδή $(Y_{(t-1)1}, Y_{(t-1)2}, Y_{(t-1)3})' = (1,0,0)'$, ενώ για την χρονική στιγμή $t-2$ έχουμε $Y_{t-2} = 2$ δηλαδή $(Y_{(t-2)1}, Y_{(t-2)2}, Y_{(t-2)3})' = (0,1,0)'$, με $X_{t-1} = x_{t-1}$ και $W_t = w_t$, τότε ο $\mathbf{Z}_{t-1}$ λαμβάνει την ακόλουθη τιμή

$$\mathbf{Z}_{t-1} = \begin{pmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ x_{t-1} & x_{t-1} & x_{t-1} \\ w_t & w_t & w_t \end{pmatrix}.$$

Για τις $\pi_{tj}, j = 1, 2, \ldots, q$ θα ισχύει

$$\pi_{tj}(\boldsymbol{\beta}) = h_j(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta})$$

όπου η $h_j : R^q \to R$ είναι μια γνωστή μονότονη συνάρτηση. Γίνεται λοιπόν φανερό πως η κάθε πιθανότητα μετάβασης συνδέεται, μέσω του διανύσματος $\boldsymbol{\beta}$, και με τα q διανύσματα τα οποία επηρεάζουν τις q μεταβλητές απόκρισης. Με σκοπό λοιπόν να μελετήσουμε ταυτόχρονα τις πιθανότητες $\pi_{t1}, \ldots, \pi_{tm}$ θεωρούμε το πολυμεταβλητό μοντέλο

$$\boldsymbol{\pi}_t(\boldsymbol{\beta}) = \begin{pmatrix} \pi_{t1}(\boldsymbol{\beta}) \\ \pi_{t2}(\boldsymbol{\beta}) \\ \vdots \\ \pi_{tq}(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} P(Y_{t1} = 1 \mid \mathcal{F}_{t-1}) \\ P(Y_{t2} = 1 \mid \mathcal{F}_{t-1}) \\ \vdots \\ P(Y_{tq} = 1 \mid \mathcal{F}_{t-1}) \end{pmatrix} = \begin{pmatrix} h_1(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) \\ h_2(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) \\ \vdots \\ h_q(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) \end{pmatrix} \tag{4.10}$$

ή πιο συνεπτυγμένα

$$\boldsymbol{\pi}_t(\boldsymbol{\beta}) = E(\mathbf{Y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{h}(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) = \boldsymbol{h}(\boldsymbol{\eta}_t). \tag{4.11}$$

Η εξίσωση (4.11) αποτελεί την γενική μορφή του πολυμεταβλητού γενικευμένου γραμμικού μοντέλου για κατηγορικές χρονοσειρές, και έχει μελετηθεί από αρκετούς συγγραφείς (*Fahrmeir* και *Kaufmann* (1987)), (*Pruscha* (1993)). Η συνάρτηση $\mathbf{h}(\cdot) =$

$(h_1(\cdot), h_2(\cdot), \ldots, h_q(\cdot))' : R^q \to R^q$ αποτελεί, όπως και στην μονοδιάστατη περίπτωση, την αντίστροφη της συνάρτησης σύνδεσης. Επειδή οι "πιθανότητες μετάβασης" ανήκουν στο διάστημα [0,1], θεωρούμε ότι η $\mathbf{h}$ απεικονίζεται αμφιμονοσήμαντα στο υποσύνολο H του R^q το οποίο ορίζεται ως ακολούθως:

$$\{(\omega_1, \omega_2, \ldots, \omega_q)' : \omega_j > 0, j = 1, 2, \ldots, q \quad \kappa\alpha\iota \quad \sum_{j=1}^{q} \omega_j < 1\}$$

(Με τον τρόπο αυτό και η π_{tm} ανήκει στο [0,1].)

Μέσω του μοντέλου (4.11) μπορούμε να παρατηρήσουμε την ειδική περίπτωση των δίτιμων χρονοσειρών που παρουσιάσαμε στο προηγούμενο κεφάλαιο. Αυτό συμβαίνει αν $m = 2$ οπότε $q = 1$ και επομένως η συμμεταβλητή διαδικασία ανάγεται στο $p-$διάστατο διάνυσμα $\mathbf{Z}_{t-1}$ και η εξίσωση (4.11) παίρνει την μορφή (3.6). Όπως θα δούμε και σε επόμενη ενότητα, η επιλογή της $q-$διάστατης συνάρτησης $\boldsymbol{h}$ οδηγεί σε μια σειρά μοντέλων παλινδρόμησης για κατηγορικές χρονοσειρές. Πρίν προχωρήσουμε στην παρουσίαση των σχετικών μοντέλων θα δείξουμε πώς πραγματοποιείται η εκτίμηση των παραμέτρων $\boldsymbol{\beta}$ του γενικού μοντέλου (4.11) μέσω της μεθοδολογικής προσέγγισης της μερικής πιθανοφάνειας.

4.4 Συμπερασματολογία όταν η Y_t έχει $m = 3$ επίπεδα

Έστω η κατηγορική χρονοσειρά $\{Y_t\}, t = 1, 2, \ldots, N$, όπου η μεταβλητή ενδιαφέροντος Y_t σε κάθε επανάληψη μπορεί να πάρει $m = 3$ δυνατές τιμές που αντικατοπρίζουν τις $m = 3$ δυνατές εκβάσεις του φαινομένου, έστω τις A_1, A_2, A_3. Όπως είδαμε η κατάσταση του φαινομένου τον χρόνο t μπορεί να αποδοθεί μέσω του τυχαίου διανύσματος $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, Y_{t3})'$ του οποίου οι συνιστώσες παίρνουν τις τιμές '1' ή '0' και το οποίο μπορεί να έχει το '1' μόνο σε μια θέση. Σύμφωνα με την σχέση (4.5) θα ισχύει

$$\mathbf{Y}_t = (Y_{t1}, Y_{t2}, Y_{t3})' \mid \mathcal{F}_{t-1} \sim M_m(1; \pi_{t1}, \pi_{t2}, \pi_{t3}) \tag{4.12}$$

με $\sum_{j=1}^{3} Y_{tj} = 1$ και $\sum_{j=1}^{3} \pi_{tj} = 1$.
Θέτοντας

$$Y_{t3} = 1 - (Y_{t1} + Y_{t2}) \tag{4.13}$$

και

$$\pi_{t3} = 1 - (\pi_{t1} + \pi_{t2}), \tag{4.14}$$

ώστε να αποφύγουμε τα προβλήματα που αναφέραμε στην Παρατήρηση 4.3.1, θα προχωρήσουμε σε συμπερασματολογία σχετικά με τις δεσμευμένες ροπές του διανύσματος $\mathbf{Y}_t = (Y_{t1}, Y_{t2})'$. Για την δεσμευμένη μέση τιμή του $\mathbf{Y}_t$ έχουμε

$$E(\mathbf{Y}_t \mid F_{t-1}) = \begin{pmatrix} E(Y_{t1} \mid \mathcal{F}_{t-1}) \\ E(Y_{t2} \mid \mathcal{F}_{t-1}) \end{pmatrix} = \begin{pmatrix} \pi_{t1} \\ \pi_{t2} \end{pmatrix} \tag{4.15}$$

αφού $Y_{tj} \mid \mathcal{F}_{t-1} \sim Bernoulli(1, \pi_{tj})$ για $j = 1, 2$ με $\pi_{tj} = P(Y_{tj} = 1 \mid \mathcal{F}_{t-1})$. Επίσης ισχύει

$$Var(\mathbf{Y}_t \mid \mathcal{F}_{t-1}) = \mathbf{\Sigma}_t = \begin{pmatrix} Var(Y_{t1} \mid \mathcal{F}_{t-1}) & Cov(Y_{t1}, Y_{t2} \mid \mathcal{F}_{t-1}) \\ Cov(Y_{t2}, Y_{t1} \mid \mathcal{F}_{t-1}) & Var(Y_{t2} \mid \mathcal{F}_{t-1}) \end{pmatrix}$$

και τελικά

$$Var(\mathbf{Y}_t \mid \mathcal{F}_{t-1}) = \begin{pmatrix} \pi_{t1}(1-\pi_{t1}) & -\pi_{t1}\pi_{t2} \\ -\pi_{t2}\pi_{t1} & \pi_{t2}(1-\pi_{t2}) \end{pmatrix}. \tag{4.16}$$

Ο καλύτερος τρόπος για αναπαράγουμε την παρατηρούμενη πολυπλοκότητα του φαινομένου είναι να εκτιμήσουμε με μεγάλη ακρίβεια τις δεσμευμένες ροπές $E(\mathbf{Y}_t \mid \mathcal{F}_{t-1})$ και $Var(\mathbf{Y}_t \mid \mathcal{F}_{t-1})$. Σύμφωνα με τις σχέσεις (4.15) και (4.16) οι δεδομένες ροπές εξαρτώνται από τις πιθανότητες π_{t1}, π_{t2} των ενδεχομένων A_1, A_2 και επομένως για να αυξήσουμε τον βαθμό της στοχαστικής πληροφόρησης είναι πλέον ανάγκη να εκτιμηθούν οι προαναφερθείσες πολυωνυμικές πιθανότητες. Οι δεδομένες όμως δεσμευμένες πιθανότητες επιτυχίας βάσει του γενικού πολυμεταβλητού μοντέλου παλινδρόμησης (4.11) εκφράζονται συναρτήσει του σταθερού διανύσματος $\boldsymbol{\beta}$ το οποίο θα εκτιμηθεί μέσω της μερικής πιθανοφάνειας της δειγματοληπτικής διαδρομής που διαθέτουμε. Η μερική πιθανοφάνεια στην προκειμένη περίπτωση ορίζεται ως εξής

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^{N} f_t(\boldsymbol{y}_t; \boldsymbol{\beta}) \tag{4.17}$$

όπου $f_t(\boldsymbol{y}_t; \boldsymbol{\beta}) \equiv f_{\mathbf{Y}_t}(\boldsymbol{y}_t; \boldsymbol{\theta} \mid \mathcal{F}_{t-1})$. Έτσι λαμβάνοντας υπόψη την (4.4) και το γεγονός ότι $\pi_{tj} \equiv \pi_{tj}(\boldsymbol{\beta})$, σύμφωνα με τα προαναφερθέντα, η (4.17) γράφεται

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^{N} \prod_{j=1}^{3} [\pi_{tj}(\boldsymbol{\beta})]^{y_{tj}}. \tag{4.18}$$

Για την εύρεση του $MPLE$ του $\boldsymbol{\beta}$ αντί της συνάρτησης μερικής πιθανοφάνειας (4.18) θα μεγιστοποιήσουμε τον λογάριθμό της που δίνεται από την σχέση

$$\ell(\boldsymbol{\beta}) \equiv \log PL(\boldsymbol{\beta}) = \sum_{t=1}^{N} \sum_{j=1}^{3} y_{tj} \log \pi_{tj}(\boldsymbol{\beta}). \tag{4.19}$$

Αναλυτικότερα για τον λογάριθμο της συνάρτησης μερικής πιθανοφάνειας θα έχουμε

$$\ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \{y_{t1} \log \pi_{t1}(\boldsymbol{\beta}) + y_{t2} \log \pi_{t2}(\boldsymbol{\beta}) + y_{t3} \log \pi_{t3}(\boldsymbol{\beta})\}$$

ή λόγω των (4.13) και (4.14)

$$\ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \{y_{t1} \log \pi_{t1}(\boldsymbol{\beta}) + y_{t2} \log \pi_{t2}(\boldsymbol{\beta}) + (1 - y_{t1} - y_{t2}) \log(1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta}))\}$$

ή

$$\begin{aligned} \ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \{ & y_{t1} \log(\frac{\pi_{t1}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})}) + y_{t2} \log(\frac{\pi_{t2}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})}) \\ & + \log(1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta}))\}. \end{aligned} \tag{4.20}$$

Στην περίπτωση των δίτιμων χρονοσειρών είχαμε δεί ότι για την φυσική παράμετρο θ_t ισχύει

$$\theta_t = \log(\frac{\pi_t(\boldsymbol{\beta})}{1 - \pi_t(\boldsymbol{\beta})}) \equiv \text{logit}(\pi_t(\boldsymbol{\beta})). \tag{4.21}$$

Σε αναλογία με την σχέση (4.21) μπορούμε να ορίσουμε το φυσικό παραμετρικό διάνυσμα

$$\begin{aligned} \boldsymbol{\theta}_t(\boldsymbol{\beta}) &= (\theta_{t1}(\boldsymbol{\beta}), \theta_{t2}(\boldsymbol{\beta}))' \\ &= \Big(\log(\frac{\pi_{t1}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})}), \log(\frac{\pi_{t2}(\boldsymbol{\beta})}{1 - \pi_{t1}(\boldsymbol{\beta}) - \pi_{t2}(\boldsymbol{\beta})}) \Big)'. \end{aligned} \tag{4.22}$$

Το δεδομένο διάνυσμα το οποίο προφανώς εξαρτάται από το $\boldsymbol{\beta}$ έχει την ίδια διάσταση με το διάνυσμα $(Y_{t1}, Y_{t2})'$. Σύμφωνα με την (4.22), η (4.20) παίρνει την μορφή

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \sum_{t=1}^{N} \{y_{t1}\theta_{t1}(\boldsymbol{\beta}) + y_{t2}\theta_{t2}(\boldsymbol{\beta}) - \log[1 + \exp(\theta_{t1}(\boldsymbol{\beta})) + \exp(\theta_{t2}(\boldsymbol{\beta}))]\} \\ &= \sum_{t=1}^{N} \ell_t(\boldsymbol{\beta}). \end{aligned} \tag{4.23}$$

Με σκοπό να ευρεθεί ο εκτιμητής μέγιστης μερικής πιθανοφάνειας του $\boldsymbol{\beta}$ απαιτείται να λυθούν οι εξισώσεις

$$\mathbf{S}_N(\boldsymbol{\beta}) = \mathbf{0} \Leftrightarrow (\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_1}, \ldots, \frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_p})' = \mathbf{0}$$

οι οποίες είναι συνήθως μη γραμμικές και απαιτούν μεθόδους αριθμητικής ανάλυσης. Γίνεται φανερό πως αρχικά πρέπει να υπολογιστούν οι συνιστώσες $\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j}, j =$

$1, 2, \ldots, p$, του διανύσματος των μερικών σκορ $\mathbf{S}_N(\boldsymbol{\beta})$. Σύμφωνα με την (4.23) θα ισχύει

$$\frac{\partial \ell(\boldsymbol{\beta})}{\partial \beta_j} = \sum_{t=1}^{N} \frac{\partial \ell_t(\boldsymbol{\beta})}{\partial \beta_j}, j = 1, 2, \ldots, p. \tag{4.24}$$

Όμως με βάση πάλι την (4.23) το $\ell_t(\boldsymbol{\beta})$ δεν είναι άμεσα εκφρασμένο ως συνάρτηση του διανύσματος $\boldsymbol{\beta}$ και προφανώς είναι ανάγκη να χρησιμοποιήσουμε έναν κανόνα αλυσίδας για τον υπολογισμό της μερικής παραγώγου του ως προς β_j. Σε αναλογία με τον κανόνα αλυσίδας που χρησιμοποιούμε στην μονοδιάστατη περίπτωση (βλέπε σχέση (A.9) του Παραρτήματος Α) για την εύρεση του $\partial \ell_t(\boldsymbol{\beta})/\partial \beta_j$, θα εφαρμόσουμε έναν κανόνα που θα αναφέρεται σε πολυμεταβλητές συναρτήσεις και ο οποίος θα λαμβάνει υπόψη την ειδική περίπτωση ποιοτικών χρονοσειρών που εξετάζουμε ($m = 3$). Έτσι θα έχουμε

$$\frac{\partial \ell_t}{\partial \boldsymbol{\beta}'} = \frac{\partial \ell_t}{\partial \boldsymbol{\theta}_t'} \frac{\partial \boldsymbol{\theta}_t}{\partial \boldsymbol{\pi}_t'} \frac{\partial \boldsymbol{\pi}_t}{\partial \boldsymbol{\eta}_t'} \frac{\partial \boldsymbol{\eta}_t}{\partial \boldsymbol{\beta}'}, \tag{4.25}$$

όπου $\frac{\partial \ell_t}{\partial \boldsymbol{\beta}'} = (\frac{\partial \ell_t}{\partial \beta_1}, \ldots, \frac{\partial \ell_t}{\partial \beta_p})$. Επειδή $m = 3 \Rightarrow q = 2$ και επομένως θεωρούμε το διάνυσμα κατάστασης $\mathbf{Y}_t = (Y_{t1}, Y_{t2})'$. Ακόμη για το διάνυσμα των πιθανοτήτων μετάβασης ισχύει $\pi_t = (\pi_{t1}, \pi_{t2})'$ ενώ για το διάνυσμα των γραμμικών προβλέψεων προκύπτει $\boldsymbol{\eta}_t = \mathbf{Z}_{t-1}'\boldsymbol{\beta} = (\eta_{t1}, \eta_{t2})'$. Για τις ποσότητες του $2^{ου}$ μέλους της σχέσης (4.25) θα έχουμε

$$\begin{aligned} \frac{\partial \ell_t}{\partial \boldsymbol{\theta}_t'} &= (\frac{\partial \ell_t}{\partial \theta_{t1}}, \frac{\partial \ell_t}{\partial \theta_{t2}})' \\ &= (Y_{t1} - \pi_{t1}, Y_{t2} - \pi_{t2}) = (Y_{t1}, Y_{t2}) - (\pi_{t1}, \pi_{t2}) = (\mathbf{Y}_t - \boldsymbol{\pi}_t)', \end{aligned} \tag{4.26}$$

$$\frac{\partial \boldsymbol{\theta}_t}{\partial \boldsymbol{\pi}_t'} = \begin{bmatrix} \frac{\partial \theta_{t1}}{\partial \pi_{t1}} & \frac{\partial \theta_{t1}}{\partial \pi_{t2}} \\ \frac{\partial \theta_{t2}}{\partial \pi_{t1}} & \frac{\partial \theta_{t2}}{\partial \pi_{t2}} \end{bmatrix} = \begin{bmatrix} \frac{1-\pi_{t2}}{\pi_{t1}(1-\pi_{t1}-\pi_{t2})} & \frac{1}{1-\pi_{t1}-\pi_{t2}} \\ \frac{1}{1-\pi_{t1}-\pi_{t2}} & \frac{1-\pi_{t1}}{\pi_{t1}(1-\pi_{t1}-\pi_{t2})} \end{bmatrix}. \tag{4.27}$$

Θέτοντας για λόγους ευκολίας τον πίνακα συνδιακύμανσης του διανύσματος $(Y_{t1}, Y_{t2})'$ δοθείσης της ιστορίας $\mathcal{F}_{t-1}$ με Σ_t, τότε σύμφωνα με τις (4.16) και (4.26) εύκολα μπορούμε να παρατηρήσουμε ότι ισχύει

$$\boldsymbol{\Sigma}_t^{-1} = \frac{\partial \boldsymbol{\theta}_t}{\partial \boldsymbol{\pi}_t'}. \tag{4.28}$$

Ακόμη,

$$\frac{\partial \boldsymbol{\pi}_t}{\partial \boldsymbol{\eta}_t'} = \begin{bmatrix} \frac{\partial \pi_{t1}}{\partial \eta_{t1}} & \frac{\partial \pi_{t1}}{\partial \eta_{t2}} \\ \frac{\partial \pi_{t2}}{\partial \eta_{t1}} & \frac{\partial \pi_{t2}}{\partial \eta_{t2}} \end{bmatrix} \overset{(12)}{=} \frac{\partial \boldsymbol{h}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t'} \equiv \mathbf{D}_t', \tag{4.29}$$

όπου $\boldsymbol{h}(\boldsymbol{\eta}_t) = (h_1(\eta_{1t}), h_2(\eta_{2t}))'$. Επιπρόσθετα

$$\frac{\partial \boldsymbol{\eta}_t}{\partial \boldsymbol{\beta}'} = \mathbf{Z}'_{t-1}. \tag{4.30}$$

Έτσι αντικαθιστώντας τις (4.26),(4.27),(4.28),(4.29),(4.30) στην (4.25) προκύπτει

$$\underbrace{\frac{\partial \ell_t}{\partial \boldsymbol{\beta}'}}_{1\times p} = \underbrace{(\mathbf{Y}_t - \boldsymbol{\pi}_t)'}_{1\times 2}\underbrace{\Sigma_t^{-1}}_{2\times 2}\underbrace{\mathbf{D}'_t}_{2\times 2}\underbrace{\mathbf{Z}'_{t-1}}_{2\times p}. \tag{4.31}$$

Ακόμη, για το παράδειγμά μας όπου $m = 3$ και $q = 2$ έχουμε

$$\mathbf{Z}_{t-1} = \begin{bmatrix} Z_{(t-1)11} & Z_{(t-1)21} \\ Z_{(t-1)12} & Z_{(t-1)22} \\ \vdots & \vdots \\ Z_{(t-1)1p} & Z_{(t-1)2p} \end{bmatrix}.$$

Επομένως για το μερικό σκορ θα ισχύει σύμφωνα με τις σχέσεις (4.25),(4.31)

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \frac{\partial \ell_t}{\partial \boldsymbol{\beta}} = \sum_{t=1}^{N} \Big(\frac{\partial \ell_t}{\partial \boldsymbol{\beta}'}\Big)' \Rightarrow$$

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} [(\mathbf{Y}_t - \boldsymbol{\pi}_t)'\boldsymbol{\Sigma}_t^{-1}\mathbf{D}'_t\mathbf{Z}'_{t-1}]' \Rightarrow$$

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta})(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})) \tag{4.32}$$

(αφού ο πίνακας $\boldsymbol{\Sigma}_t^{-1}$ είναι συμμετρικός). Μια εναλλακτική μορφή για το διάνυσμα των μερικών σκορ είναι

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})) \tag{4.33}$$

όπου

$$\mathbf{U}_t(\boldsymbol{\beta}) = \mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{u}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t}. \tag{4.34}$$

Η συνάρτηση $\boldsymbol{u}(\cdot)$ είναι ανάλογη της u που ορίζουμε στην μονοδιάστατη περίπτωση (βλέπε σχέσεις (A.7) και (A.8) στο Παράρτημα A2.) Στην δεδομένη περίπτωση η $\boldsymbol{u}(\cdot)$ είναι διδιάστατη και αποτελεί την σύνθεση των $\boldsymbol{\theta}_t$ και $\boldsymbol{h}$. Συγκεκριμένα θα ισχύει

$$\begin{aligned} \boldsymbol{u} &= (u_1, u_2)' = (\theta_{t1}(\boldsymbol{h}), \theta_{t2}(\boldsymbol{h}))' = \\ &= \Big(\log\Big(\frac{h_1(\boldsymbol{\eta}_t)}{1 - h_1(\boldsymbol{\eta}_t) - h_2(\boldsymbol{\eta}_t)}\Big), \log\Big(\frac{h_2(\boldsymbol{\eta}_t)}{1 - h_1(\boldsymbol{\eta}_t) - h_2(\boldsymbol{\eta}_t)}\Big)\Big)' \\ &\overset{(12)}{=} \Big(\log\Big(\frac{\pi_{t1}(\boldsymbol{\eta}_t)}{1 - \pi_{t1}(\boldsymbol{\eta}_t) - \pi_{t2}(\boldsymbol{\eta}_t)}\Big), \log\Big(\frac{\pi_{t2}(\boldsymbol{\eta}_t)}{1 - \pi_{t1}(\boldsymbol{\eta}_t) - \pi_{t2}(\boldsymbol{\eta}_t)}\Big)\Big)'. \end{aligned} \tag{4.35}$$

Η απόδειξη της (4.34) παρατίθεται στο Παράρτημα Γ.

Για τον αθροιστικό κατά συνθήκη πίνακα πληροφορίας (*cumulative conditional information matrix*) που είναι διάστασης $p \times p$ ισχύει

$$\mathbf{G}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t(\boldsymbol{\beta})\mathbf{U}_t'(\boldsymbol{\beta})\mathbf{Z}_{t-1}'. \tag{4.36}$$

Απόδειξη της (4.36).

$$\begin{aligned}
\mathbf{G}_N(\boldsymbol{\beta}) &= Cov(\mathbf{S}_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}) \\
&\overset{(4.32)}{=} Cov(\sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta})(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})) \mid \mathcal{F}_{t-1}) \\
&= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta})Cov[\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})) \mid \mathcal{F}_{t-1}](\mathbf{Z}_{t-1}\mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}))' \\
&= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta})\boldsymbol{\Sigma}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta})\mathbf{D}_t'(\boldsymbol{\beta})\mathbf{Z}_{t-1}' \\
&\overset{(4.34)}{=} \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t(\boldsymbol{\beta})\mathbf{U}_t'(\boldsymbol{\beta})\mathbf{Z}_{t-1}'.
\end{aligned}$$

Για τον παρατηρούμενο πίνακα πληροφορίας (*observed information matrix*) $\mathbf{H}_N(\boldsymbol{\beta})$, διάστασης επίσης $p \times p$, με στοιχεία $\mathbf{H}_{N,ij}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j}$ ισχύει η σχέση

$$\mathbf{H}_N(\boldsymbol{\beta}) = -\nabla\nabla'\ell(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) - \mathbf{R}_N(\boldsymbol{\beta}), \tag{4.37}$$

όπου, με βάση την (4.35),

$$\mathbf{R}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N}\sum_{r=1}^{q=2} \mathbf{Z}_{t-1}\mathbf{W}_{tr}(\boldsymbol{\beta})\mathbf{Z}_{t-1}'(Y_{tr} - \pi_{tr}(\boldsymbol{\beta})) \tag{4.38}$$

με $\mathbf{W}_{tr}(\boldsymbol{\beta}) = \frac{\partial^2 u_r(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t \boldsymbol{\eta}_t'}$ για $r = 1, 2 = q$. Πιο αναλυτικά, για τον πίνακα $\mathbf{W}_{tr}(\boldsymbol{\beta})$ που στην προκειμένη περίπτωση είναι διάστασης 2×2, θα έχουμε

$$\mathbf{W}_{tr}(\boldsymbol{\beta}) = \frac{\partial^2 u_r(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t \partial \boldsymbol{\eta}_t'} = \frac{\partial}{\partial \boldsymbol{\eta}_t}\Big[\frac{\partial u_r(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t'}\Big] =$$

$$\begin{bmatrix} \frac{\partial}{\partial \eta_1} \\ \frac{\partial}{\partial \eta_2} \end{bmatrix} \begin{bmatrix} \frac{\partial u_r}{\partial \eta_1} & \frac{\partial u_r}{\partial \eta_2} \end{bmatrix} = \begin{bmatrix} \frac{\partial^2 u_r}{\partial \eta_1^2} & \frac{\partial^2 u_r}{\partial \eta_1 \partial \eta_2} \\ \frac{\partial^2 u_r}{\partial \eta_1 \partial \eta_2} & \frac{\partial^2 u_r}{\partial \eta_2^2} \end{bmatrix}.$$

Ο πίνακας $\mathbf{R}_N(\boldsymbol{\beta})$ στην σχέση (4.38) θυμίζει την μορφή του $\mathbf{R}_N(\boldsymbol{\beta})$ στην μονοδιάστατη περίπτωση που δίνεται από

$$\mathbf{R}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t(\phi)} \sum_{t=1}^{N} \mathbf{Z}_{t-1} d_t(\boldsymbol{\beta}) \mathbf{Z}_{t-1}'(Y_t - \mu_t(\boldsymbol{\beta}))$$

όπου $d_t(\boldsymbol{\beta}) = [\frac{\partial^2 u(\eta_t)}{\partial \eta_t^2}]$ με την συνάρτηση u να ορίζεται αναλυτικά σύμφωνα με τον τύπους (Α.7) και (Α.8) του Παραρτήματος Α2.

4.5 Συμπερασματολογία για $m > 3$

Τα προαναφερθέντα αποτελέσματα γενικεύονται εύκολα και στην περίπτωση $m > 3$ ακολουθώντας τα ίδια βήματα. Έστω λοιπόν $\{Y_t\}$ η κατηγορική χρονοσειρά με $m = q + 1$ επίπεδα. Για κάθε χρόνο $t = 1, 2, \ldots, N$ θεωρούμε το διάνυσμα $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tq})'$ όπου

$$Y_{tj} = \begin{cases} 1, & \text{αν η } j\text{-οστή κατηγορία παρατηρείται τον χρόνο } t \\ 0, & \text{αλλιώς.} \end{cases}$$

Ακόμη έστω $\boldsymbol{\pi}_t(\boldsymbol{\beta}) = (\pi_{t1}, \pi_{t2}, \ldots, \pi_{tq})'$ το διάνυσμα των δεσμευμένων πιθανοτήτων όπου $\pi_{tj} = P(Y_{tj} = 1 \mid \mathcal{F}_{t-1}), \quad j = 1, 2, \ldots, q$. Αναλόγως με την (4.18) η μερική πιθανοφάνεια θα είναι

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^{N} \prod_{j=1}^{m} \pi_{tj}(\boldsymbol{\beta})^{y_{tj}} \tag{4.39}$$

και επομένως ο λογάριθμός της θα δίνεται από την σχέση

$$\ell(\boldsymbol{\beta}) = \log PL(\boldsymbol{\beta}) = \sum_{t=1}^{N} \sum_{j=1}^{m} y_{tj} \log \pi_{tj}(\boldsymbol{\beta}). \tag{4.40}$$

Σε αναλογία με την μορφή της $\ell(\boldsymbol{\beta})$ στην περίπτωση που $m = 3$ (βλέπε (4.20)) η συνάρτηση του λογαρίθμου της μερικής πιθανοφάνειας τώρα θα λάβει την μορφή

$$\begin{aligned} \ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \{ & y_{t1} \log \Big(\frac{\pi_{t1}(\boldsymbol{\beta})}{1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})} \Big) + y_{t2} \log \Big(\frac{\pi_{t2}(\boldsymbol{\beta})}{1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})} \Big) + \ldots \\ & \ldots + y_{tq} \log \Big(\frac{\pi_{tq}(\boldsymbol{\beta})}{1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})} \Big) + \log(1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})) \}, \end{aligned} \tag{4.41}$$

με $q = m - 1$. Έτσι το φυσικό παραμετρικό διάνυσμα θα είναι

$$\begin{aligned} \boldsymbol{\theta}_t(\boldsymbol{\beta}) &= (\theta_{t1}(\boldsymbol{\beta}), \ldots, \theta_{tq}(\boldsymbol{\beta}))' \\ &= \Big(\log \Big(\frac{\pi_{t1}(\boldsymbol{\beta})}{1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})} \Big), \ldots, \log \Big(\frac{\pi_{tq}(\boldsymbol{\beta})}{1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})} \Big) \Big)' \end{aligned} \tag{4.42}$$

και τελικά η (4.41) θα πάρει την μορφή

$$\begin{aligned} \ell(\boldsymbol{\beta}) &= \{ \sum_{j=1}^{q} y_{tj} \theta_{tj}(\boldsymbol{\beta}) - \log[1 + \sum_{j=1}^{q} \exp(\theta_{tj}(\boldsymbol{\beta}))] \} \\ &= \sum_{t=1}^{N} \ell_t(\boldsymbol{\beta}). \end{aligned} \tag{4.43}$$

Παρατήρηση 4.5.1 Το διάνυσμα $\boldsymbol{\theta}_t(\boldsymbol{\beta})$ αποτελεί την τιμή της q-διάστατης συνάρτησης *logit* για $\boldsymbol{x} = \boldsymbol{\pi}_t(\boldsymbol{\beta})$. Αυτό εύκολα μπορούμε να το διαπιστώσουμε από τον ορισμό της συνάρτησης *logit* που είναι

$$\text{logit}(\boldsymbol{x}) = \Big(\log\Big(\frac{x_1}{1-\sum_{j=1}^{q} x_j}\Big), \ldots, \log\Big(\frac{x_1}{1-\sum_{j=1}^{q} x_j}\Big)\Big)' \tag{4.44}$$

με το $\boldsymbol{x}$ να ανήκει στο σύνολο $\{(x_1, x_2 \ldots, x_q)' : x_j > 0, j = 1, 2, \ldots, q, \sum_{j=1}^{q} x_j < 1\}$. Ακόμη αξίζει να αναφέρουμε ότι η συνάρτηση *logit* είναι η αντίστροφη συνάρτηση του κανονικού συνδέσμου στην πολυωνυμική κατανομή.

Υποθέτοντας διαφορισιμότητα, ο MPLE $\hat{\boldsymbol{\beta}}$ (εφόσον υπάρχει) βρίσκεται από την λύση των εξισώσεων μερικών σκόρ (*partial score equations*)

$$\nabla \ell(\boldsymbol{\beta}) = \nabla \log PL(\boldsymbol{\beta}) = \mathbf{0}$$

μέσω του αλγορίθμου *Fisher Scoring* (βλέπε Παράρτημα Β).
Για το διάνυσμα των μερικών σκόρ (*partial scor vector*), όπως και στην περίπτωση που $m = 3$, θα ισχύει

$$\mathbf{S}_N(\boldsymbol{\beta}) = \nabla \ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{D}_t(\boldsymbol{\beta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta})(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})). \tag{4.45}$$

Για τον $q \times q$ πίνακα $\mathbf{D}_t(\boldsymbol{\beta})$ στην προκειμένη περίπτωση θα ισχύει

$$\mathbf{D}_t(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{h}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t} = \begin{bmatrix} \frac{\partial \pi_{t1}}{\partial \eta_{t1}} & \frac{\partial \pi_{t2}}{\partial \eta_{t1}} & \cdots & \frac{\partial \pi_{tq}}{\partial \eta_{t1}} \\ \frac{\partial \pi_{t1}}{\partial \eta_{t2}} & \frac{\partial \pi_{t2}}{\partial \eta_{t2}} & \cdots & \frac{\partial \pi_{tq}}{\partial \eta_{t2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial \pi_{t1}}{\partial \eta_{tq}} & \frac{\partial \pi_{t2}}{\partial \eta_{tq}} & \cdots & \frac{\partial \pi_{tq}}{\partial \eta_{tq}} \end{bmatrix}.$$

Θέτουμε

$$\mathbf{U}_t(\boldsymbol{\beta}) = \mathbf{D}_t(\boldsymbol{\beta}) \boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}),$$

όπου $\boldsymbol{\Sigma}_t(\boldsymbol{\beta})$ είναι ο δεσμευμένος πίνακας διακυμάνσεων-συνδιακυμάνσεων του $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tq})'$ διάστασης $q \times q$ με στοιχεία

$$\sigma_t^{(ij)}(\boldsymbol{\beta}) = \begin{cases} -\pi_{ti}(\boldsymbol{\beta})\pi_{tj}(\boldsymbol{\beta}) & ,\text{αν } i \neq j \\ \pi_{ti}(\boldsymbol{\beta})(1 - \pi_{ti}(\boldsymbol{\beta})) & ,\text{αν } i = j \end{cases}$$

για $i, j = 1, 2, \ldots, q$. Επομένως το διάνυσμα των μερικών σκόρ μπορεί να εκφρασθεί και μέσω της σχέσης

$$\mathbf{S}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N} \mathbf{Z}_{t-1} \mathbf{U}_t(\boldsymbol{\beta})(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})), \tag{4.46}$$

όπου

$$\mathbf{U}_t(\boldsymbol{\beta}) = \frac{\partial \boldsymbol{u}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t}$$

είναι πλέον ένας $q \times q$ πίνακας με $\boldsymbol{\eta}_t = \mathbf{Z}'_{t-1}(\boldsymbol{\beta})$. Η q-διάστατη συνάρτηση $\boldsymbol{u} = (u_1, u_2, \ldots, u_q)'$ αποτελεί την σύνθεση της $\boldsymbol{h}$ (που είδαμε στην σχέση (4.11)) με την συνάρτηση logit (σχέση (4.44)). Δηλαδή θα έχουμε

$$\mathbf{u} = \Big(\log\Big(\frac{h_1(\eta_t)}{1 - \sum_{j=1}^{q} h_j(\eta_t)}\Big), \ldots, \log\Big(\frac{h_q(\eta_t)}{1 - \sum_{j=1}^{q} h_j(\eta_t)}\Big)\Big)'.$$

Για τον αθροιστικό κατα συνθήκη πίνακα πληροφορίας θα ισχύει

$$\begin{aligned} \mathbf{G}_N(\boldsymbol{\beta}) &= \sum_{t=1}^{N} Cov[\mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})(\mathbf{Y}_t - \boldsymbol{\pi}_t(\boldsymbol{\beta})) \mid \mathcal{F}_{t-1}] \\ &= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\mathbf{U}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t(\boldsymbol{\beta})\mathbf{U}'_t(\boldsymbol{\beta})\mathbf{Z}'_{t-1}. \end{aligned} \tag{4.47}$$

Ο αδέσμευτος πίνακας πληροφορίας δίνεται από την σχέση

$$\mathbf{F}_N(\boldsymbol{\beta}) = E[\mathbf{G}_N(\boldsymbol{\beta})].$$

Τέλος σε αναλογία με την (4.38), ο πίνακας $\mathbf{R}_N(\boldsymbol{\beta})$ δίνεται από την σχέση

$$\mathbf{R}_N(\boldsymbol{\beta}) = \sum_{t=1}^{N}\sum_{r=1}^{q} \mathbf{Z}_{t-1}\mathbf{W}_{tr}(\boldsymbol{\beta})\mathbf{Z}'_{t-1}(\mathbf{Y}_{tr} - \pi_{tr}(\boldsymbol{\beta})) \tag{4.48}$$

με

$$\mathbf{W}_{tr}(\boldsymbol{\beta})\frac{\partial^2 u_r(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t \partial \boldsymbol{\eta}'_t},$$

για $r = 1, 2, \ldots, q$.

4.6 Ασυμπτωτικά αποτελέσματα

Όπως και στην περίπτωση των δίτιμων χρονοσειρών έτσι και στις ποιοτικές χρονολογικές σειρές ο εκτιμητής μέγιστης μερικής πιθανοφάνειας ($MPLE$), εφόσον υπάρχει, έχει τις ακόλουθες ασυμπτωτικές ιδιότητες

(i) είναι συνεπής (*consistent*)

(ii) είναι ασυμπτωτικά κανονικός (*asymptotically normal*).

Ακόμη θα πρέπει να αναφέρουμε ότι η ύπαρξη του εκτιμητή $\hat{\boldsymbol{\beta}}$ εξασφαλίζεται μέσω κατάλληλων συνθηκών ομαλότητας και κανονικότητας (*Kaufmann* (1987)). Το Θεώρημα (Α.3.1) του Παραρτήματος Α3 ισχύει και ουσιαστικά υπαινίσσεται ότι αν ο λογάριθμος της κάθε συντεταγμένης της συνάρτησης σύνδεσης $\boldsymbol{h}$ είναι κοίλη συνάρτηση, τότε η πιθανότητα ότι υπάρχει μοναδικός εκτιμητής μέγιστης μερικής πιθανοφάνειας συγκλίνει στο 1 (*Pratt* (1981)). Οποιαδήποτε τέτοια ακολουθία εκτιμητών είναι συνεπής και ασυμπτωτικά κανονική.

Άμεση εφαρμογή του Θεωρήματος (Α.3.1) οδηγεί στην κατασκευή διαστήματος εμπιστοσύνης για το διάνυσμα των πιθανοτήτων μετάβασης $\boldsymbol{\pi}_t(\boldsymbol{\beta})$. Για την δημιουργία του δεδομένου διαστήματος στηριζόμαστε στο γεγονός ότι

$$\sqrt{N}\Big(\boldsymbol{\pi}_t(\hat{\boldsymbol{\beta}}) - \boldsymbol{\pi}_t(\boldsymbol{\beta})\Big) \xrightarrow{d} N_q(0, \mathbf{Z}_{t-1}\mathbf{D}_t(\boldsymbol{\beta})\mathbf{G}^{-1}(\boldsymbol{\beta})\mathbf{D}_t'(\boldsymbol{\beta})\mathbf{Z}_{t-1}') \tag{4.49}$$

καθώς $N \to \infty$, που προκύπτει με εφαρμογή της μεθόδου δέλτα (*Rao*, (1973), σελίδα 338).

4.7 Έλεγχος Υποθέσεων

Στα μοντέλα παλινδρόμησης κατηγορικών χρονοσειρών συχνά επιθυμούμε να ελέγξουμε υποθέσεις σχετικά με «κρίσιμες» τιμές που μπορεί να λάβουν κάποιες από τις παραμέτρους. Ο χαρακτηρισμός «κρίσιμες» έχει να κάνει με το γεγονός ότι η αποδοχή των δεδομένων τιμών για κάποιες από τις παραμέτρους παλινδρόμησης οδηγεί στην αναδιαμόρφωση του μοντέλου που θα χρησιμοποιήσουμε για στατιστική συμπερασματολογία, το οποίο πλέον αποκτά απλούστερη δομή. Ενδεικτικά αναφέρουμε το ακόλουθο παράδειγμα. Έστω η ποιοτική χρονοσειρά $\{Y_t\}, t = 1, 2, \ldots, N$, και $\mathbf{Z}_{t-1} = (Y_{t-1}, Y_{t-2}, X_t, W_t)'$ το διάνυσμα των συμμεταβλητών με X_t και Y_t να είναι συνεχείς. Υποθέτουμε ότι για την στατιστική ανάλυση της συγκεκριμένης σειράς το μοντέλο *logit*

$$\text{logit}(\pi_t(\boldsymbol{\beta})) = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 X_t + \beta_4 W_t$$

με $\boldsymbol{\beta} = (\beta_0, \beta_1, \beta_2, \beta_3, \beta_4)'$, είναι κατάλληλο. Στην προκειμένη περίπτωση η αποδοχή της μηδενικής υπόθεσης $\mathbf{H}_0 : \beta_2 = 0$ δηλώνει ότι η παρουσία της μεταβλητής Y_{t-2} δεν είναι απαραίτητη στο μοντέλο και επομένως πρέπει να απομακρυνθεί.

Στην δεδομένη ενότητα θα ασχοληθούμε με πιο σύνθετους ελέγχους υποθέσεων σχετικά με το παραμετρικό διάνυσμα $\boldsymbol{\beta}$ του γενικού μοντέλου παλινδρόμησης των κατηγορικών χρονοσειρών (4.11).

Έτσι σε προβλήματα ποιοτικών σειρών συχνά μας απασχολεί έλεγχος της γενικής γραμμικής υπόθεσης

$$\mathbf{H}_0 : \mathbf{C}\boldsymbol{\beta} = \boldsymbol{\beta}_0 \quad \text{κατά} \quad \mathbf{H}_1 : \mathbf{C}\boldsymbol{\beta} \neq \boldsymbol{\beta}_0, \tag{4.50}$$

όπου $\mathbf{C}$ είναι δεδομένος πίνακας πλήρους τάξης, έστω $r \leq p$. Τα πιο συχνά χρησιμοποιούμενα τέστ για τον έλεγχο της υπόθεσης (4.50) είναι,

- Ο λόγος μερικής πιθανοφάνειας

$$\lambda_N = 2\Big\{\ell(\hat{\boldsymbol{\beta}}) - \ell(\tilde{\boldsymbol{\beta}})\Big\}, \tag{4.51}$$

- Το στατιστικό του *Wald*

$$w_N = \Big\{\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\Big\}' \Big\{\mathbf{C}\mathbf{G}^{-1}(\hat{\boldsymbol{\beta}})\mathbf{C}'\Big\}^{-1} \Big\{\mathbf{C}\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}_0\Big\} \tag{4.52}$$

- Το στατιστικό του μερικού σκόρ

$$c_N = \frac{1}{N}\mathbf{S}'_N(\tilde{\boldsymbol{\beta}})\mathbf{G}^{-1}(\tilde{\boldsymbol{\beta}})\mathbf{S}_N(\tilde{\boldsymbol{\beta}}), \tag{4.53}$$

όπου $\tilde{\boldsymbol{\beta}}$ είναι ο εκτιμητής μέγιστης μερικής πιθανοφάνειας του $\boldsymbol{\beta}$ υπό την μηδενική υπόθεση, ενώ $\hat{\boldsymbol{\beta}}$ είναι ο γενικός εκτιμητής μέγιστης μερικής πιθανοφάνειας (υπό την $\mathbf{H}_0 \cup \mathbf{H}_1$).

Θεώρημα 4.7.1 *Κάτω από συγκεκριμένες συνθήκες ομαλότητας και κανονικότητας οι στατιστικοί έλεγχοι λ_N, w_N και c_N είναι ασυμπτωτικά ισοδύναμοι. Επιπλέον, υπό την μηδενική υπόθεση (4.50), η ασυμπτωτική κατανομή τους είναι χι-τετράγωνο με r βαθμούς ελευθερίας.*

Για περισσότερες λεπτομέρειες σχετικά με την ασυμπτωτική κατανομή των προαναφερθέντων στατιστικών ο ενδιαφερόμενος αναγνώστης παραπέμπεται στον *Fahrmeir* (1987). Η συμπεριφορά των προαναφερθέντων στατιστικών εξετάστηκε σε μια σειρά εναλλακτικών καταστάσεων από τους *L.Fahrmeir* και *H.Kaufmann* (1987). Οι συγκεκριμένοι συγγραφείς εξέτασαν με το προαναφερθέν θεώρημα την ομοιογένεια και την τάξη μιας Μαρκοβιανής Αλυσίδας, καθώς και την δομική αλλαγή σε συνδυασμό με την ανεξαρτησία δυο παράλληλων χρονοσειρών.

4.8 Έλεγχοι Καλής Προσαρμογής

Γενικά στην θεωρία των γραμμικών μοντέλων όσο και των γενικευμένων γραμμικών μοντέλων μετά την εκτίμηση του μοντέλου εμφανίζεται το πρόβλημα της καλής προσαρμογής του στα δεδομένα. Ενα μοντέλο θα λέγεται «καλό» εφόσον κατορθώνει να ερμηνεύσει τα δεδομένα με όσο το δυνατόν μεγαλύτερη ακρίβεια, ικανοποιώντας όμως και τις παραδοχές που μας οδήγησαν στην κατασκευή του. Στα γενικευμένα γραμμικά μοντέλα για ανεξάρτητες παρατηρήσεις με σκοπό να ελέγξουμε την καλή προσαρμογή των υποδειγμάτων χρησιμοποιούσαμε το στατιστικό X^2 του *Pearson* καθώς και την *scaled deviance* (*McCullagh and Nelder* (1989)). Στην περίπτωση των κατηγορικών χρονοσειρών εξακολουθούμε να χρησιμοποιούμε τα προαναφερθέντα στατιστικά αφού πρώτα τα τροποποιήσουμε ώστε να λαμβάνουν υπόψη την ιστορία του φαινομένου μέχρι τον χρόνο t(δηλαδή το $\mathcal{F}_{t-1}$). Έτσι (βλέπε *Fokianos and Kedem*, 2002, σελίδα 110) η *scaled deviance* παίρνει την μορφή

$$D = -2\sum_{t=1}^{N}\sum_{j=1}^{m} Y_{tj}\log\pi_{tj}(\hat{\boldsymbol{\beta}}), \tag{4.54}$$

ενώ το στατιστικό του *Pearson* γράφεται

$$\chi^2 = \sum_{t=1}^{N}\Big(\mathbf{Y}_t - \pi_t(\hat{\boldsymbol{\beta}}\Big)'\boldsymbol{\Sigma}_t^{-1}(\hat{\boldsymbol{\beta}})\Big(\mathbf{Y}_t - \pi_t(\hat{\boldsymbol{\beta}})\Big)$$

$$= \sum_{t=1}^{N}\sum_{j=1}^{m}\frac{\Big(Y_{tj} - \pi_{tj}(\hat{\boldsymbol{\beta}})\Big)^2}{\pi_{tj}(\hat{\boldsymbol{\beta}})}. \tag{4.55}$$

Αποδεικνύεται ότι υπό κατάλληλες συνθήκες ομαλότητας και κανονικότητας, η ασυμπτωτική κατανομή των (4.54) και (4.55) προσεγγίζει την χι-τετράγωνο κατανομή με $Nq-p$ βαθμούς ελευθερίας, όπου p είναι το πλήθος των παραμέτρων του μοντέλου. Η δεδομένη προσέγγιση είναι προβληματική καθώς απαιτεί πολύ μεγάλο N. Έτσι έχουν αναπτυχθεί εναλλακτικές μέθοδοι που βασίζονται στην ταξινόμηση της απόκρισης ή στην *Power Divergence*. Η δεδομένη μεθοδολογική προσέγγιση αρχικά εισήχθη για ανεξάρτητα δεδομένα ως γενίκευση του στατιστικού του γενικευμένου λόγου πιθανοφανειών και του ελέγχου του *Pearson* (*Cressie and Read* (1984)). Για περισσότερες λεπτομέρειες για την δεδομένη τεχνική βλέπε *Fokianos and Kedem*, (2002, σελίδα 112).

Για τον έλεγχο της επάρκειας ενός μοντέλου παλινδρόμησης για κατηγορικές χρονοσειρές δεν αρκούμαστε μόνο στα στατιστικά καλής προσαρμογής. Επιπλέον τα κρι-

τήρια πληροφορίας καθώς και η ανάλυση των υπολοίπων αποτελούν σημαντικά διαγνωστικά εργαλεία στην στατιστική ανάλυση των ποιοτικών σειρών.

Το σημαντικότερο κριτήριο επιλογής υποδείγματος είναι το AIC (Ακαικε Ινφορματιον Κριτεριον) που εισήχθη από τον *Akaike* (*Akaike* (1973)), (*Akaike* (1974)) και το οποίο δίνεται από την σχέση

$$AIC = D + 2p, \tag{4.56}$$

όπου D είναι η *scaled deviance* (4.54), ενώ το p δηλώνει το πλήθος των εκτιμημένων παραμέτρων του μοντέλου. Επειδή το AIC δεν παρέχει συνεπείς εκτιμητές του p καθώς το μήκος N της δειγματοληπτικής διαδρομής αυξάνει, προτάθηκαν διάφορες τροποποιήσεις του. Η σημαντικότερη από αυτές οφείλεται στον *Schwarz* ο οποίος πρότεινε το αποκαλούμενο Μπευζιανό κριτήριο πληροφορίας BIC (*Schwarz* (1978)) που ορίζεται από τον τύπο

$$BIC = D + p \log N. \tag{4.57}$$

Το BIC, στις περισσότερες περιπτώσεις, παρέχει συνεπείς εκτιμητές της τάξης του μοντέλου. Για περισσότερες πληροφορίες σχετικά με τα κριτήρια πληροφορίας σημαντική αναφορά αποτελεί ο (*Choi* (1992)).

Τέλος, η ανάλυση των υπολοίπων των κατηγορικών χρονοσειρών βασίζεται στα *raw* υπόλοιπα

$$\hat{\mathbf{e}}_t = \begin{pmatrix} \hat{e}_{t1} \\ \hat{e}_{t2} \\ \vdots \\ \hat{e}_{tq} \end{pmatrix} = \mathbf{Y}_t - \hat{\boldsymbol{\pi}}_t = \begin{pmatrix} Y_{t1} - \hat{\pi}_{t1} \\ Y_{t2} - \hat{\pi}_{t2} \\ \vdots \\ Y_{tq} - \hat{\pi}_{tq} \end{pmatrix}, \tag{4.58}$$

ή στα *squared Pearson* υπόλοιπα

$$\hat{r}_t = (\mathbf{Y}_t - \hat{\boldsymbol{\pi}}_t)' \hat{\boldsymbol{\Sigma}}_t^{-1} (\mathbf{Y}_t - \hat{\boldsymbol{\pi}}_t), \tag{4.59}$$

όπου $\hat{\boldsymbol{\Sigma}}_t = \boldsymbol{\Sigma}_t(\hat{\boldsymbol{\beta}})$ (*Pierce and Schafer* (1986)). Ένα μοντέλο θα είναι επαρκές εφόσον τα τετραγωνικά υπόλοιπα του *Pearson* συμπεριφέρονται ως λευκός θόρυβος (*white noise*) (*Li* (1991)).

Κεφάλαιο 5

Μοντέλα Παλινδρόμησης για Ονοματικές και Διατάξιμες Κατηγορικές Χρονοσειρές

5.1 Εισαγωγή

Όπως είδαμε στο Κεφάλαιο 4, για την μελέτη διαχρονικών φαινομένων των οποίων ο μηχανισμός τύχης δεν εμπίπτει στα πλαίσια του τυχαίου πειράματος και η μεταβλητή ενδιαφέροντος Y_t είναι πολυωνυμική με m επίπεδα, στηριζόμαστε στο γενικό μοντέλο της (4.11).

Από το μοντέλο (4.11) και επιλέγοντας, ανάλογα με την φύση του προβλήματος, κατάλληλες $q = (m-1)$-διάστατες συναρτήσεις σύνδεσης οδηγούμαστε σε μια σειρά μοντέλων που περιγράφουν τις κατηγορικές χρονολογικές σειρές. Αναλυτικότερα, συχνά συναντάμε φαινόμενα στα οποία η κατηγορική αποκριτική μεταβλητή είναι διατάξιμη (*ordinal*), γεγονός το οποίο θα μας οδηγήσει στην χρήση υποδειγμάτων τα οποία θα αξιοποιούν την δεδομένη πληροφορία και τα οποία θα διαφοροποιούνται από τα μοντέλα που θα αναφέρονται σε ονοματικές (*nominal*) μεταβλητές. Γίνεται λοιπόν σαφές πως η επιλογή μοντέλου εξαρτάται από τις πιθανές κλίμακες μέτρησης της κατηγορικής χρονοσειράς μας, που είναι η ονοματική (*nominal*), η διατακτική (*ordinal*) και η διαστηματική (*interval*).

Επειδή οι διαστηματικές (*interval*) μεταβλητές μπορούν να προσεγγιστούν από μοντέλα που αναφέρονται σε διατεταγμένα δεδομένα (*ordinal data*), στην ανάλυση που θα ακολουθήσει θα παρουσιάσουμε μοντέλα παλινδρόμησης για ονοματικές και διατακτικές χρονολογικές σειρές. Συνάμα αξίζει να αναφέρουμε ότι τα μοντέλα που θα αναφέρουμε γενικεύουν την λογιστική παλινδρόμηση (*logistic regression*) για πολυωνυμικές αποκριτικές μεταβλητές σε φαινόμενα που επιδεικνύουν διαχρονική εξάρτηση.

Στην Παράγραφο 5.2.1 θα αναφερθούμε στο *Baseline-Category Logit* μοντέλο για ονοματικές κατηγορικές σειρές. Στην Ενότητα 5.2.2 θα μιλήσουμε για την διαδικασία εύρεσης του $MPLE$ του δεδομένου υποδείγματος. Στην Παράγραφο 5.3.1 θα παρουσιάσουμε το *Proportional odds* μοντέλο, για διατάξιμες ποιοτικές χρονοσειρές, δίδοντας κάποιες ιδιότητές του στην Ενότητα 5.3.2. Στην Παράγραφο 5.3.3 θα δώσουμε την συνάρτηση της PL για αυτό. Μια εναλλακτική προσέγγιση του *Proportional odds* μοντέλου μέσω μιας βοηθητικής μεταβλητής δίνεται στην Παράγραφο 5.3.4. Ολοκληρώνοντας την ανάλυση μας, στην Ενότητα 4 παρέχονται εναλλακτικές μέθοδοι μοντελοποίησης διατάξιμων εξαρτημένων δεδομένων. Αξίζει να τονίσουμε ότι για τα μοντέλα *Baseline-Category Logit* και *Proportional odds* θα δείξουμε ότι αποτελούν ειδική περίπτωση του γενικού μοντέλου (4.11).

5.2 Ονοματικές Χρονολογικές Σειρές

Έστω Y_t η κατηγορική μεταβλητή με m κατηγορίες στις οποίες δεν εμπεριέχεται η έννοια της διάταξης. Στην δεδομένη παράγραφο θα παρουσιάσουμε τα *multicategory* ή *polytomous logit* μοντέλα για ονοματικές αποκρίσεις, τα οποία μοντελοποιούν ταυτόχρονα τους λογαρίθμους των *odds* για όλα τα δυνατά ζεύγη των m κατηγοριών. Το πλήθος αυτών των ζευγών είναι $\frac{m!}{2!(m-2)!} = \frac{m(m-1)}{2}$. Μέσω κατάλληλων περιορισμών και με σκοπό την μείωση της διάστασης του προβλήματος, τα δεδομένα μοντέλα αρκούνται ουσιαστικά στην περιγραφή $q = m - 1$ κατηγοριών της μεταβλητής Y_t.

5.2.1 *Baseline-Category Logit Models*

Έστω η κατηγορική ονοματική χρονοσειρά $Y_t, \quad t = 1, 2, \ldots, N$. Την χρονική στιγμή t, όπως είδαμε και στο Κεφάλαιο 4, στην αποκριτική μεταβλητή Y_t αντιστοιχεί το διάνυσμα $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tm})'$ το οποίο παριστάνει την πολυωνυμική δοκιμή στον δεδομένο χρόνο. Συγκεκριμένα, ισχύει

$$Y_{tj} = \begin{cases} 1, & \text{αν η } j\text{-οστή κατηγορία παρατηρείται τον χρόνο } t \\ & \quad \text{για} \qquad j = 1, 2 \ldots, q, \quad t = 1, 2, \ldots N \\ 0, & \text{αλλίως.} \end{cases} \quad . \tag{5.1}$$

Προφανώς $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tm})' \mid \mathcal{F}_{t-1} \sim M_m(1; \pi_{t1}, \pi_{t2}, \ldots, \pi_{tm})$ όπου $\pi_{tj} = Pr(Y_{tj} = 1 \mid \mathcal{F}_{t-1}) = Pr(Y_t = j \mid \mathcal{F}_{t-1}),\ j = 1, 2, \ldots, m$ με $\sum_{j=1}^{m} \pi_{tj} = 1$ και $\sum_{j=1}^{m} Y_{tj} = 1$. Έχοντας ορίσει $Y_{tm} = 1 - \sum_{j=1}^{q} Y_{tj}$ και $\pi_{tm} = 1 - \sum_{j=1}^{q} \pi_{tj}$ το μοντέλο

που θα παρουσιάσουμε για την στατιστική ανάλυση του διαχρονικού φαινομένου, θα περιγράφει πλέον το διάνυσμα των δεσμευμένων πιθανοτήτων $\boldsymbol{\pi}_t = (\pi_{t1}, \pi_{t2}, \ldots, \pi_{tq})'$, που τις ονομάσαμε πιθανότητες μετάβασης, στηριζόμενοι πλέον στα διανύσματα $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tq})'$ όπου $q = m - 1$. Έτσι για την π_{tj} επιπλέον μπορούμε να έχουμε

$$\pi_{tj}(\boldsymbol{\beta}) = Pr(Y_{tj} = 1 \mid \mathbf{Z}_{t-1}) = Pr(Y_t = j \mid \mathbf{Z}_{t-1}). \tag{5.2}$$

Επιλέγοντας ως κατηγορία αναφοράς (*baseline category*) την τελευταία, δηλαδή την $j = m$, το πολυωνυμικό logit ορίζεται ως εξής

$$\log \frac{\pi_{tj}(\boldsymbol{\beta})}{\pi_{tm}(\boldsymbol{\beta})} = \alpha_j + \widetilde{\boldsymbol{\beta}}_j' \cdot \widetilde{\mathbf{Z}}_{t-1}, \quad j = 1, 2, \ldots, q = m - 1 \tag{5.3}$$

και αποτελεί επέκταση του λογιστικού μοντέλου (3.8), με $\boldsymbol{\beta}$ να είναι το διάνυσμα των παραμέτρων και των q υποδειγμάτων. Το μοντέλο (5.3) περιγράφει ταυτόχρονα τις επιδράσεις του διανύσματος των επεξηγηματικών μεταβλητών $\widetilde{\mathbf{Z}}_{t-1} = (Z_{(t-1)2}, Z_{(t-1)3}, \ldots, Z_{(t-1)p})'$ στα $m - 1$ logits τα οποία έχουν την δική τους σταθερά α_j και το δικό τους διάνυσμα παραμέτρων $\widetilde{\boldsymbol{\beta}}_j = (\beta_{j2}, \beta_{j3}, \ldots, \beta_{jp})'$, σε σχέση με το $\widetilde{\mathbf{Z}}_{t-1}$.

Θα πρέπει να αναφέρουμε ότι ο λόγος $\frac{\pi_{tj}}{\pi_{tm}}$ αποτελεί την σχετική πιθανότητα (*odds*) εμφάνισης της j κατηγορίας ως προς την πιθανότητα εμφάνισης της m κατηγορίας, που είναι η κατηγορία αναφοράς. Εκτιμώντας για κάθε logit τις παραμέτρους α_j και $\widetilde{\boldsymbol{\beta}}_j$, $j = 1, 2, \ldots, m - 1$, εκτιμούμε και τον παραπάνω λόγο με αποτέλεσμα να είμαστε σε θέση την χρονική στιγμή t να αποφανθούμε, με κάποιο βαθμό βεβαιότητας, κατά πόσο θα πραγματοποιηθεί το j-οστό ή το m-οστό ενδεχόμενο. Επειδή προφανώς

$$\log \frac{\pi_{tc}(\boldsymbol{\beta})}{\pi_{tb}(\boldsymbol{\beta})} = \log \frac{\pi_{tc}(\boldsymbol{\beta})}{\pi_{tm}(\boldsymbol{\beta})} - \log \frac{\pi_{tb}(\boldsymbol{\beta})}{\pi_{tm}(\boldsymbol{\beta})} \tag{5.4}$$

όπου $c, b \in \{1, 2, \ldots, m-1\}$, βλέπουμε ότι μέσω δυο *baseline-category logits* μπορούμε να υπολογίσουμε την σχετική πιθανότητα εμφάνισης του ενδεχομένου c ως προς το ενδεχόμενο b. Η σχέση (5.4) μέσω της (5.3) γράφεται

$$\log \frac{\pi_{tc}(\widetilde{\mathbf{Z}}_{t-1})}{\pi_{tb}(\widetilde{\mathbf{Z}}_{t-1})} = (\alpha_c - \alpha_b) + (\widetilde{\boldsymbol{\beta}}_c' - \widetilde{\boldsymbol{\beta}}_b')\widetilde{\mathbf{Z}}_{t-1}. \tag{5.5}$$

Παρατηρούμε λοιπόν ότι ο λόγος $\frac{\pi_{tc}}{\pi_{tb}}$ είναι ο ίδιος άσχετα με τον συνολικό αριθμό m των κατηγοριών της Y_t. Αυτή η ιδιότητα καλείται «ανεξαρτησία άσχετων εναλλακτικών» (*independence of irrelevant alternatives*) (*Luce* (1959)).

Το πολυωνυμικό μοντέλο logit για κατηγορικές χρονολογικές σειρές σε όρους των πολυωνυμικών πιθανοτήτων π_{tj} (*Agresti* (2002)) γράφεται

$$\pi_{tj}(\boldsymbol{\beta}) = \frac{\exp(\alpha_j + \widetilde{\boldsymbol{\beta}}_j' \widetilde{\mathbf{Z}}_{t-1})}{1 + \sum_{\ell=1}^{q} \exp(\alpha_\ell + \widetilde{\boldsymbol{\beta}}_\ell' \widetilde{\mathbf{Z}}_{t-1})} \tag{5.6}$$

για $j = 1, 2, \ldots, q = m - 1$, με τους περιορισμούς $\alpha_m = 0$ και $\widetilde{\boldsymbol{\beta}}_m = \mathbf{0}$. Πράγματι, από την (5.3) έχουμε

$$\frac{\pi_{tj}}{\pi_{tm}} = \exp(\alpha_j + \widetilde{\boldsymbol{\beta}}_j' \widetilde{\mathbf{Z}}_{t-1}) \Rightarrow \pi_{tj} = \pi_{tm} \exp(\alpha_j + \widetilde{\boldsymbol{\beta}}_j' \widetilde{\mathbf{Z}}_{t-1}) \tag{5.7}$$

$$\Rightarrow \sum_{j=1}^{m} \frac{\pi_{tj}}{\pi_{tm}} = \sum_{j=1}^{m} \exp(\alpha_j + \widetilde{\boldsymbol{\beta}}_j' \widetilde{\mathbf{Z}}_{t-1}) \Rightarrow \frac{1}{\pi_{tm}} = \sum_{j=1}^{m} \exp(\alpha_j + \widetilde{\boldsymbol{\beta}}_j' \widetilde{\mathbf{Z}}_{t-1})$$

$$\Rightarrow \pi_{tm} = \frac{1}{\sum_{\ell=1}^{q} \exp(\alpha_\ell + \widetilde{\boldsymbol{\beta}}_\ell' \widetilde{\mathbf{Z}}_{t-1}) + \exp(\alpha_m + \widetilde{\boldsymbol{\beta}}_m' \widetilde{\mathbf{Z}}_{t-1})}.$$

Σύμφωνα όμως με τους περιορισμούς $\alpha_m = 0$ και $\widetilde{\boldsymbol{\beta}}_m = \mathbf{0}$ για την π_{tm} θα ισχύει

$$\pi_{tm} = \frac{1}{1 + \sum_{\ell=1}^{q} \exp(\alpha_\ell + \widetilde{\boldsymbol{\beta}}_\ell' \widetilde{\mathbf{Z}}_{t-1})}. \tag{5.8}$$

Επομένως, αντικαθιστώντας στην (5.7) προκύπτει το ζητούμενο. Το μοντέλο (5.6) εναλλακτικά προκύπτει από την μεγιστοποίηση μιας τυχαίας ποσότητας. Ο αναγνώστης για την δεδομένη προσέγγιση παραπέμπεται στον *McFadden* (1973).

Το μοντέλο (5.6) αποτελεί ειδική περίπτωση του γενικού πολυμεταβλητού μοντέλου παλινδρόμησης για την ανάλυση κατηγορικών χρονοσειρών (4.11). Αρκεί απλά να ορίσει κανείς το διάνυσμα $\boldsymbol{\beta}$ του γενικού μοντέλου (4.11) ως $(\alpha_1, \widetilde{\boldsymbol{\beta}}_1', \alpha_2, \widetilde{\boldsymbol{\beta}}_2', \ldots, \alpha_q, \widetilde{\boldsymbol{\beta}}_q')$ και να θέσει $\mathbf{Z}_{t-1} = (1, \widetilde{\mathbf{Z}}_{t-1})'_{p \times 1}$. Αναλυτικότερα, το $\boldsymbol{\beta}$ θα έχει την μορφή

$$\boldsymbol{\beta} = (\alpha_1, \beta_{12}, \beta_{13}, \ldots, \beta_{1p}, \alpha_2, \beta_{22}, \beta_{23}, \ldots, \beta_{2p}, \ldots, \alpha_q, \beta_{q2}, \beta_{q3}, \ldots, \beta_{qp})'$$

και προφανώς θα έχει διάσταση $q \cdot p$. Ακόμη ορίζουμε τον $q \cdot p \times q$ πίνακα (*Fokianos-Kedem*) ο οποίος έχει την μορφή

$$\mathbf{Z}_{t-1} = \begin{bmatrix} \boldsymbol{z}_{t-1} & \mathbf{0} & \ldots & \mathbf{0} \\ \mathbf{0} & \boldsymbol{z}_{t-1} & \ldots & \mathbf{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{z}_{t-1} \end{bmatrix},$$

καταλήγοντας έτσι στην (4.11).

5.2.2 Μερική Πιθανοφάνεια στο *Baseline-Category Logit* μοντέλο

Στο σημείο αυτό θα εξετάσουμε διεξοδικά την διαδικασία συμπερασματολογίας, μέσω της μερικής πιθανοφάνειας για το μοντέλο (5.6). Για την απλούστευση των αποτελεσμάτων θα τροποποιήσουμε τους συμβολισμούς μας. Έστω λοιπόν η δειγματοληπτική

διαδρομή $\{Y_t\}, t = 1, 2, \ldots, N$, ενός μη τυχαίου πειράματος με m δυνατά αποτελέσματα σε κάθε επανάληψη. Για τον χρόνο t το διάνυσμα $\mathbf{Y}_t = (Y_{t1}, Y_{t2}, \ldots, Y_{tm})'$ όπως έχουμε ήδη αναφέρει παριστά την πολυωνυμική δοκιμή στην εν λόγω χρονική στιγμή. Έστω $\mathbf{Z}_{t-1} = (Z_{(t-1)1}, Z_{(t-1)2}, \ldots, Z_{(t-1)p})'_{p\times 1}$ το διάνυσμα σταθερής μορφής των τυχαίων χρονοεξαρτώμενων συμμεταβλητών που αναφέρεται στα παραμετρικά διανύσματα $\boldsymbol{\beta}_j = (\beta_{j1}, \beta_{j2}, \ldots, \beta_{jp})'$, $j = 1, 2, \ldots, q = m-1$, των $m-1$ logit που υπάρχουν για το εν λόγω φαινόμενο. Τα διανύσματα $\boldsymbol{\beta}_j$ δεν περιλαμβάνουν τις σταθερές $\alpha_j, j = 1, 2, \ldots, q$. Έχοντας ορίσει

$$\pi_{tm} = 1 - \sum_{j=1}^{q} \pi_{tj} \quad \text{και} \quad Y_{tm} = 1 - \sum_{j=1}^{q} Y_{tj} \quad \text{με} \quad q = m-1,$$

η συνεισφορά του χρόνου t κατα την οποία λαμβάνει χώρα το φαινόμενο στην συνάρτηση του λογαρίθμου της μερικής πιθανοφάνειας θα είναι

$$\log f(\boldsymbol{y}_t \mid \mathcal{F}_{t-1}) = \log\Big\{ \underbrace{\frac{1!}{y_{t_1}! y_{t_2}! \cdots y_{t_m}!}}_{1} \prod_{j=1}^{m} (\pi_{tj}(\boldsymbol{\beta}))^{y_{tj}} \Big\} =$$

$$= \log\Big\{ \prod_{j=1}^{m} \pi_{tj}(\boldsymbol{\beta})^{y_{tj}} \Big\} = \sum_{j=1}^{q} y_{tj} \log \pi_{tj}(\boldsymbol{\beta}) + y_{tm} \log \pi_{tm}(\boldsymbol{\beta}) =$$

$$\sum_{j=1}^{m} y_{tj} \log \pi_{tj}(\boldsymbol{\beta}) + (1 - \sum_{j=1}^{q} y_{tj}) \log(1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})) =$$

$$= \sum_{j=1}^{q} y_{tj} \log \Big(\frac{\pi_{tj}(\boldsymbol{\beta})}{1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})} \Big) + \log[1 - \sum_{j=1}^{q} \pi_{tj}(\boldsymbol{\beta})],$$

όπου το διάνυσμα $\boldsymbol{\beta}$ θα περιέχει όλες τις παραμέτρους των $m-1$ logits. Έστω ότι έχουμε καταγράψει την εξέλιξη του φαινομένου για N φορές οι οποίες εξαρτώνται μεταξύ τους. Τότε ο λογάριθμος της μερικής πιθανοφάνειας υπολογίζεται ως εξής

$$\ell(\boldsymbol{\beta}) = \log \prod_{t=1}^{N} f(\boldsymbol{y}_t; \boldsymbol{\beta} \mid \mathcal{F}_{t-1}) = \sum_{t=1}^{N} \log f(\boldsymbol{y}_t; \boldsymbol{\beta} \mid \mathcal{F}_{t-1}) \Rightarrow$$

$$\ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \Big\{ \sum_{j=1}^{q} y_{tj} \log \frac{\pi_{tj}}{\pi_{tm}} + \log \pi_{tm} \Big\} \overset{(5.3)}{\Rightarrow}$$

$$\ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \Big\{ \sum_{j=1}^{q} y_{tj}(\alpha_j + \boldsymbol{\beta}_j' \mathbf{Z}_{t-1}) - \log[1 + \sum_{j=1}^{q} \exp(\alpha_j + \boldsymbol{\beta}_j' \mathbf{Z}_{t-1})] \Big\}. \tag{5.9}$$

Όμως,

$$\sum_{t=1}^{N} \Big\{ \sum_{j=1}^{q} y_{tj}(\alpha_j + \boldsymbol{\beta}_j' \mathbf{Z}_{t-1}) \Big\} = \sum_{t=1}^{N} \Big(\sum_{j=1}^{q} y_{tj} \{ \alpha_j + \sum_{i=1}^{p} \beta_{ji} \cdot Z_{(t-1)i} \} \Big) =$$

$$\sum_{t=1}^{N}\Big\{\sum_{j=1}^{q} y_{tj}\alpha_j + \sum_{j=1}^{q} y_{tj}\cdot\sum_{i=1}^{p}\beta_{ji}\cdot Z_{(t-1)i}\Big\} =$$

$$\sum_{j=1}^{q}\Big\{\sum_{t=1}^{N} y_{tj}\alpha_j + \sum_{t=1}^{N} y_{tj}\cdot\sum_{i=1}^{p}\beta_{ji}\cdot Z_{(t-1)i}\Big\} =$$

$$\sum_{j=1}^{q}\Big\{\alpha_j\Big(\sum_{t=1}^{N} y_{tj}\Big) + \sum_{i=1}^{p}\beta_{ji}\Big(\sum_{t=1}^{N} y_{tj}Z_{(t-1)i}\Big)\Big\}. \tag{5.10}$$

Άρα η (5.9) μέσω της (5.10) γράφεται

$$\ell(\boldsymbol{\beta}) = \sum_{j=1}^{q}\Big\{\alpha_j\Big(\sum_{t=1}^{N} y_{tj}\Big) + \sum_{i=1}^{p}\beta_{ji}\Big(\sum_{t=1}^{N} Z_{(t-1)i}y_{tj}\Big)\Big\}$$

$$-\sum_{t=1}^{N}\log\Big\{1+\sum_{j=1}^{q}\exp(\alpha_j+\boldsymbol{\beta}_j'\mathbf{Z}_{t-1})\Big\}. \tag{5.11}$$

Σύμφωνα με την (5.11) παρατηρούμε ότι το επαρκές στατιστικό για την παράμετρο β_{ji} είναι το $\sum_{t=1}^{N} Z_{(t-1)i}y_{tj}$ για $j=1,2,\ldots,q=m-1$ και $i=1,2,\ldots,p$ (*Birch*, (1964α)). Συνάμα, το επαρκές στατιστικό για το α_j είναι το $\sum_{t=1}^{N} y_{tj} = \sum_{t=1}^{N} Z_{(t-1)0}y_{tj}$ με $Z_{(t-1)0}=1$. Ουσιαστικά το $\sum_{t=1}^{N} y_{tj}$ αποτελεί τον συνολικό αριθμό αποτελεσμάτων της j κατηγορίας για τις N παρατηρήσεις της χρονοσειράς μας.

5.3 Διατάξιμες Χρονολογικές Σειρές

Έστω η κατηγορική χρονοσειρά $\{Y_t\}$, $t=1,2,\ldots,N$ στην οποία η μεταβλητή ενδιαφέροντος Y_t μετριέται σε κλίμακα που εμπεριέχεται η φυσική διάταξη. Για την μελέτη του αντίστοιχου διαχρονικού στοχαστικού φαινομένου το οποίο «γέννησε» τα δεδομένα, δηλαδή την δειγματοληπτική διαδρομή, καλούμαστε να χρησιμοποιήσουμε διαφορετικά στατιστικά μοντέλα από εκείνα που συναντήσαμε στις ονοματικές χρονοσειρές. Αναλυτικότερα για την στατιστική ανάλυση των διατάξιμων κατηγορικών χρονοσειρών θα αξιοποιήσουμε υποδείγματα τα οποία αντανακλούν διατάξιμα χαρακτηριστικά, όπως είναι η μονότονη τάση (*monotone trend*), αφού αυτά έχουν αυξημένη ισχύ, παρέχοντας μοντέλα με μικρό αριθμό συμμεταβλητών (*Brillinger* (1996)).

5.3.1 *Proportional odds model*

Για την μελέτη του δεδομένου μοντέλου απαιτείται πρώτα να εξηγήσουμε την έννοια του *Cumulative Logit* μοντέλου (*McCullagh* (1980)), (*Snell* (1964)).

Ένας τρόπος για να αξιοποιήσουμε την κατηγορική διάταξη της μεταβλητής ενδιαφέροντος είναι το μοντέλο logit των *αθροιστικών πιθανοτήτων* (*cumulative probabilities*) οι οποίες έχουν την μορφή

$$P(Y_t \leq j \mid \mathbf{Z}_{t-1}) = \pi_{t1}(\mathbf{Z}_{t-1}) + \ldots + \pi_{tj}(\mathbf{Z}_{t-1}), \quad j = 1, 2, \ldots, m.$$

Επειδή όμως το διάνυσμα $\mathbf{Z}_{t-1}$ των τυχαίων χρονοεξαρτώμενων συμμεταβλητών εμπεριέχει την ιστορία του φαινομένου $\mathcal{F}_{t-1}$, για τις αθροιστικές πιθανότητες θα έχουμε ισοδύναμα

$$P(Y_t \leq j \mid \mathcal{F}_{t-1}) = \pi_{t1}(\boldsymbol{\beta}) + \pi_{t2}(\boldsymbol{\beta}) + \ldots + \pi_{tj}(\boldsymbol{\beta}), \quad j = 1, 2, \ldots, m.$$

Για το εν λόγω διαχρονικό φαινόμενο που υπάρχουν m δυνατά αποτελέσματα σε κάθε επανάληψη θα αντιστοιχούν $m-1$ *cumulative logits* μοντέλα που ορίζονται ως εξής

$$\text{logit}[P(Y_t \leq j \mid \mathcal{F}_{t-1})] = \log \frac{P(Y_t \leq j \mid \mathcal{F}_{t-1})}{1 - P(Y_t \leq j \mid \mathcal{F}_{t-1})} =$$

$$\log \frac{P(Y_t \leq j \mid \mathcal{F}_{t-1})}{P(Y_t > j \mid \mathcal{F}_{t-1})} = \log \frac{\pi_{t1} + \pi_{t2} + \ldots + \pi_{tj}}{\pi_{t(j+1)} + \pi_{t(j+2)} + \ldots + \pi_{tm}} \quad j = 1, 2, \ldots, m-1. \tag{5.12}$$

Σύμφωνα με την (5.12) γίνεται φανερό ότι κάθε *cumulative logit* χρησιμοποιεί και τις m κατηγορίες απόκρισης.
Το μοντέλο το οποίο αξιοποιεί ταυτόχρονα τα $q = m-1$ *cumulative logits* της σχέσης (5.12) είναι το

$$\text{logit}[P(Y_t \leq j \mid F_{t-1})] = \log \frac{P(Y_t \leq j \mid F_{t-1})}{P(Y_t > j \mid F_{t-1})} = \theta_j + \boldsymbol{\gamma}' \mathbf{Z}_{t-1}, \quad j = 1, 2, \ldots, q \tag{5.13}$$

όπου $\boldsymbol{\gamma} = (\gamma_1, \gamma_2, \ldots, \gamma_p)'$. Από την (5.13) βλέπουμε ότι κάθε *cumulative logit* έχει την δική του σταθερά θ_j, ενώ οι επιδράσεις του $\mathbf{Z}_{t-1}$ είναι ίδιες για τα $m-1$ logits και εκφράζονται μέσω του σταθερού διανύσματος $\boldsymbol{\gamma}$ κατα γραμμικό τρόπο. Την χρονική στιγμή t το διάνυσμα $\mathbf{Z}_{t-1}$ παίρνει μια σταθερή τιμή που είναι κοινή και για τα $m-1$ logits της (5.13). Επομένως ο όρος $\boldsymbol{\gamma}' \mathbf{Z}_{t-1}$τον δεδομένο χρόνο είναι σταθερός. Για να έχει νόημα το μοντέλο (13) απαιτείται τον χρόνο t τα θ_j, $j = 1, 2, \ldots, m-1$ να αυξάνουν ως προς j. Αυτό συμβαίνει διότι το $1^ο$ μέλος της (5.13), την δεδομένη χρονική στιγμή, είναι αύξουσα συνάρτηση του j. Πράγματι η $P(Y_t \leq j \mid F_{t-1}) \equiv p(j)$ αυξάνει ως προς j και επομένως η συνάρτηση $\log \frac{p(j)}{1-p(j)}$ είναι αύξουσα συνάρτηση του j. Έτσι, ορίζοντας $\theta_0 = -\infty$ και $\theta_m = +\infty$ για τα $\theta_j, j = 1, 2, \ldots, m$, αναγκαστικά θα ισχύει

$$-\infty = \theta_0 < \theta_1 < \ldots < \theta_{m-1} < \theta_m = +\infty.$$

5.3.2 Ιδιότητες του μοντέλου (5.13)

Για τον χρόνο t θεωρούμε δυο διαφορετικές τιμές του συγκεκριμένου χρονοεξαρτώμενου διανύσματος των συμμεταβλητών $\mathbf{Z}_{t-1} = (Z_{(t-1)1}, Z_{(t-1)2}, \ldots, Z_{(t-1)p})'$, που αντιστοίχως είναι

$$\boldsymbol{z}_{t-1,1} = (z_{(t-1)11}, z_{(t-1)12}, \ldots, z_{(t-1)1p})',$$

$$\boldsymbol{z}_{t-1,2} = (z_{(t-1)21}, z_{(t-1)22}, \ldots, z_{(t-1)2p})'.$$

Για τις προαναφερθείσες πραγματοποιήσεις του $\mathbf{Z}_{t-1}$, οι οποίες αντικατοπρίζουν ουσιαστικά με δυο διαφορετικούς τρόπους την ιστορία $\mathcal{F}_{t-1}$, ισχύει

$$\text{logit}[P(Y_t \leq j \mid \boldsymbol{z}_{t-1,1})] - \text{logit}[P(Y_t \leq j \mid \boldsymbol{z}_{t-1,2})] =$$

$$= \log \frac{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,1})/P(Y_t > j \mid \boldsymbol{z}_{t-1,1})}{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,2})/P(Y_t > j \mid \boldsymbol{z}_{t-1,2})} = \theta_j + \boldsymbol{\gamma}' \cdot \boldsymbol{z}_{t-1,1} - (\theta_j + \boldsymbol{\gamma}' \cdot \boldsymbol{z}_{t-1,2})$$

$$\Rightarrow \log \frac{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,1})/P(Y_t > j \mid \boldsymbol{z}_{t-1,1})}{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,2})/P(Y_t > j \mid \boldsymbol{z}_{t-1,2})} = \boldsymbol{\gamma}'(\boldsymbol{z}_{t-1,1} - \boldsymbol{z}_{t-1,2}) \tag{5.14}$$

Στην περίπτωση μας ο λόγος των σχετικών πιθανοτήτων (*odds ratio*)

$$or = \frac{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,1})/P(Y_t > j \mid \boldsymbol{z}_{t-1,1})}{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,2})/P(Y_t > j \mid \boldsymbol{z}_{t-1,2})}$$

επειδή αναφέρεται σε αθροιστικές πιθανότητες ονομάζεται λόγος σχετικών αθροιστικών πιθανοτήτων (*cumulative odds ratio*).

Σύμφωνα με την (5.14) ο λογάριθμος του δεδομένου *odds ratio* είναι ανάλογος της απόστασης μεταξύ των τιμών $\boldsymbol{z}_{t-1,1}$ και $\boldsymbol{z}_{t-1,2}$. Η (5.14) γράφεται ισοδύναμα

$$or = \exp\{\boldsymbol{\gamma}'[\boldsymbol{z}_{t-1,1} - \boldsymbol{z}_{t-1,2}]\} \Rightarrow$$

$$\frac{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,1})}{1 - P(Y_t \leq j \mid \boldsymbol{z}_{t-1,1})} = \underbrace{\exp\{\boldsymbol{\gamma}'[\boldsymbol{z}_{t-1,1} - \boldsymbol{z}_{t-1,2}]\}}_{>0} \cdot \Big(\frac{P(Y_t \leq j \mid \boldsymbol{z}_{t-1,2})}{1 - P(Y_t \leq j \mid \boldsymbol{z}_{t-1,2})}\Big),$$

από όπου συμπεραίνουμε ότι την χρονική στιγμή t η σχετική πιθανότητα η τιμή της αποκριτικής μεταβλητής να είναι $\leq j$ για την τιμή $\boldsymbol{z}_{t-1,1}$ του $\mathbf{Z}_{t-1}$ είναι $\exp\{\boldsymbol{\gamma}_j'[\boldsymbol{z}_{t-1,1} - \boldsymbol{z}_{t-1,2}]\}$ φορές μεγαλύτερη από την σχετική πιθανότητα του ενδεχομένου $\{Y_t \leq j\}$ για $\mathbf{Z}_{t-1} = \boldsymbol{z}_{t-1,2}$. Λόγω των παραπάνω ιδιοτήτων το μοντέλο (13) πέρα από *cumulative Logit* ονομάστηκε από τον *McCullagh* (1980) *proportional odds model.*

Παρατήρηση 5.3.1 Θεωρούμε ότι $\mathbf{Z}_{t-1} = Z_{t-1}$ είναι συνεχής ερμηνευτική μεταβλητή. Για συγκεκριμένη κατηγορία $j = 1, 2, \ldots, m-1$, η καμπύλη απόκρισης (*response*

curve) $f(Z_{t-1}) = P(Y_t \leq j \mid Z_{t-1}), \quad t = 1, 2, \ldots, N$, είναι η καμπύλη λογιστικής παλινδρόμησης για μια δίτιμη απόκριση έστω W_t, με $W_t = 1$ αν $Y_t \leq j$ και $W_t = 0$ όταν $Y_t > j$. Πράγματι, το μοντέλο (5.13) στην δεδομένη περίπτωση μπορεί να γραφεί

$$P(Y_t \leq j \mid Z_{t-1}) = \frac{\exp(\theta_j + \gamma Z_{t-1})}{1 + \exp(\theta_j + \gamma Z_{t-1})}. \tag{5.15}$$

από την (5.15) μπορούμε να συμπεράνουμε ότι οι καμπύλες απόκρισης για κάθε $j = 1, 2, \ldots, m-1$ θα έχουν ακριβώς το ίδιο σχήμα. Συγκεκριμένα, θα έχουν την ίδια μονοτονία σε όλο το πεδίο τιμών της f αλλά θα είναι οριζόντια μετατοπισμένες. Έστω λοιπόν ότι έχουμε τις κατηγορίες j και k της Y_t με $j < k$. Τότε για τις πιθανότητες $P(Y_t \leq j \mid Z_{t-1})$ και $P(Y_t \leq k \mid Z_{t-1})$ θα ισχύει η (5.15) με $\theta_j < \theta_k$. Στην δεδομένη περίπτωση η καμπύλη $P(Y_t \leq k \mid Z_{t-1})$ προκύπτει από την μετατόπιση της $P(Y_t \leq j \mid Z_{t-1})$ κατά $\frac{\theta_k - \theta_j}{\gamma}$ μονάδες αριστερά στον άξονα της Z_{t-1}. Πράγματι

$$P(Y_t \leq j \mid Z_{t-1} = Z_{t-1} + \frac{\theta_k - \theta_j}{\gamma}) = \frac{\exp(\theta_j + \gamma(Z_{t-1} + \frac{\theta_k - \theta_j}{\gamma}))}{1 + \exp(\theta_j + \gamma(Z_{t-1} + \frac{\theta_k - \theta_j}{\gamma}))}$$

$$\frac{\exp(\theta_j + \gamma Z_{t-1} + \theta_k - \theta_j)}{1 + \exp(\theta_j + \gamma Z_{t-1} + \theta_k - \theta_j)} = \frac{\exp(\theta_k + \gamma Z_{t-1})}{1 + \exp(\theta_k + \gamma Z_{t-1})} =$$

$$= P(Y_t \leq k \mid Z_{t-1}).$$

5.3.3 Συνάρτηση Μερικής Πιθανοφάνειας του μοντέλου (5.13)

Η αντίστοιχη συνάρτηση μερικής πιθανοφάνειας που στηρίζεται στο *proportional odds model* προκύπτει ως ακολούθως

$$PL(\theta_1, \theta_2, \ldots, \theta_{m-1}, \boldsymbol{\gamma}) = \prod_{t=1}^{N} \prod_{j=1}^{m} \pi_{tj}^{y_{tj}} = \prod_{t=1}^{N} \prod_{j=1}^{m} P(Y_t = j \mid \mathbf{Z}_{t-1})$$

$$= \prod_{t=1}^{N} \Big\{ \prod_{j=1}^{m} \Big(P(Y_t \leq j \mid \mathbf{Z}_{t-1}) - P(Y_t \leq j-1 \mid \mathbf{Z}_{t-1}) \Big)^{y_{tj}} \Big\}$$

$$= \prod_{t=1}^{N} \Big\{ \prod_{j=1}^{m} \Big(\frac{\exp(\theta_j + \boldsymbol{\gamma}' \mathbf{Z}_{t-1})}{1 + \exp(\theta_j + \boldsymbol{\gamma}' \mathbf{Z}_{t-1})} \Big) - \frac{\exp(\theta_{j-1} + \boldsymbol{\gamma}' \mathbf{Z}_{t-1})}{1 + \exp(\theta_{j-1} + \boldsymbol{\gamma}' \mathbf{Z}_{t-1})} \Big)^{y_{tj}} \Big\}.$$

5.3.4 Προσέγγιση του *proportional odds* μέσω κρυφής μεταβλητής

Η περιγραφή του *proportional odds model* γίνεται περισσότερο κατανοητή μέσω μιας βοηθητικής συνεχούς μεταβλητής. Αναλυτικότερα για την δεδομένη μεταβλητή που κρύβεται πίσω από την Y_t , θεωρούμε ένα μοντέλο παλινδρόμησης το οποίο δείχνει την

κοινή επίδραση του $\boldsymbol{\gamma}$ για τα διαφορετικά j στο μοντέλο (5.13). Έστω λοιπόν η X_t που συνιστά την μη παρατηρήσιμη στοχαστική διαδικασία $\{X_t\}$ που δημιούργησε την $\{Y_t\}, t = 1, 2, \ldots, N$. Η συγκεκριμένη μεταβλητή λέγεται κρυφή (*latent*) μεταβλητή και έχει αθροιστική συνάρτηση κατανομής (*cdf*) $F(x_t+\eta)$. Οι τιμές της X_t διαφέρουν γύρω από μια παράμετρο θέσης η (π.χ ο μέσος), η οποία εξαρτάται από το διάνυσμα των συμμεταβλητών $\mathbf{Z}_{t-1}$ σύμφωνα με την σχέση $\eta(\mathbf{Z}_{t-1}) = \boldsymbol{\gamma}'\mathbf{Z}_{t-1}$. Έστω $-\infty = \theta_0 < \theta_1 < \ldots < \theta_m = \infty$ τα σημεία διακοπής της συνεχούς κλίμακας, τα οποία καθορίζουν $m+1$ διαστήματα τιμών για την X_t, τέτοια ώστε

$$Y_t = j \Leftrightarrow Y_{tj} = 1 \Leftrightarrow \theta_{j-1} \leq X_t < \theta_j \quad \mu\epsilon \quad I_j = [\theta_{j-1}, \theta_j), \quad j = 1, 2 \ldots, m.$$

Δηλαδή μπορούμε να πούμε ότι η Y_t θα πάρει την κατηγορία j αν και μόνο αν η X_t λάβει τιμή στο $j-$οστό διάστημα τιμών της. Επομένως για την αθροιστική πιθανότητα θα ισχύει

$$P(Y_t \leq j \mid F_{t-1}) = P(X_t < \theta_j \mid \mathcal{F}_{t-1}) = F(\theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) \tag{5.16}$$

για $j = 1, 2, \ldots, m$. Ακόμη για την π_{tj} θα έχουμε

$$\pi_{tj} = P(Y_t = j \mid \mathcal{F}_{t-1}) = P(\theta_{j-1} \leq X_t < \theta_j \mid \mathcal{F}_{t-1}) =$$

$$P(X_t < \theta_j \mid \mathcal{F}_{t-1}) - P(X_t \leq \theta_{j-1} \mid \mathcal{F}_{t-1}) \overset{(5.16)}{\Rightarrow}$$

$$\pi_{tj} = F(\theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) - F(\theta_{j-1} + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) \tag{5.17}$$

για $j = 1, 2, \ldots, m$. Στην σχέση (5.16) αν επιλέξουμε η F να είναι η λογιστική συνάρτηση $F_\ell = \frac{e^x}{1+e^x}$ τότε αυτή γράφεται

$$P(Y_t \leq j \mid \mathcal{F}_{t-1}) = \frac{\exp(\theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1})}{1 + \exp(\theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1})} \Rightarrow \log\Big\{\frac{P(Y_t \leq j \mid \mathcal{F}_{t-1})}{P(Y_t > j \mid \mathcal{F}_{t-1})}\Big\} = \theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}.$$

για $j = 1, 2, \ldots, m$, που είναι το *proportional odds* μοντέλο.

Άλλες επιλογές για την F περιλαμβάνουν την συνάρτηση κατανομής της τυπικής κανονικής κατανομής

$$F \equiv \Phi,$$

την συνάρτηση κατανομής της *extreme minimal* κατανομής

$$F(x) \equiv 1 - \exp(-\exp(x)),$$

και τη συνάρτηση κατανομής της *extreme maximal* κατανομής

$$F(x) \equiv \exp(-\exp(-x)).$$

Γενικά οποιαδήποτε συνάρτηση σύνδεσης που χρησιμοποιείται στις δίτιμες χρονοσειρές είναι κατάλληλη όταν εισάγουμε ένα μοντέλο *cumulative logit*.

Τα προαναφερθέντα ουσιαστικά προέκυψαν θεωρώντας το ακόλουθο μοντέλο παλινδρόμησης για την X_t

$$X_t = -\boldsymbol{\gamma}'\mathbf{Z}_{t-1} + e_t \tag{5.18}$$

όπου e_t είναι μια ακολουθία ανεξάρτητων και ισόνομων τυχαίων μεταβλητών με συνεχή αθροιστική συνάρτηση κατανομής την F. Πράγματι

$$P(e_t \leq x) = P(X_t + \boldsymbol{\gamma}'\mathbf{Z}_{t-1} \leq x) = P(X_t \leq -\boldsymbol{\gamma}'\mathbf{Z}_{t-1} + x) =$$

$$F(-\boldsymbol{\gamma}'\mathbf{Z}_{t-1} + x + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) = F(x).$$

Επειδή το διάνυσμα $\boldsymbol{\gamma}$ έχει την ίδια διάσταση με το συμμεταβλητό διάνυσμα $\mathbf{Z}_{t-1}$ μπορούμε να οδηγηθούμε στα ίδια αποτελέσματα δουλεύοντας πλέον με την *cdf* F του e_t αντί της $F(x_t + \eta)$ της X_t. Ενδεικτικά θα δείξουμε πώς μπορεί να προκύψει μέσω της F η (5.17) για τις π_{tj}.

$$\pi_{tj} = P(Y_t = j \mid \mathcal{F}_{t-1}) = P(\theta_{j-1} \leq X_t < \theta_j \mid \mathcal{F}_{t-1}) =$$

$$P(X_t < \theta_j \mid \mathcal{F}_{t-1}) - P(X_t < \theta_{j-1} \mid \mathcal{F}_{t-1}) =$$

$$P(-\boldsymbol{\gamma}'\mathbf{Z}_{t-1} + e_t < \theta_j \mid \mathcal{F}_{t-1}) - P(-\boldsymbol{\gamma}'\mathbf{Z}_{t-1} + e_t < \theta_{j-1} \mid \mathcal{F}_{t-1}) =$$

$$P(e_t < \theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1} \mid \mathcal{F}_{t-1}) - P(e_t < \theta_{j-1} + \boldsymbol{\gamma}'\mathbf{Z}_{t-1} \mid \mathcal{F}_{t-1}) =$$

$$F(\theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) - F(\theta_{j-1} + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}).$$

Συνάμα το γενικό μοντέλο παλινδρόμησης για διατάξιμες κατηγορικές χρονοσειρές της (5.16) θα προκύψει εναλλακτικά ως εξής

$$P(Y_t \leq j \mid \mathcal{F}_{t-1}) = P(Y_t < j \mid \mathcal{F}_{t-1}) + P(Y_t = j \mid \mathcal{F}_{t-1}) =$$

$$P(X_t < \theta_{j-1} \mid \mathcal{F}_{t-1}) + \pi_{tj} = P(-\boldsymbol{\gamma}'\mathbf{Z}_{t-1} + e_t < \theta_{j-1}) + \pi_{tj} =$$

$$P(e_t < \boldsymbol{\gamma}'\mathbf{Z}_{t-1} + \theta_{j-1} \mid \mathcal{F}_{t-1}) + \pi_{tj} = F(\theta_{j-1} + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) + \pi_{tj} \overset{(5.17)}{=}$$

$$F(\theta_{j-1} + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) + F(\theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) - F(\theta_{j-1} + \boldsymbol{\gamma}'\mathbf{Z}_{t-1}) = F(\theta_j + \boldsymbol{\gamma}'\mathbf{Z}_{t-1})$$

για $j = 1, 2, \ldots, m$.

Παρατήρηση 5.3.2 Το μοντέλο (5.16) αποτελεί ειδική περίπτωση του γενικού μοντέλου παλινδρόμησης για κατηγορικές χρονοσειρές που συναντήσαμε στο Κεφάλαιο 4 και το οποίο δίνεται από την σχέση

$$\boldsymbol{\pi}_t(\boldsymbol{\beta}) = E(\mathbf{Y}_t \mid \mathcal{F}_{t-1}) = \boldsymbol{h}(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) = \boldsymbol{h}(\boldsymbol{\eta}_t).$$

ή πιο αναλυτικά

$$\boldsymbol{\pi}_t(\boldsymbol{\beta}) = \begin{pmatrix} \pi_{t1}(\boldsymbol{\beta}) \\ \pi_{t2}(\boldsymbol{\beta}) \\ \vdots \\ \pi_{tq}(\boldsymbol{\beta}) \end{pmatrix} = \begin{pmatrix} E(Y_{t1} \mid \mathcal{F}_{t-1}) \\ E(Y_{t2} \mid \mathcal{F}_{t-1}) \\ \vdots \\ E(Y_{tq} \mid \mathcal{F}_{t-1}) \end{pmatrix} = \begin{pmatrix} P(Y_{t1} = 1 \mid \mathcal{F}_{t-1}) \\ P(Y_{t2} = 1 \mid \mathcal{F}_{t-1}) \\ \vdots \\ P(Y_{tq} = 1 \mid \mathcal{F}_{t-1}) \end{pmatrix} = \begin{pmatrix} h_1(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) \\ h_2(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) \\ \vdots \\ h_q(\mathbf{Z}'_{t-1} \cdot \boldsymbol{\beta}) \end{pmatrix}.$$

Για το σκοπό αυτό θεωρούμε το διάνυσμα $\boldsymbol{\beta}$ διάστασης $q + d$

$$\boldsymbol{\beta} = (\theta_1, \theta_2, \ldots, \theta_q, \boldsymbol{\gamma}')'.$$

Ακόμη θεωρούμε τον πίνακα

$$\mathbf{Z}_{t-1} = \begin{pmatrix} 1 & 0 & \ldots & 0 \\ 0 & 1 & \ldots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \ldots & 1 \\ \boldsymbol{z}_{t-1} & \boldsymbol{z}_{t-1} & \ldots & \boldsymbol{z}_{t-1} \end{pmatrix}$$

όπου $\boldsymbol{z}_{t-1} = (z_{(t-1)1}, z_{(t-1)2}, \ldots, z(t-1)d)'$ είναι ο συντελεστής του διανύσματος $\boldsymbol{\gamma}$ στην (5.16). Επιπλέον θεωρούμε την q-διάστατη συνάρτηση $\boldsymbol{h} = (h_1, \ldots, h_q)'$, με

$$\pi_{t1}(\boldsymbol{\beta}) = h_1(\boldsymbol{\eta}_t) = F(\eta_{t1}),$$

$$\pi_{tj}(\boldsymbol{\beta}) = h_j(\boldsymbol{\eta}_t) = F(\eta_{tj}) - F(\eta_{t(j-1)}), \quad j = 1, 2, \ldots, q = m - 1.$$

Για το q-διάστατο διάνυσμα $\boldsymbol{\eta}_t$ ισχύει

$$\boldsymbol{\eta}_t = (\eta_{t1}, \eta_{t2}, \ldots, \eta_{tq})' = \mathbf{Z}'_{t-1}\boldsymbol{\beta} = \begin{pmatrix} \theta_1 + \boldsymbol{\gamma}'\boldsymbol{z}_{t-1} \\ \theta_2 + \boldsymbol{\gamma}'\boldsymbol{z}_{t-1} \\ \vdots \\ \theta_q + \boldsymbol{\gamma}'\boldsymbol{z}_{t-1} \end{pmatrix}.$$

Σύμφωνα με τα προαναφερθέντα και θέτοντας $p = q + d$ συμπεραίνουμε ότι το μοντέλο (5.16) ικανοποιεί το γενικό μοντέλο παλινδρόμησης για κατηγορικές χρονοσειρές.

5.4 Εναλλακτικά Μοντέλα στις Διατάξιμες Χρονοσειρές

Για την ανάλυση διατάξιμων κατηγορικών χρονοσειρών αξίζει να αναφέρουμε και τα ακόλουθα μοντέλα.

Μοντέλο *continuation ratio*.

Το δεδομένο μοντέλο ορίζεται σύμφωνα με την σχέση

$$F^{-1}\Big(\frac{\pi_{tj}(\boldsymbol{\beta})}{\pi_{t(j+1)}(\boldsymbol{\beta})+\ldots+\pi_{tm}(\boldsymbol{\beta})}\Big)=\boldsymbol{\beta}'\boldsymbol{z}_{t-1}, \tag{5.19}$$

Μοντέλο *adjacent categories logits*.

Το συγκεκριμένο μοντέλο ορίζεται από την σχέση

$$P(Y_t=j \mid Y_t\in\{r,r+1\},\mathcal{F}_{t-1})=F(\boldsymbol{\beta}'\boldsymbol{z}_{t-1}), \tag{5.20}$$

όπου η F είναι συνεχής συνάρτηση κατανομής, $\boldsymbol{z}_{t-1}$ το διάνυσμα των συμμεταβλητών και $\boldsymbol{\beta}$ το διάνυσμα των παραμέτρων (*Agresti*, (2002), κεφάλαιο 7). Ο ενδιαφερόμενος αναγνώστης για περισσότερες πληροφορίες σχετικά με τις εναλλακτικές τεχνικές μοντελοποίσης των εξαρτημένων διατακτικών και διαστηματικών μεταβλητών παραπέμπεται στους (*Fahrmeir and Tutz*, (2001), κεφάλαιο 3) και (*Johnson and Albert* (1999)).

Κεφάλαιο 6

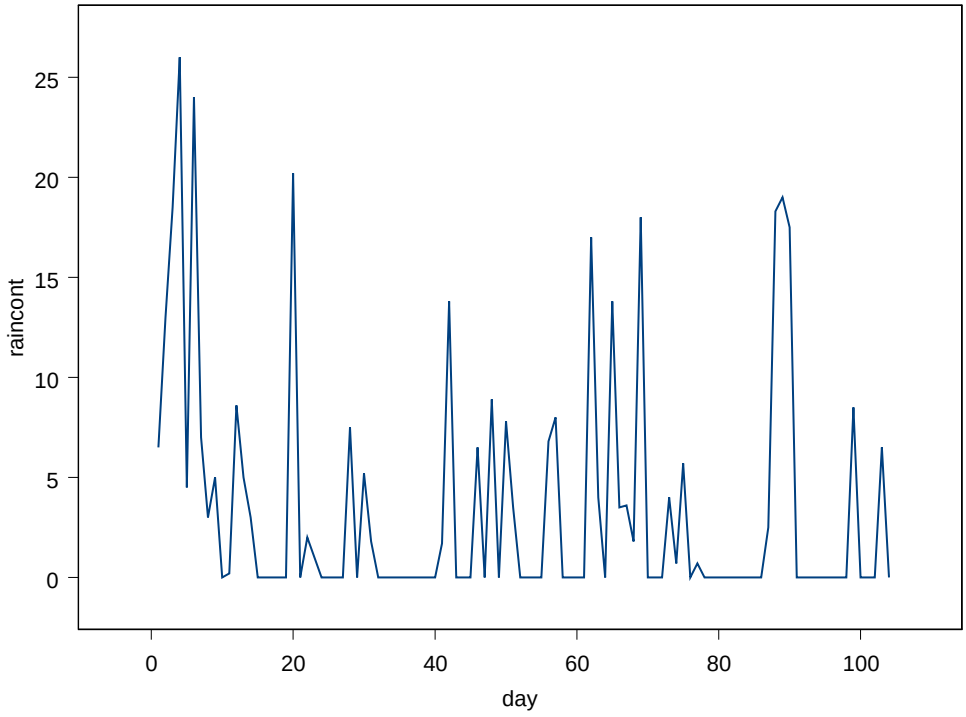
Μοντελοποίηση Δεδομένων Βροχόπτωσης

6.1 Εισαγωγή

Στο δεδομένο κεφάλαιο θα προσπαθήσουμε να επιλέξουμε ένα ή περισσότερα μοντέλα τα οποία θα περιγράφουν σε ικανοποιητικό βαθμό την συμπεριφορά του φαινομένου της βροχόπτωσης στο νομό Ιωαννίνων.

Αναμφισβήτητα το δεδομένο φαινόμενο είναι στοχαστικό αφού η μελλοντική του εξέλιξη δεν μπορεί να προβλεφθεί με ακρίβεια. Η δομή του καθίσταται ακόμη πιο πολύπλοκη επειδή οι βροχοπτώσεις εξαρτώνται σε μεγάλο βαθμό απο την γεωγραφική περιοχή που αναφέρονται. Για παράδειγμα στον Ελλαδικό χώρο, που απο μόνος του αποτελεί ένα περιορισμένο πεδίο, τα επίπεδα της βροχής στους δυτικούς ηπειρωτικούς νομούς διαφέρουν σημαντικά απο τους ανατολικούς. Σε πολλές περιπτώσεις, ακόμη και γειτονικοί νομοί είτε της δυτικής είτε της ανατολικής ηπειρωτικής Ελλάδας παρουσιάζουν τελείως διαφορετική συμπεριφορά ως προς τις βροχοπτώσεις τους. Στο σημείο αυτό θα πρέπει να τονίσουμε πως η βροχή σχετίζεται πρωτίστως με την ατμοσφαιρική κυκλοφορία. Το γεγονός αυτό σε συνδυασμό με την σύνδεση των βροχοπτώσεων και με τοπικούς παράγοντες διαμορφώνει επιπλέον μια σχέση της βροχής με την θερμοκρασία, την υγρασία και την ατμοσφαιρική πίεση της κάθε περιοχής (Κατσούλης και Μπαρτζώκας (2003)).

Σύμφωνα με τα προαναφερθέντα γίνεται φανερό, πως το φαινόμενο της βροχής είναι ιδιαίτερα σύνθετο και απαιτείται ιδιαίτερη προσοχή κατά την διαδικασία μοντελοποίησης του. Ο στατιστικός πρέπει να περιορίζει την έρευνα του σε συγκεκριμένη περιοχή ζητώντας από τον μετεωρολόγο να τον πληροφορήσει για τους τοπικούς παράγοντες βροχόπτωσης.



Σχήμα 6.1: Χρονοσειρά βροχοπτώσεων για $N = 104$ ημέρες

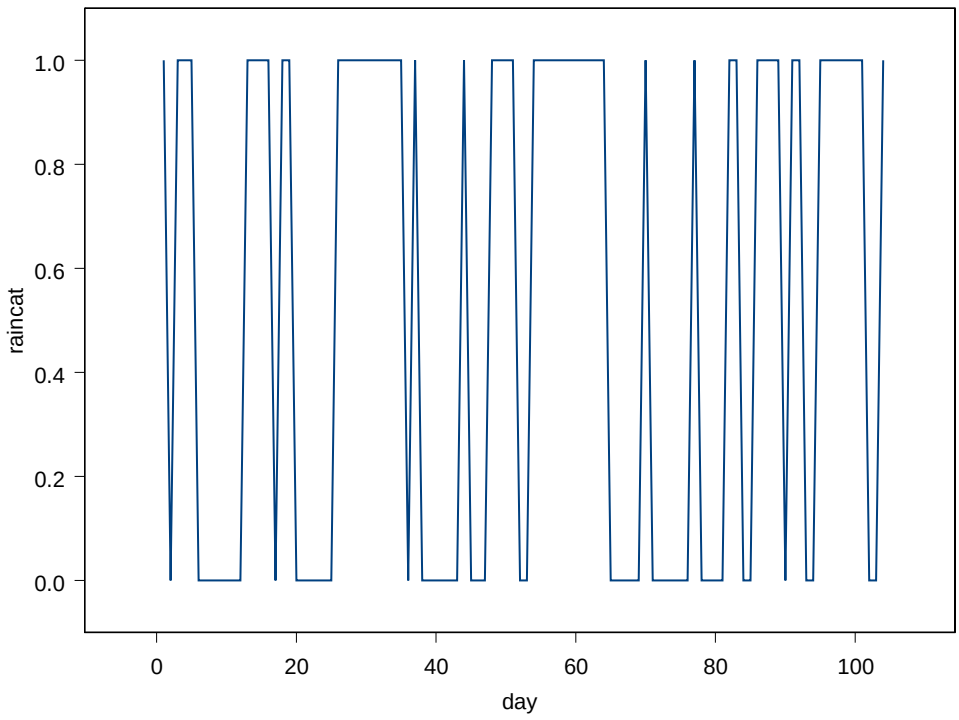
6.2 Βροχοπτώσεις στο νομό Ιωαννίνων

6.2.1 Παρουσίαση των δεδομένων

Με σκοπό την μεντελοποίηση των βροχοπτώσεων στον νομό Ιωαννίνων διαθέτουμε την ημερήσια βοχόπτωση σε 'mm' για μια περίοδο $N = 1823$ ημερών απο τις 4.1.1995 έως την 31.12.1999. Η δεδομένη χρονοσειρά παρουσιάζεται στο Σχήμα 6.1. Αναθέτοντας στην μεταβλητή Y_t την τιμή '1' εφόσον η ημερήσια βροχόπτωση είναι '≥ 0.1 mm' (που σημαίνει ότι έβρεξε) ή την τιμή '0' εφόσον το επίπεδο της βροχής είναι '< 0.1 mm' (που σημαίνει ότι δεν έβρεξε) προκύπτει μια αντίστοιχη δίτιμη χρονοσειρά που βρίσκεται στο Σχήμα 6.2. Συνάμα διαθέτουμε για τις 1823 ημέρες μετρήσεις της θερμοκρασίας (T), της υγρασίας (H) και της ατμοσφαιρικής πίεσης (A). Αναλυτικότερα για την θερμοκρασία ανά ημέρα γνωρίζουμε την μέση, την ελάχιστη και την μέγιστη τιμή της. Όσον αφορά στην υγρασία και την ατμοσφαιρική πίεση για κάθε ημέρα διαθέτουμε την τιμή τους ανά δυο ώρες[1].

Για την αξιοποίηση της πληροφορίας των συμμεταβλητών μας, άρα και την καλύ-

[1] Δηλαδή έχουμε 12 μετρήσεις υγρασίας και ατμοσφαιρικής πίεσης ημερησίως.



Σχήμα 6.2: Δίτιμη χρονοσειρά βροχοπτώσεων για $N = 104$ ημέρες

τερη προβλεψιμότητα του μοντέλου που θα καταλήξουμε χρησιμοποιήσαμε το εύρος της θερμοκρασίας ανά ημέρα, ενώ για την υγρασία και την ατμοσφαιρική πίεση αντιστοίχως υπολογίσαμε τις μέσες τιμές των 12 μετρήσεων που διαθέτουμε για αυτές ημερησίως. Η επιλογή του εύρους της θερμοκρασίας επιτρέπει την καλύτερη ερμηνεία του φαινομένου της βροχής και δικαιολογείται από μετεωρολογική σκοπιά. Πράγματι σύμφωνα με τα κλιματολογικά αποτελέσματα, τις ημέρες που έχουμε συννεφιά (και στις οποίες συνήθως βρέχει), παρατηρείται σχετικά μικρότερο εύρος θερμοκρασίας από αυτό που καταγράφεται τις ημέρες που επικρατεί ξαστεριά. Τέλος επειδή διαθέτουμε δίτιμα εξαρτημένα δεδομένα θα χρησιμοποιήσουμε το μοντέλο (3.8) λογιστικής παλινδρόμησης που παρουσιάσαμε στο $2^{ο}$ κεφάλαιο. Βάσει των προαναφερθέντων θα προχωρήσουμε στην διαδικασία μοντελοποίησης των βροχοπτώσεων του νομού Ιωαννίνων.

6.3 Μοντελοποίηση ολόκληρης της χρονοσειράς

Σύμφωνα με την δεδομένη προσέγγιση θα διερευνήσουμε το φαινόμενο της βροχόπτωσης καθόλη την διάρκεια του χρόνου, χωρίς να λαμβάνουμε υπόψη τις 4 εποχές του

για τις οποίες όπως γνωρίζουμε οι συχνότητες και οι ποσότητες των βροχοπτώσεων διαφέρουν. Επειδή το φαινόμενο που μελετάμε παρουσιάζει διαχρονική εξάρτηση στις συμμεταβλητές θα συμπεριλάβουμε τις χρονικές υστερήσεις $1^{ης}$, $2^{ης}$ και $3^{ης}$ τάξης τόσο της Y_t όσο και της θερμοκρασίας (T), της υγρασίας (H) καθώς και της ατμοσφαιρικής πίεσης (A). Έτσι σε πρώτη φάση το διάνυσμα $\mathbf{Z}_{t-1}$ θα αποτελείται από τις χρονοεξαρτώμενες μεταβλητές

$$Y_{t-1}, Y_{t-2}, Y_{t-3}, T_t, T_{t-1}, T_{t-2}, T_{t-3}, H_t, H_{t-1}, H_{t-2}, H_{t-3}, A_t, A_{t-1}, A_{t-2}, A_{t-3}.$$

Αναμφισβήτητα το αντίστοιχο υπόδειγμα, που το ονομάζουμε M1, και το οποίο θα προκύψει σύμφωνα με την σχέση (3.8) είναι ακατάλληλο για την ανάλυση των δεδομένων αφού παραβιάζει την βασική στατιστική αρχή της 'οικονομικής' μοντελοποίησης (*parsimonious modelling*). Η προσαρμογή του δεδομένου μοντέλου θα γίνει για την διευκόλυνση της παρουσίασης επιτρέποντας συγκριτικά αποτελέσματα.

Για την επιλογή του βέλτιστου μοντέλου θα στηριχτούμε στο κριτήριο πληροφορίας του AIC. Σύμφωνα με το συγκεκριμένο διαγνωστικό κριτήριο την καλύτερη προσαρμογή στα δεδομένα παρουσιάζει το μοντέλο με την μικρότερη τιμή του AIC. Στο *S-PLUS* η δεδομένη διαδικασία πραγματοποιείται αυτόματα μέσω της συνάρτησης *stepAIC*. Τρέχοντας το κατάλληλο *script* (βλέπε παράρτημα Δ) το λογιστικό μοντέλο που καταλήγουμε περιέχει τις συμμεταβλητές

$$Y_{t-1}, Y_{t-2}, Y_{t-3}, T_t, T_{t-1}, T_{t-2}, H_t, H_{t-1}, H_{t-2}, H_{t-3}, A_t, A_{t-1}, A_{t-2}$$

και το συμβολίζουμε με M2. Για τα μοντέλα M1 και M2 συνοπτικά θα έχουμε τα ακόλουθα

Πίνακας 6.1 Διαγνωστικά κριτήρια για τα M1 και M2

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
M1	17	0.1058	1625.847	1222.910	1806	1256.91	1350.55
M2	14	0.1062	1600.326	1226.939	1809	1254.939	1332.0543

Σύμφωνα με τα παραπάνω γίνεται φανερό ότι ούτε το μοντέλο M2, το οποίο έχει μικρότερο AIC από το M1, δεν μπορεί να χρησιμοποιηθεί για στατιστική επεξεργασία αφού διαθέτει αρκετά μεγάλο αριθμό παραμέτρων.

Εναλλακτική προσέγγιση για την μοντελοποίηση των βροχών είναι εκείνη η οποία στηρίζεται αποκλειστικά στις χρονικές υστερήσεις της μεταβλητής Y_t (βλέπε *Fokiano*

and Kedem, (2002), σελίδα 72). Έχοντας θεωρήσει ότι η βροχή επηρεάζεται από τις βροχοπτώσεις των τριών προηγούμενων ημερών θα προσαρμόσουμε αυτοπαλίνδρομα λογιστικά μοντέλα, των οποίων οι γραμμικές προβλέψεις η_t θα δίνονται στον ακόλουθο πίνακα

Πίνακας 6.2 Γραμμικές προβλέψεις αυτοπαλίνδρομων λογιστικών μοντέλων

M1F	$\beta_0 + \beta_1 Y_{t-1}$
M2F	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2}$
M3F	$\beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \beta_3 Y_{t-3}$
M4F	$\beta_0 + \beta_1 Y_{t-1} + \beta_3 Y_{t-3}$

Για τα δεδομένα υποδείγματα έχουμε (βλέπε παράρτημα Δ)

Πίνακας 6.3 Διαγνωστικά κριτήρια για τα αυτοπαλίνδρομα μοντέλα

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
M1F	2	0.1788	1811.814	1967.757	1821	1971.757	1982.773
M2F	3	0.1769	1794.295	1945.115	1820	1951.115	1967.6397
M3F	4	0.1768	1791.496	1942.965	1819	1590.965	1972.9979
M4F	3	0.1782	1803.470	1959.167	1820	1965.167	1981.6917

Απο τον τελευταίο πίνακα διαπιστώνουμε πως την καλύτερη προσαρμογή στα δεδομένα την παρουσιάζει το μοντέλο M3F. Αυτό συμβαίνει διότι παρουσιάζει την μικρότερη τιμή για το AIC σε σχέση με τα υπόλοιπα μοντέλα. Το ίδιο συμβαίνει και με τα υπόλοιπα διαγνωστικά κριτήρια. Το μοντέλο M3F όμως εξακολουθεί να έχει μεγαλύτερη τιμή για το AIC, καθώς και για τους υπόλοιπους διαγνωστικούς ελέγχους, από εκείνη που παρουσιάζει το υπόδειγμα M2. Έτσι, το M3F παρόλλο που έχει μικρότερο αριθμό παραμέτρων δεν θα χρησιμοποιηθεί για πρόβλεψη.

Συνοψίζοντας την ανάλυση μας βλέπουμε πως τα αποτελέσματα που προέκυψαν συμφωνούν με την αρχική μας παρατήρηση, πως το φαινόμενο της βροχής είναι εξαιρετικά πολύπλοκο και για την ορθότερη διερεύνηση του ο στατιστικός απαιτείται να διαθέτει επαρκή στοιχεία των τοπικών παραγόντων που το επηρεάζουν. Έτσι στο σημείο αυτό θα προχωρήσουμε την ανάλυση μας αξιοποιώντας τα κλιματολογικά χαρακτηριστικά του νομού Ιωαννίνων.

Σύμφωνα λοιπόν με τα μετεωρολογικά δεδομένα ο φυσικός μηχανισμός που διαμορφώνει τις βροχοπτώσεις στην ευρύτερη περιοχή των Ιωαννίνων διαφέρει από χειμώνα

σε καλοκαίρι. Αναλυτικότερα στους χειμερινούς μήνες οι βροχές οφείλονται σε συστήματα χαμηλών πιέσεων τα οποία κινούνται από την δυτική προς την ανατολική Μεσόγειο ενώ το καλοκαίρι οι βροχοπτώσεις προκαλούνται από τοπικούς παράγοντες θερμικής φύσης. Ειδικά τους μήνες Απρίλιο και Οκτώβρη που είναι οι μεταβατικοί περίοδοι από το χειμώνα στο καλοκαίρι και από το καλοκαίρι στο χειμώνα αντιστοίχως η συμπεριφορά των μετεωρολογικών φαινομένων γενικά καθίσταται ακόμη πιο πολύπλοκη. Οι δεδομένες ζώνες, που ονομάζονται '*buffer*' επιδεικνύουν έντονο μετεωρολογικό ενδιαφέρον (Κατσούλης και Μπαρτζώκας (2003)). Παρόλλα αυτά δεν θα τις λάβουμε υπόψη στην ανάλυση μας ώστε να έχουμε μια ξεκάθαρη διαφοροποίηση μεταξύ των χειμερινών και θερινών περιόδων.

Σύμφωνα με τα προαναφερθέντα θα μοντελοποιήσουμε ξεχωριστά τους χειμώνες από τα καλοκαίρια για κάθε έτος, εξαιρώντας από την ανάλυση μας τα δεδομένα του Απρίλη και του Οκτώβρη. Ειδικότερα η χειμερινή περίοδος θα ξεκινά τον Νοέμβρη και θα ολοκληρώνεται στο τέλος Μαρτίου ενώ η καλοκαιρινή θα διαρκεί από τον Μάιο μέχρι το τέλος Σεπτέμβρη.

Στο σημείο αυτό γεννάται το ερώτημα γιατί να μην δούμε όλους τους χειμώνες και όλα τα καλοκαίρια αντιστοίχως ως δυο εννοιέες χρονοσειρές. Μια τέτοια προσέγγιση με μια πρώτη ματιά φαίνεται προβληματική αφού μπορούμε να ισχυριστούμε ότι οι χειμώνες και τα καλοκαίρια από έτος σε έτος δεν παρουσιάζουν εξάρτηση, μια που μεταξύ τους μεσολαβεί ένα μεγάλο χρονικό διάστημα. Λαμβάνοντας όμως υπόψη ότι οι χρονοσειρές αναφέρονται ακόμη και σε ετήσιες παρατηρήσεις (π.χ. καταγραφή του πληθωρισμού ανα έτος), συμπεραίνουμε ότι οι βροχοπτώσεις από καλοκαίρι σε καλοκαίρι και από τον ένα χειμώνα στον επόμενο θα παρουσιάζουν διαχρονική εξάρτηση. Επομένως έχει νόημα να θεωρήσουμε ένα μοντέλο που θα περιγράφει τις βροχές από κοινού για όλους τους χειμώνες καθώς και ένα αντίστοιχο για τα καλοκαίρια. Η αδυναμία της δεδομένης προσέγγισης έγγυται στο γεγονός ότι πρέπει να ληφθεί υπόψη το κενό μεταξύ των χειμερινών και των θερινών ζωνών. Το *S-PLUS* καθώς και άλλα γνωστά στατιστικά πακέτα δεν μπορούν να λάβουν υπόψη τους αυτό το κενό.

6.4 Μοντελοποίηση των χειμερινών περιόδων

Επειδή είδαμε ότι τον χειμώνα οι βροχές διαμορφώνονται κυρίως από συστήματα χαμηλών πιέσεων, στα μοντέλα που θα παρουσιάσουμε για κάθε έτος, θα απομακρύνουμε τις χρονοεξαρτώμενες μεταβλητές $T_t, T_{t-1}, T_{t-2}, T_{t-3}$. Με τον τρόπο αυτό θα οδηγηθούμε

σε υποδείγματα με σχετικά μικρή διάσταση. Για κάθε χειμώνα θα προσαρμόζουμε αρχικά το μοντέλο με τις συμμεταβλητές $Y_{t-1}, Y_{t-2}, Y_{t-3}, H_t, H_{t-1}, H_{t-2}, H_{t-3}, A_t, A_{t-1}, A_{t-2}, A_{t-3}$ το οποίο θα καλούμε αρχικό μοντέλο και θα το συμβολίζουμε με το γράμμα M ακολουθούμενο με το αντίστοιχο έτος. Για την επιλογή βέλτιστου υποδείγματος και πάλι θα αξιοποιήσουμε το κριτήριο AIC.

Χειμώνας 95-96 (1.11.95 έως 31.3.1995)

Μέσω της συνάρτησης *stepAIC* το βέλτιστο υπόδειγμα είναι εκείνο με συμμεταβλητές

$$Y_{t-3}, H_t, H_{t-1}, H_{t-2}, A_{t-1}, A_{t-2}$$

το οποίο συμβολίζουμε με MX95.96*step*. Για τα μοντέλα MX95.96 (αρχικό) και MX95.96*step* προέκυψαν

Πίνακας 6.4 Διαγνωστικά κριτήρια για τα MX95.96*step* και MX95.96

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MX95.96	12	0.1040	164.4343	97.263	133	121.263	156.984
MX95.96*step*	7	0.1101	122.887	99.71135	138	113.711	134.548

Χειμώνας 96-97 (1.11.96 έως 31.3.1997)

Μέσω της συνάρτησης *stepAIC* το καλύτερο μοντέλο είναι αυτό με συμμεταβλητές

$$H_{t-1}, A_t, A_{t-1}.$$

Για το αρχικό μοντέλο της δεδομένης περιόδου MX96.97 καθώς και αυτό που μας έδωσε η συνάρτηση *stepAIC* MX96.97*step* προέκυψαν τα εξής

Πίνακας 6.5 Διαγνωστικά κριτήρια για τα MX96.97*step* και MX96.97

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MX96.97	12	0.0658	56.856	56.666	132	80.666	116.304
MX96.97*step*	4	0.07	65.607	61.336	140	69.336	81.215

Χειμώνας 97-98 (1.11.97 έως 31.3.1998)

Το μοντέλο με το μικρότερο AIC την συγκεκριμένη χειμερινή ζώνη θα έχει τις ακόλουθες χρονοεξαρτώμενες συμμεταβλητές

$$H_t, H_{t-1}, A_t, A_{t-1}$$

Για το δεδομένο υπόδειγμα που το συμβολίζουμε με MX97.98*step* καθώς και το αρχικό MX97.98 έχουμε

Πίνακας 6.6 Διαγνωστικά κριτήρια για τα MX97.98*step* και MX97.98

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MX97.98	12	0.1099	128.5615	100.839	132	124.839	160.477
MX97.98*step*	5	0.1169	126.399	106.585	139	116.5848	131.434

Χειμώνας 98-99 (1.11.98 έως 31.3.1999)

Το βέλτιστο υπόδειγμα για τον χειμώνα 1998-1999 βάση του AIC θα είναι εκείνο με τις συμμεταβλητές

$$Y_{t-1}, Y_{t-2}, H_{t-1}, H_{t-2}, A_t.$$

Για το συγκεκριμένο υπόδειγμα MX98.99*step* καθώς και το αρχικό μοντέλο MX98.99 έχουμε

Πίνακας 6.7 Διαγνωστικά κριτήρια για τα MX98.99*step* και MX98.99

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
M98.99	12	0.1284	182.0125	120.1459	132	144.146	179.7836
M98.99*step*	6	0.1305	225.232	124.236	138	136.236	154.0547

Βάσει των προαναφερθέντων παρατηρούμε ότι όλα τα λογιστικά μοντέλα που λάβαμε μέσω της συνάρτησης *stepAIC* έχουν καλύτερη προσαρμογή από όλα τα αρχικά μοντέλα σε κάθε χειμερινή περίοδο. Ειδικά τον χειμώνα του 1996-1997 το υπόδειγμα MX96.97*stepAIC* έχει την καλύτερη προσαρμογή σε σχέση με τα μοντέλα των υπολοίπων χειμερινών περιόδων (MX95.96*stepAIC*, MX97.98*stepAIC*, MX98.99*stepAIC*). Το συγκεκριμένο μοντέλο θα το συγκρίνουμε με τα αυτοπαλίνδρομα λογιστικά μοντέλα που θεωρήσαμε και στην περίπτωση της δειγματοληπτικής διαδρομής μεγέθους

N=1823. Για την περίπτωση μας, δηλαδή τον χειμώνα 1996-97 τα δεδομένα μοντέλα από M1F, M2F, M3F, M4F θα τα συμβολίζω με MX1F, MX2F, MX3F, MX4F. Μέσω του *S-PLUS* προέκυψαν τα ακόλουθα

Πίνακας 6.8 Διαγνωστικά κριτήρια για τα αυτοπαλίνδρομα μοντέλα του χειμώνα 1996-97

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MX1F	2	0.1633	142.159	145.046	142	149.046	144.986
MX2F	3	0.1619	138.926	142.226	141	148.226	157.135
MX3F	4	0.1590	138.786	139.956	140	147.956	159.835
MX4F	3	0.1596	140.657	141.085	141	147.085	155.995

Το καλύτερο αυτοπαλίνρομο λογιστικό υπόδειγμα σύμφωνα με τα παραπάνω είναι το MX4F. Το δεδομένο υπόδειγμα αν και δεν είναι επαρκέστερο από το MX96.97*stepAIC*, θα προτιμηθεί για πρόβλεψη επειδή δεν απαιτεί επιπλέον την πρόβλεψη των συμμεταβλητών H_t και A_t που είναι τυχαίες χρονοεξαρτώμενες. Το μοντέλο MX96.97*stepAIC* σε περίπτωση που δεν γνωρίζουμε τις μελλοντικές τιμές των H_t και A_t, παρέχει την πρόβλεψη της Y_t μόνο για την επόμενη μέρα.

Προβλέψεις 3 ημερών για τον χειμώνα 1996-97 μέσω του μοντέλου MX4F

Το μοντέλο MX4F αναφέρεται σε μια δειγματοληπτική διαδρομή μήκους N=144. Μέσω αυτού του υποδείγματος θα προβλέψουμε αν θα βρέξει ή όχι τις 3 επόμενες ημέρες. Ουσιαστικά θα προβλέψουμε τις δεσμευμένες πιθανότητες βροχής δοθέντος της ιστορίας του φαινομένου. Τα αποτελέσματα που θα προκύψουν θα συγκριθούν με τις τιμές τις Y_t, τις δεδομένες ημέρες που τις γνωρίζουμε. Το διάνυσμα των παραμέτρων του μοντέλου MX4F, όπως προέκυψε από τις 144 ημέρες είναι

$$\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1, \beta_3)' = (-1.889, 1.963, 0.887)'$$

και επομένως η δεσμευμένη πιθανότητα βροχής την 145^{η} ημέρα σύμφωνα με την σχέση (3.7) θα δίνεται ως

$$\hat{P}(Y_{145} = 1 \mid \mathcal{F}_{144}) = \frac{\exp(-1.889 + 1.963Y_{144} + 0.887Y_{142})}{1 + \exp(-1.889 + 1.963Y_{144} + 0.887Y_{142})}$$

$$= \frac{\exp(-1.889 + 1.963 \times 0 + 0.887 \times 0)}{1 + \exp(-1.889 + 1.963 \times 0 + 0.887 \times 0)} \Rightarrow$$

$$\hat{P}(Y_{145} = 1 \mid \mathcal{F}_{144}) = 0.13. \tag{6.1}$$

Σύμφωνα με την σχέση (6.1) την 145^{η} ημέρα η πιθανότητα βροχής είναι αρκετά μικρή (< 0.5) και επομένως μπορούμε να θεωρήσουμε για την μέρα εκείνη ότι $\hat{Y}_{145} = 0$.

Η πρόβλεψη της δεσμευμένης πιθανότητας βροχής την 146^{η} ημέρα θα προκύψει αξιοποιώντας την τιμή $\hat{Y}_{145} = 0$. Πράγματι έχουμε

$$\hat{P}(Y_{146} = 1 \mid \mathcal{F}_{145}) = 0.13. \tag{6.2}$$

Το δεδομένο αποτέλεσμα ήταν αναμενόμενο αφού $(\hat{Y}_{144}, \hat{Y}_{145}) = (0, 0)$ όπως και $(Y_{143}, Y_{144}) = (0, 0)$. *Σύμφωνα με την σχέση (6.2) την* 146^{η} *ημέρα η πιθανότητα βροχής είναι αρκετά μικρή* (< 0.5) *και επομένως μπορούμε να θεωρήσουμε για την μέρα εκείνη ότι* $\hat{Y}_{146} = 0$. *Στο σημείο αυτό βλέπουμε πως το μοντέλο μας εκτίμησε λάθος την δεσμευμένη πιθανότητα βροχής την* 146^{η} *ημέρα αφού τότε η πραγματική τιμή του* Y_{146} *ισούται με 1. Συνεχίζοντας την διαδικασία πρόβλεψης κατα αυτό τον τρόπο εύκολα θα διαπιστώσουμε ότι το μοντέλο μας πάντα θα δίνει ένδειξη ότι δεν θα βρέξει γεγονός που το καθιστά προβληματικό. Όπως θα δούμε όμως στα ομαδοποιημένα δεδομένα αυτό δεν σημαίνει πάντα ότι δεν επιλέξαμε τις κατάλληλες συμμεταβλητές.*

Ολοκληρώνοντας την ανάλυση μας για τους χειμώνες έχει ενδιαφέρον να εξετάσουμε την συμπεριφορά του μοντέλου MX96.97*stepAIC* στις υπόλοιπες χειμερινές ζώνες. Τα διαγνωστικά κριτήρια του δεδομένου υποδείγματος για όλες τις χειμερινές περιόδους δίνονται στον ακόλουθο πίνακα

Πίνακας 6.9 Διαγνωστικά κριτήρια από την προσαρμογή του MX96.97*step* σε όλους του χειμώνες.

Χειμώνας	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
95-96	4	0.1297	137.066	116.607	141	124.607	135.514
96-97	4	0.07	65.607	66.336	140	69.336	81.215
97-98	4	0.1347	145.021	122.708	140	130.708	142.587
98-99	4	0.1534	216.164	137.036	140	145.036	156.915

Παρατηρώντας τον παραπάνω πίνακα, καθώς και τους πίνακς 6.4-6.7, διαπιστώνουμε ότι η προσαρμογή του υποδείγματος MX96.97*stepAIC* δεν αλλοιώνει σημαντικά την επάρκεια των μοντέλων που είχαμε καταλήξει για τους υπόλοιπους χειμώνες, βάσει του κριτηρίου *AIC* (τα οποία ήταν MX95.96*stepAIC*, MX97.98*stepAIC*, MX98.99*stepAIC*). Το γεγονός αυτό δείχνει ότι ίσως θα ήταν ρεαλιστικό να θεωρήσουμε το μοντέλο MX96.97*stepAIC* κατάλληλο για την μοντελοποίηση των βροχοπτώσεων για κάθε χειμώνα. Σίγουρα κατι τέτοιο θα ήταν επιθυμητό αφού το δεδομένο

υπόδειγμα είναι μικρής διάστασης, συμφωνώντας παράλληλα με τα μετεωρολογικά δεδομένα τα οποία δηλώνουν τα συστήματα χαμηλών πιέσεων ως το βασικότερο αίτιο των χειμερινών βροχοπτώσεων στον νομό Ιωαννίνων. Για τις παραμέτρους του μοντέλου MX96.97*stepAIC* για όλες τις χειμερινές ζώνες έχουμε

Πίνακας 6.10 Εκτιμήσεις των παραμέτρων του MX96.97*step* για κάθε χειμώνα

Παράμετροι	95-96	96-97	97-98	98-99
$\hat{\beta}_0$	178.182	411.115	193.649	200.056
$\hat{\beta}_1$ (H_{t-1})	0.112	0.166	0.103	0.147
$\hat{\beta}_2$ (A_t)	-0.213	-0.198	-0.152	-0.109
$\hat{\beta}_3$ (A_{t-1})	0.029	-0.218	-0.047	-0.100

Απο τον τελευταίο πίνακα βλέπουμε ότι τα παραμετρικά διανύσματα του υποδείγματος MX96.97*step* για όλους τους χειμώνες δεν διαφέρουν σημαντικά. Οι αποκλίσεις που παρατηρούνται είναι μικρές. Παράλληλα διαπιστώνουμε ότι οι συμμεταβλητές A_t και A_{t-1} επιδρούν αρνητικά στις βροχές σε όλους τους χειμώνες, σε αντίθεση με την H_{t-1} η οποία επηρεάζει θετικά τις βροχοπτώσεις, συμφωνώντας με τις μετεωρολογικές γνώσεις. Γίνεται λοιπόν φανερό πως το μοντέλο με συμμεταβλητές H_{t-1}, A_t, A_{t-1} μπορεί να χρησιμοποιηθεί για να ερμηνεύσει τις βροχοπτώσεις για κάθε χειμώνα στον νομό Ιωαννίνων.

Τα προαναφερθέντα συνηγορούν προς την κατασκευή ενός υποδείγματος το οποιο θα χρησιμοποιεί ταυτόχρονα τις 4 χειμερινές χρονοσειρές (περίοδοι 95-96, 96-97, 97-98, 98-99) για την εννιαία εκτίμηση του διανύσματος $\boldsymbol{\beta}$, του λογιστικού μοντέλου (3.7) το οποίο στην προκειμένη περίπτωση έχει διάνυσμα συμμεταβλητών $\mathbf{Z}_{t-1} = (H_{t-1}, A_t, A_{t-1})'$. Για την συγκεκριμένη στρατηγική μοντελοποίησης ούτε το $S-PLUS$ ούτε το $SPSS$, όπως αναφέραμε, παρέχουν τις κατάλληλες επιλογές. Το δεδομένο πρόβλημα θα αποτελέσει αντικείμενο μελλοντικής έρευνας.

6.5 Μοντελοποίηση των καλοκαιρινών περιόδων

Για την μοντελοποίση των βροχοπτώσεων κατα τους καλοκαιρινούς μήνες επειδή αυτές διαμορφώνονται από τοπικούς παράγοντες θερμικής φύσεως, στα μοντέλα λογιστικής παλινδρόμησης που θα παρουσιάσουμε, με σκοπό να πετύχουμε την μείωση της διάστασης του προβλήματος δεν θα χρησιμοποιήσουμε τις τυχαίες χρονοεξαρτώμενες

συμμεταβλητές $A_t, A_{t-1}, A_{t-2}, A_{t-3}$. Τα βήματα της ανάλυση μας είναι ανάλογα με εκείνα που ακολουθήσαμε κατα τους καλοκαιρινούς μήνες.

Καλκαίρι 95 (1.5.95 έως 30.9.1995)

Για το αρχικό μοντέλο MK95 και εκείνο που προέκυψε μέσω της συνάρτησης *stepAIC* MK95*step* (T_t, H_{t-1}, H_{t-2}) έχουμε

Πίνακας 6.11 Διαγνωστικά κριτήρια για τα MK95*step* και MK95

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MK95	12	0.1226	154.3244	116.811	134	140.811	176.6145
MK95*step*	4	0.1323	154.7179	122.564	142	130.564	122.498

Καλοκαίρι 96 (1.5.96 έως 30.9.1996)

Για το αρχικό μοντέλο MK96 και εκείνο που προέκυψε μέσω του κριτήριου *AIC* MK96*step* (T_t, H_{t-1}, H_{t-2}) προέκυψαν

Πίνακας 6.12 Διαγνωστικά κριτήρια για τα MK96*step* και MK96

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MK96	12	0.0838	92.07987	76.694	134	100.694	136.497
MK96*step*	4	0.0921	92.24663	83.115	142	91.115	103.049

Καλοκαίρι 97 (1.5.97 έως 30.9.1997)

Για τα υποδείγματα MK97 και MK97*step* $(Y_{t-3}, T_t, T_{t-1}, H_t, H_{t-1}, H_{t-3})$ έχουμε

Πίνακας 6.13 Διαγνωστικά κριτήρια για τα MK97*step* και MK97

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MK97	12	0.0363	28.94072	31.153	134	55.153	90.956
MK97*step*	7	0.0381	35.00144	33.577	139	47.577	68.462

Καλοκαίρι 98 (1.5.98 έως 30.9.1998)

Για τα μοντέλα MK98 και MK98*step* (Y_{t-1}, H_t, H_{t-1}) προέκυψαν

Πίνακας 6.14 Διαγνωστικά κριτήρια για τα MK98*step* και MK99

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MK98	12	0.0792	577.0517	81.502	134	105.502	141.305
MK98*step*	4	0.0832	516.1075	83.823	142	91.823	103.757

Καλοκαίρι 99 (1.5.99 έως 30.9.1999)

Τέλος για τα υποδείγματα MK99 και MK99*step* (H_t, H_{t-1}, H_{t-2}) ισχύουν

Πίνακας 6.15 Διαγνωστικά κριτήρια για τα MK99*step* και MK99

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MK99	12	0.0699	83.2289	61.416	124	85.416	120.367
MK99*step*	4	0.0791	72.4846	66.022	132	74.022	85.673

Σύμφωνα με τα προαναφερθέντα το υπόδειγμα με την καλύτερη προσαρμογή στα δεδομένα είναι το MK97*step*. Το δεδομένο όμως μοντέλο για τους ίδιους λόγους που αναφέραμε στο MX96.97*step* δεν θα προτιμηθεί για πρόβλεψη. Εναλλακτικά για την μοντελοποίηση της συγκεκριμένης καλοκαιρινής περιόδου θα αξιολογήσουμε και τα αυτοπαλίνδρομα λογιστικά μοντέλα που είδαμε και στον χειμώνα 1996-97, τα οποία πλέον θα τα συμβολίσουμε με MK1F (Y_{t-1}), MK2F (Y_{t-1}, Y_{t-2}), MK3F $(Y_{t-1}, Y_{t-2}, Y_{t-3})$, MK4$F$ (Y_{t-1}, Y_{t-3}). Για τα τελευταία υποδείγματα προέκυψαν

Πίνακας 6.16 Διαγνωστικά κριτήρια για τα αυτοπαλίνδρομα υποδείγματα του καλοκαιριού του 1997

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
MK1F	2	0.0735	145.942	78.608	144	82.608	88.575
MK2F	3	0.0741	142.223	78.533	143	84.533	93.484
MK3F	4	0.0736	141.265	77.509	142	85.509	97.444
MK4F	3	0.0735	141.373	77.516	143	83.516	92.467

Παρατηρούμε ότι την καλύτερη προσαρμογή στα δεδομένα παρουσιάζει το υπόδειγμα MK1F.

Προβλέψεις 3 ημερών για το καλοκαίρι 1997 μέσω του μοντέλου MK1F

Το μοντέλο MK1F αναφέρεται σε μια δειγματοληπτική διαδρομή μήκους N=146. Το διάνυσμα των παραμέτρων του μοντέλου MK1F, όπως προέκυψε από τις 146 ημέρες είναι

$$\hat{\boldsymbol{\beta}} = (\beta_0, \beta_1)' = (-2.741, 2.153)'$$

και επομένως η δεσμευμένη πιθανότητα βροχής την 147^{η} ημέρα σύμφωνα με την σχέση (3.7) θα δίνεται ως

$$\hat{P}(Y_{147} = 1 \mid \mathcal{F}_{146}) = \frac{\exp(-2.741 + 2.153 \times Y_{146})}{1 + \exp(-2.741 + 2.153 \times Y_{146})}$$

$$= \frac{\exp(-2.741 + 2.153 \times 0)}{1 + \exp(-2.741 + 2.153 \times 0)} \Rightarrow$$

$$\hat{P}(Y_{147} = 1 \mid \mathcal{F}_{146}) = 0.0606. \tag{6.3}$$

Σύμφωνα με την σχέση (6.3) την 147^{η} ημέρα η πιθανότητα βροχής είναι αρκετά μικρή (< 0.5) και επομένως μπορούμε να θεωρήσουμε για την μέρα εκείνη ότι $\hat{Y}_{147} = 0$.

Η πρόβλεψη της δεσμευμένης πιθανότητας βροχής την 148^{η} ημέρα θα προκύψει αξιοποιώντας την τιμή $\hat{Y}_{147} = 0$. Πράγματι έχουμε

$$\hat{P}(Y_{148} = 1 \mid \mathcal{F}_{147}) = 0.0606. \tag{6.4}$$

Σύμφωνα με την σχέση (6.4) την 148^{η} ημέρα η πιθανότητα βροχής είναι αρκετά μικρή (< 0.5) και επομένως μπορούμε να θεωρήσουμε για την μέρα εκείνη ότι $\hat{Y}_{148} = 0$. Με όμοιο τρόπο βρίσκουμε για την 149^{η} ημέρα ότι $\hat{Y}_{149} = 0$. Παρόλλο που για τις πραγματικές τιμές της Y_t, για τις ημέρες που κάναμε πρόβλεψη, ισχύει $Y_{147} = 0, Y_{148} = 0, Y_{149} = 0$ δεν μπορούμε να ισχυριστούμε πως το μοντέλο μας παρέχει ασφαλείς προβλέψεις, αφου είναι αρκετά ευαίσθητό σε κοντινές παρελθούσες τιμές. Έτσι αν λάβει την τιμή '0' ακόμη και για ένα, σχετικά μικρό, αριθμό διαδοχικών ημερών, ενδεχομένως να μην μπορέσει να προβλέψει μια αμέσως επόμενη βροχή.

Ολοκληρώνοντας την ανάλυση μας για τα καλοκαίρια έχει ενδιαφέρον να εξετάσουμε την συμπεριφορά του μοντέλου MK97*step* στις υπόλοιπες καλοκαιρινές περιόδους, ακολουθώντας τα ίδια βήματα με εκείνα κατα την αξιολόγηση του υποδείγματος MX96.97*step*. Τα διαγνωστικά κριτήρια του δεδομένου υποδείγματος για όλες τις καλοκαιρινές ζώνες δίνονται στον ακόλουθο πίνακα.

Πίνακας 6.17 Διαγνωστικά κριτήρια από την προσαρμογή του MK97*step* για όλα τα καλοκαίρια

Καλοκαίρι	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
95	7	0.1328	174.082	126.391	139	140.391	161.276
96	7	0.0919	81.296	81.029	139	95.029	115.914
97	7	0.0381	35.001	35.577	139	47.577	68.462
98	7	0.0833	417.842	84.380	139	98.380	119.265
99	7	0.0810	72.924	67.461	129	81.461	101.849

Παρατηρώντας τον παραπάνω πίνακα σε συνδυασμό με τους 6.11-6.15, βλέπουμε ότι το μοντέλο MK97*step* δεν αλλάζει σημαντικά την προσαρμογή των υποδειγμάτων που προέκυψαν για τις άλλες καλοκαιρινές ζώνες (που ήταν MK95*step*, MK96*step*, MK98*step*, MK99*step*), μέσω της συνάρτησης *stepAIC*. Το γεγονός αυτο μας κινεί πάλι την υποψία , όπως έγινε και στους χειμώνες, μήπως είναι δυνατό να χρησιμοποιηθεί το λογιστικό μοντέλο με συμμεταβλητές $Y_{t-3}, T_t, T_{t-1}, H_t, H_{t-1}, H_{t-3}$ και για τις 5 καλοκαιρινές περιόδους, το οποίο θα δώσει κοινό $\boldsymbol{\beta}$. Παρατηρώντας τα διανύσματα των παραμέτρων του μοντέλου MK97*step* για όλα τα καλοκαίρια, που δίνονται στον πίνακα

Πίνακας 6.18 Εκτιμήσεις των παραμέτρων του MK97*step* σε όλα τα καλοκαίρια

Παράμετροι	95	96	97	98	99
$\hat{\beta}_0$	-0.522	-8.846	-7.170	-4.666	-15.732
$\hat{\beta}_1$ (Y_{t-3})	-0.434	1.142	3.248	-0.166	0.512
$\hat{\beta}_2$ (T_t)	-0.249	-0.231	-0.301	-0.088	0.055
$\hat{\beta}_3$ (T_{t-1})	-0.065	0.028	-0.293	-0.001	-0.044
$\hat{\beta}_4$ (H_t)	-0.016	0.051	0.159	0.163	0.084
$\hat{\beta}_5$ (H_{t-1})	0.118	0.225	0.233	-0.100	0.174
$\hat{\beta}_6$ (H_{t-3})	-0.034	-0.088	-0.1199	-0.004	-0.049

διαπιστώνουμε πως η αρχική μας υποψία δεν είναι και τόσο βάσιμη, αφού από καλοκαίρι σε καλοκαίρι το $\boldsymbol{\beta}$ διαφοροποιείται σημαντικά.

Στο σημείο αυτό είναι σημαντικό να αναφέρουμε ότι το υπόδειγμα MK97*step* παρόλλο που παρουσιάζει την καλύτερη προσαρμογή στα δεδομένα, δεν επιτρέπει την εννιαία εκτίμηση του $\boldsymbol{\beta}$, από όλα τα καλοκαίρια και συνάμα έχει σχετικά μεγάλη διάσταση. Το μοντέλο με την δεύτερη καλύτερη προσαρμογή και το οποίο όμως περιλαμβάνει την θερμοκρασία είναι το MK96*step* (T_t, H_{t-1}, H_{t-2}) που αναφέρεται στο καλοκαίρι του 1996. Παρατηρούμε ακόμη ότι το υπόδειγμα MK96*step* έχει ακριβώς

τις ίδιες συμμεταβλητές με το μοντέλο MK95*step* του καλοκαιριού του 1995. Βάσει αυτών των διαπιστώσεων θα εξετάσουμε κατα πόσο μπορεί το MK96*step* (ισοδύναμα το MK95*step*) να χρησιμοποιηθεί για κάθε θερινή περίοδο, οδηγώντας σε εννιαία εκτίμηση του $\boldsymbol{\beta}$ μέσω της ταυτόχρονης αξιοποίησης όλων των καλοκαιρινών χρονοσειρών. Για τα διαγνωστικά κριτήρια λοιπόν του δεδομένου μοντέλου για όλες τις καλοκαιρινές ζώνες έχουμε

Πίνακας 6.19 Διαγνωστικά κριτήρια από την προσαρμογή του MK96*step* για όλα τα καλοκαίρια

Καλοκαίρι	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
95	4	0.1323	154.718	122.564	142	130.564	122.498
96	4	0.0921	99.246	83.115	142	91.115	103.049
97	4	0.0467	117.963	47.904	142	55.904	67.838
98	4	0.111	134.985	106.345	142	114.345	126.279
99	4	0.0827	70.767	68.614	132	76.614	88.265

Βάσει του πίνακα 6.19 σε συνδυασμό με τους 6.11-6.15, βλέπουμε πως η προσαρμογή του λογιστικού μοντέλου με συμμεταβλητές T_t, H_{t-1}, H_{t-2} για τα καλοκαίρια 1997, 1998, 1999 διαφέρει σε πολύ μικρό βαθμό από την προσαρμογή που επεδεικνύουν αντιστοίχως τα υποδείγματα MK97*step*, MK98*step* και MK99*step*. Το γεγονός αυτό μας δείχνει ότι μπορούμε να χρησιμοποιήσουμε το μοντέλο (3.7) με διάνυσμα συμμεταβλητών $\mathbf{Z}_{t-1} = (T_t, H_{t-1}, H_{t-2})'$, για την μοντελοποίηση των θερινών βροχοπτώσεων στον νομό Ιωαννίνων, αξιοποιώντας και τις 5 καλοκαιρινές ζώνες που διαθέτουμε. Το δεδομένο υπόδειγμα θα παρέχει κοινό $\boldsymbol{\beta}$ για κάθε περίοδο. Πράγματι η δεδομένη θέση ενισχύεται και από τον παρακάτω πίνακα

Πίνακας 6.20 Εκτιμήσεις των παραμέτρων του MK96*step* σε όλα τα καλοκαίρια

Παράμετροι	95	96	97	98	99
$\hat{\beta}_0$	-1.911	-8.061	-7.651	0.325	-14.598
$\hat{\beta}_1$ (Y_{t-3})	-0.261	-0.240	-0.284	-0.189	-0.043
$\hat{\beta}_2$ (T_t)	0.153	0.289	0.250	0.008	0.286
$\hat{\beta}_3$ (T_{t-1})	-0.077	-0.096	-0.092	-0.001	-0.082

σύμφωνα με τον οποίο οι παράμετροι $\beta_j, j = 0, 1, 2, 3$ έχουν κοντινές τιμές για όλα τα καλοκαίρια (αν εξαιρέσουμε το καλοκαίρι του 1998). Όπως έχουμε ήδη αναφέρει το

$S-PLUS$ και το $SPSS$ δεν παρέχουν την δυνατότητα εννιαίας μοντελοποιήσης των καλοκαιρινών περιόδων.

6.6 Μοντελοποίηση των βροχοπτώσεων για ομαδοποιημένα δεδομένα

Από την ανάλυση που προηγήθηκε γίνεται φανερό πως η προσπάθεια επιλογής 'καλού' μοντέλου το οποίο θα περιγράφει επαρκώς τις βροχοπτώσεις, του νομού Ιωαννίνων, παρέχοντας ικανοποιητικές προβλέψεις είναι εξαιρετικά δύσκολη. Η πολυπλοκότητα του φαινομένου της βροχής το οποίο παρουσιάζει εξαιρετικές κυμάνσεις και μεταπτώσεις, όχι μόνο από καλοκαίρι σε χειμώνα αλλά και από μέρα σε μέρα της ίδιας εποχιακής ζώνης, έχει οδηγήσει τους μετεωρολόγους σε εναλλακτικές τεχνικές για την μοντελοποίηση του. Μια σημαντική στρατηγική είναι η ομαδοποίηση των ημερών σε δεκαπενθήμερα τα οποία πλέον αντιμετωπίζονται ως νέες μονάδες παρατήρησης, πάνω στις οποίες χτίζονται τα ήδη γνωστά μοντέλα. Τα νέα υποδείγματα απορροφώνας την μεταβλητότητα που παρουσιάζει η βροχή από ημέρα σε ημέρα παρέχουν ακριβέστερες προβλέψεις σε σχέση με τα αρχικά μοντέλα και συνηθίζονται στην πράξη. Αναμφισβήτητα η πληροφόρηση σχετικά με τις βροχοπτώσεις το επόμενο ή το μεθεπόμενο δεκαπενθήμερο κρίνεται σημαντική για περιοχές, όπως τα Ιωάννινα, οι οποίες στηρίζονται στην αγροτική οικονομία.

Με βάση τα προαναφερθέντα θα μετασχηματίσουμε τις αρχικές συνεχείς μετρήσεις που αναφέρονται στην ημερήσια βροχόπτωση σε μια δίτιμη χρονοσειρά $Y_t, t = 1, 2, \ldots$ ώστε η κάθε μέτρηση να αναφέρεται σε ένα δεκαπενθήμερο. Έστω λοιπόν ότι διαθέτουμε 1620 τιμές για την ποσότητα βροχής ανα ημέρα σε 'mm'. Με τον τρόπο αυτό θα προκύψει μια δίτιμη δειγματοληπτική διαδρομή, δεκαπενθήμερων, μήκους N=108. Η ανάθεση της τιμής '1' ή '0' στην Y_t για κάθε δεκαπενθήμερο θα γίνει σύμφωνα με τον ακόλουθο κανόνα: Αν για ένα δίαστημα 15 ημερών το άθροισμα των ποσοτήτων της βροχής από κάθε ημέρα είναι $\leq$ 5'mm' **και** για κάθε ημέρα από τις 15 η αντίστοιχη βροχόπτωση ήταν $\leq$ 1'mm' (ώστε να εξασφαλίσω ότι σε καμία μέρα από τις 15 δεν έβρεξε) τότε το αντίστοιχο Y_t θα πάρει την τιμή '0'. Γίνεται φανερό πως η τιμή '0' δηλώνει δεκαπενθήμερο στο οποίο δεν σημειώνονται βροχοπτώσεις. Σε αντίθετη περίπτωση από αυτή που μόλις αναφέραμε το Y_t θα λάβει την τιμή '1', και στο αντίστοιχο διάστημα των 15 ημερών θα σημειωθούν βροχές.

Για την στατιστική ανάλυση των δεδομένων μας θα στηριχτούμε στα αυτοπαλίν-

δρομα μοντέλα λογιστικής παλινδρόμησης που εισαγάγαμε στην περίπτωση της ενιαίας σειράς (N=1823) και τα οποία πλέον θα τα συμβολίσουμε ως εξής: M1F15D (Y_{t-1}), M2F15D (Y_{t-1}, Y_{t-2}), M3F15D (Y_{t-}, Y_{t-2}, Y_{t-3}), M4F15D (Y_{t-1}, Y_{t-3}). Για τα δεδομένα υποδείγματα προέκυψαν τα εξής.

Πίνακας 6.21 Διαγνωστικά κριτήρια για τα αυτοπαλίνδρομα μοντέλα στα δεκαπενθήμερα

Μοντέλο	p	MSE	$\mathcal{X}^2$	D	df	AIC	BIC
M1F15D	2	0.0926	107.145	70.348	106	74.348	79.890
M2F15D	3	0.0926	107.145	70.348	105	76.348	84.661
M3F15D	4	0.0896	95.153	66.574	104	74.574	85.656
M4F15D	3	0.0894	95.999	66.693	105	72.693	81.005

Με βάση τον παραπάνω πίνακα βλέπουμε πως την καλύτερη προσαρμογή στα δεδομένα την παρουσιάζει το μοντέλο M4F15D (Y_{t-1}, Y_{t-3}). Για την διεξαγωγή προβλέψεων θα στηριχτούμε στο δεδομένο αυτοπαλίνδρομο υπόδειγμα για το οποίο οι εκτιμήσεις των παραμέτρων του είναι

$$\hat{\boldsymbol{\beta}} = (\hat{\beta}_0, \hat{\beta}_1, \hat{\beta}_3)' = (7.895, 1.872, -7.489)',$$

με αντίστοιχο πίνακα διακυμάνσεων-συνδιακυμάνσεων

$$\mathbf{G}_N^{-1}(\hat{\boldsymbol{\beta}}) = \begin{bmatrix} 279.303 & -0.241 & -279.122 \\ -0.241 & 0.554 & -0.175 \\ -279.122 & -0.175 & 279.357 \end{bmatrix}.$$

Για την πρόβλεψη της δεσμευμένης πιθανότητας το 109^o δεκαπενθήμερο να είναι βροχερό, δοθέντος της ιστορίας του φαινομένου μέχρι το 108^o δεκαπενθήμερο (δηλαδή δοθέντος του $\mathcal{F}_{108}$) έχουμε

$$\hat{P}(Y_{109} = 1 \mid \mathcal{F}_{108}) = \frac{\exp(7.895 + 1.872 \times Y_{108} - 7.489 \times Y_{106})}{1 + \exp(7.895 + 1.872 \times Y_{108} - 7.489 \times Y_{106})}$$

$$= \frac{\exp(7.895 + 1.872 \times 1 - 7.489 \times 1)}{1 + \exp(7.895 + 1.872 \times 1 - 7.489 \times 1)} \Rightarrow$$

$$\hat{P}(Y_{109} = 1 \mid \mathcal{F}_{108}) = 0.907. \tag{6.5}$$

Σύμφωνα με την σχέση (6.5) επειδή η πιθανότητα το 109^o δεκαπενθήμερο να είναι βροχερό βρέθηκε αρκετά ψηλή συμπεραίνουμε ότι για το δεδομένο διάστημα θα ισχύει

$\hat{Y}_{109} = 1$. Το συγκεκριμένο αποτέλεσμα παρόλλο που συμφωνεί με το πραγματικό Y_{109}, δεν μας εξασφαλίζει ότι το μοντέλο μας διαθέτει ικανοποιητική προβλεψιμότητα. Η δεδομένη πρόβλεψη μπορούμε να πούμε ότι ήταν αναμενόμενη αν σκεφτούμε ότι το μοντέλο μας λαμβάνει υπόψη τις υστερήσεις $1^{ης}$ και $3^{ης}$ τάξης της Y_t, με $Y_{108} = 1$ και $Y_{106} = 1$. Γίνεται λοιπόν φανερό πως ιδιαίτερο ενδιφέρον παρουσιάζει επιπλέον η πρόβλεψη των πιθανοτήτων

$$P(Y_t = 1 \mid Y_{t-1} = 0, Y_{t-3} = 0),$$

$$P(Y_t = 1 \mid Y_{t-1} = 1, Y_{t-3} = 0),$$

$$P(Y_t = 1 \mid Y_{t-1} = 0, Y_{t-3} = 1).$$

Για τον υπολογισμό των προαναφερθέντων δεσμευμένων πιθανοτήτων θα χρησιμοποιήσουμε κατάλληλες θέσεις της δειγματοληπτικής διαδρομής, εξακολουθώντας να αξιοποιούμε το προσαρμοσμένο μοντέλο M4F15D που προέκυψε από την αρχική χρονοσειρά μήκους N=108. Έτσι θα έχουμε

- για την $P(Y_t = 1 \mid Y_{t-1} = 0, Y_{t-3} = 0)$

$$\hat{P}(Y_{85} = 1 \mid Y_{84} = 0, Y_{82} = 0) = \frac{\exp(7.895 + 1.872 \times 0 - 7.489 \times 0)}{1 + \exp(7.895 + 1.872 \times 0 - 7.489 \times 0)} \Rightarrow$$
$$\hat{P}(Y_{85} = 1 \mid \mathcal{F}_{84}) = 0.999.$$

- για την $P(Y_t = 1 \mid Y_{t-1} = 1, Y_{t-3} = 0)$

$$\hat{P}(Y_{113} = 1 \mid Y_{112} = 1, Y_{110} = 0) = \frac{\exp(7.895 + 1.872 \times 1 - 7.489 \times 0)}{1 + \exp(7.895 + 1.872 \times 1 - 7.489 \times 0)} \Rightarrow$$
$$\hat{P}(Y_{113} = 1 \mid \mathcal{F}_{112}) = 0.999.$$

- για την $P(Y_t = 1 \mid Y_{t-1} = 0, Y_{t-3} = 1)$

$$\hat{P}(Y_{111} = 1 \mid Y_{110} = 0, Y_{108} = 1) = \frac{\exp(7.895 + 1.872 \times 0 - 7.489 \times 1)}{1 + \exp(7.895 + 1.872 \times 0 - 7.489 \times 1)} \Rightarrow$$
$$\hat{P}(Y_{111} = 1 \mid \mathcal{F}_{111}) = 0.6.$$

Παρατηρώντας τις παραπάνω πιθανότητες συμπεραίνουμε ότι αν ένα δεκαπενθήμερο είναι βροχερό τότε είναι πολύ πιθανό και στο αμέσως επόμενο να συμβούν βροχοπτώσεις, ακόμη και αν στο προπροηγούμενο δεν σημειώθηκαν βροχές. Αυτή η πιθανότητα

μειώνεται, παραμένοντας όμως μεγαλύτερη του 0.5, όταν το προηγούμενο δεκαπενθήμερο δεν έβρεξε παρόλλο που το προπροηγούμενο ήταν βροχερό. Σύμφωνα με τα προαναφερθέντα το μοντέλο μας, όταν το προηγούμενο ή το προπροηγούμενο δεκαπενθήμερο είναι βροχερό, έχει την τάση να δώσει ένδειξη βροχής για το αμέσως επόμενο διάστημα. Ενδεικτικά στηριζόμενοι στην τιμή $\hat{Y}_{109} = 1$ για την Y_{110} θα έχουμε

$$\hat{P}(Y_{110} = 1 \mid \hat{Y}_{109} = 1, Y_{107} = 1) = \frac{\exp(7.895 + 1.872 \times 1 - 7.489 \times 1)}{1 + \exp(7.895 + 1.872 \times 1 - 7.489 \times 1)} \Rightarrow$$

$$\hat{P}(Y_{110} = 1 \mid \mathcal{F}_{110}) = 0.907.$$

Δηλαδή μπορούμε να πούμε ότι $\hat{Y}_{110} = 1$, ενώ ισχύει $Y_{110} = 0$. Διαπιστώνουμε λοιπόν ότι οι δεσμευμένες πιθανότητες βροχής δεκαπενθημέρου είναι όλες αρκετά υψηλές (> 0.5), αδυνατώντας να προβλέψει ξηρό δεκαπενθήμερο.

Η προαναφερθείσα αδυναμία του υποδείγματος M4F15D δεν έχει να κάνει με την παράλειψη κάποιας σημαντικής ερμηνευτικής μεταβλητής, κατά την διαδικασία μοντελοποίησης, αλλά με την φύση των δεδομένων. Πράγματι κοιτάζοντας προσεκτικότερα την δειγματοληπτική διαδρομή των δεκαπενθημέρων, σε συνδυασμό με την αρχική χρονοσειρά των ημερήσιων βροχοπτώσεων (N=1823), θα δούμε πως στην πλειοψηφία των περιπτώσεων είχαμε βροχοπτώσεις, κάτι το οποίο είναι σύνηθες φαινόμενο για τον νομό Ιωαννίνων. Αναλυτικότερα για την πραγματοποίηση των N=108 δεκαπενθημέρων προέκυψε ο ακόλουθος πίνακας συχνοτήτων για τα ενδεχόμενα, 'βροχερό δεκαπενθήμερο' ($Y_t = 1$) και 'δεκαπενθήμερο χωρίς βροχοπτώσεις' ($Y_t = 0$)

$(Y_t = 1)$:	96
$(Y_t = 0)$:	12

Στο σημείο αυτό μπορούμε να εξηγήσουμε αναλυτικότερα γιατί το μοντέλο M4F15D (Y_{t-1}, Y_{t-3}) δεν μπορεί να προβλέψει τα μη βροχερά διαστήματα των 15 ημερών. Ορίζουμε τα ακόλουθα ενδεχόμενα:

$$A_{00} = \{Y_t = 0 \mid Y_{t-1} = 0, Y_{t-3} = 0\}$$
$$A_{01} = \{Y_t = 0 \mid Y_{t-1} = 0, Y_{t-3} = 1\}$$
$$A_{10} = \{Y_t = 0 \mid Y_{t-1} = 1, Y_{t-3} = 0\}$$
$$A_{11} = \{Y_t = 0 \mid Y_{t-1} = 1, Y_{t-3} = 1\}.$$

Μέσω της δειγματοληπτικής διαδρομής λαμβάνουμε τις συχνότητες των παραπάνω ενδεχομένων οι οποίες αντιστοίχως είναι

$$n_{00} = 0, n_{01} = 4, n_{10} = 0, n_{11} = 8.$$

Έτσι σύμφωνα με τον Νόμο των Μεγάλων Αριθμών, οι πιθανότητες

$$P(A_{00}), P(A_{01}), P(A_{10}), P(A_{11})$$

θα είναι πολύ μικρές, με αποτέλεσμα οι πιθανότητες

$$1 - P(A_{00}), 1 - P(A_{01}), 1 - P(A_{10}), 1 - P(A_{11}),$$

που παρέχει το υπόδειγμα M4F15D, να είναι αρκετά υψηλές.

6.7 Συμπεράσματα

Γενικά, σύμφωνα με την ανάλυση που προηγήθηκε, γίνεται φανερό πως είναι εξαιρετικά δύσκολο να μοντελοποιήσουμε τις βροχοπτώσεις στηριζόμενοι σε αυτοπαλίνδρομα λογιστικά υποδείγματα. από την άλλη τα λογιστικά μοντέλα που περιέχουν τυχαίες χρονοεξαρτώμενες συμμεταβλητές όπως είναι η ατμοσφαιρική πίεση, η θερμοκρασία και η υγρασία παρόλλο που έχουν καλύτερη προσαρμογή στα δεδομένα μπορούν να αξιοποιηθούν άμεσα μόνο για την πρόβλεψη της αμέσως επόμενης περιόδου. Για πιο μακρινές προβλέψεις μέσω των δεδομένων υποδειγμάτων θα πρέπει πρώτα να προβλέψουμε τις μελλοντικές τιμές των συνεχών μεταβλητών T_t, H_t, A_t χρησιμοποιώντας μεθόδους της φασματικής ανάλυσης. Ο ενδιαφερόμενος αναγνώστης για περισσότερες πληροφορίες σχετικά με την μοντελοποίηση των βροχοπτώσεων καθώς και την αναπαραγωγή της μεταβλητότητας που επιδεικνύουν γενικά τα κλιματολογικά φαινόμενα, μέσω των GLM, παραπέμπεται στους *Chandler* και *Wheater* (1998) καθώς και στον *Chandler* (2003).

Κεφάλαιο 7

Σύγχρονες Προσεγγίσεις για την Στατιστική Συμπερασματολογία των Κατηγορικών Χρονοσειρών

7.1 Εισαγωγή

Η μοντελοποίηση των κατηγορικών χρονοσειρών μέσω μοντέλων παλινδρόμησης που στηρίζονται στα GLM και την PL αναμφισβήτητα είναι αρκετά διαδεδομένη, αφού κατόρθωσε να ξεπεράσει αρκετές δυσκολίες που επιδεικνύουν τα Μαρκοβιανά μοντέλα. Η ανάπτυξη όμως και η πληρέστερη θεμελίωση άλλων στατιστικών πεδίων, ιδιαίτερα τα τελευταία 20 χρόνια, συντέλεσαν στην δημιουργία νέων αποτελεσματικών τεχνικών για την στατιστική ανάλυση των ποιοτικών σειρών.

Στο κεφάλαιο αυτό θα παρουσιάσουμε αρχικά τα λεγόμενα μοντέλα του Χώρου-Καταστάσεων (*State Space Models*). Εν συνεχεία θα αναφερθούμε στα Μπευζιανά ημιπαραμετρικά μοντέλα παλινδρόμησης (*Bayesian semiparametric regression models*) τα οποία χρησιμοποιούνται για την μοντελοποίηση πολυκατηγορικών *time-space* δεδομένων. Παράλληλα θα δώσουμε κάποιες γενικές πληροφορίες για την μέθοδο *Bootstrap* για στάσιμες κατηγορικές χρονοσειρές. Τέλος στα πλαίσια των σύγχρονων προσεγγίσεων στις ποιοτικές χρονολογικές σειρές θα αναφερθούμε στην κυματοειδή (*wavelet*) ανάλυση των κατηγορικών χρονοσειρών.

7.2 Παρουσίαση των Νεων Μεθόδων

7.2.1 *State Space* ανάλυση των κατηγορικών χρονοσειρών

Τα μοντέλα του Χώρου Καταστάσεων ξεκίνησαν πρωτοεμφανίστηκαν στη στατιστική βιβλιογραφία τις δεκαετίες του 1960 και 1970 μέσω της δουλειάς ερευνητών που ενδιαφέρονταν για την πρόβλεψη, και ειδικά την μπευζιανή πρόβλεψη, μη στάσιμων διαδικασιών (*Harrison and Stevens* (1976)), (*West and Harrison* (1997)). Παράλληλα η ανάπτυξη αυτού του στατιστικού πεδίου συντελέστηκε επειδή τα δεδομένα μοντέλα βασίζονται σε επαναληπτικές σχέσεις πυκνοτήτων πιθανότητας των οποίων τα ολοκληρώματα είναι ιδιαίτερα χρήσιμα για μη κανονικές χρονοσειρές. Το γεγονός αυτό οδήγησε στην χρήση των μοντέλων του χώρου καταστάσεων για την στατιστική συμπερασματολογία των κατηγορικών σειρών.

Σύμφωνα με την δεδομένη μεθοδολογική προσέγγιση υποθέτουμε ότι η παρατηρούμενη χρονοσειρά σχετίζεται με μια ή περισσότερες σειρές των οποίων τα επίπεδα δεν μπορούν να παρατηρηθούν. Αυτές οι 'κρυφές' ('*latent*') στοχαστικές διαδικασίες καθορίζουν τα δυναμικά στοιχεία του συστήματος (δηλαδή την πορεία του συστήματος μέσα στον χρόνο) μέσω ενός μοντέλου παρατήρησης. Το δυναμικό μέρος αναπαρίσταται μέσω μιας εξίσωσης μετάβασης. Οι εξισώσεις παρατήρησης και μετάβασης μαζί συνιστούν το κανονικό μοντέλο του Χώρου Καταστάσεων. Οι εκτιμήσεις των μη παρατηρούμενων επιπέδων μπορούν να ληφθούν από την εφαρμογή τεχνικών 'φιλτραρίσματος' ('*filtering*') και εξομάλυνσης ('*smoothing*') στην παρατηρούμενη σειρά (*Rijn and Molenaar* (2003)). Πιο συγκεκριμένα για την ανάλυση των κατηγορικών χρονοσειρών χρησιμοποιείται η μέθοδος που ανέπτυξε ο *Fahrmeir* (1992*a*, 1992*b*) η οποία κρίνεται καταλληλότερη εκείνη που πρότειναν οι *Durbin* και *Koopman* (2001). Σημαντικό εργαλείο στην προσπάθεια μοντελοποίησης είναι το γενικευμένο επεκταμένο φίλτρο του *Kalman* το οποίο αξιοποιείται κυρίως στην περίπτωση πολυμεταβλητών ποιοτικών σειρών.

7.2.2 Μπευζιανή ημιπαραμετρική ανάλυση παλινδρόμησης πολυκατηγορικών *time-space* δεδομένων

Η δεδομένη μέθοδος στηρίζεται στα τυχαία Μαρκοβιανά πεδία και επιδιώκει την ανάλυση της εξάρτησης των πολυκατηγορικών αποκριτικών μεταβλητών ως προς τον χώρο, τον χρόνο και άλλες συμμεταβλητές. Το γενικό μοντέλο επεκτείνει το δυναμικό ή το μοντέλο του χώρου καταστάσεων για κατηγορικές χρονοσειρές, περιλαμβάνοντας

χωρικές επιδράσεις (*spatial effects*) καθώς και μη γραμμικές επιδράσεις μετρικών συμμεταβλητών, σύμφωνα με ευέλικτες ημιπαραμετρικές μορφές. Η τάση και οι εποχικές συνιστώσες καθώς και οι χωρικές επιδράσεις αντιμετωπίζονται μέσω της ανάθεσης κατάλληλων *priors* που έχουν διαφορετική μορφή ή βαθμούς εξομάλυνσης. Η συμπερασματολογία είναι πλήρως Μπευζιανή και χρησιμοποιεί *Monte Carlo Markov Chain* ($MCMC$) τεχνικές για την *posterior* ανάλυση (*Fahrmeir and Lang* (2000)).

7.2.3 *Sieve Bootstrap with Variable Length Marcov Chains for Stationary Categorical Times Series*

Για την στατιστική συμπερασματολογία των κατηγορικών χρονοσειρών τα τελευταία χρόνια χρησιμοποιείται η μέθοδος *Bootstrap*, κατάλληλα τροποποιημένη. Αναλυτικότερα η δεδομένη προσέγγιση στηρίζεται στην μέθοδο του 'κοσκινίσματος' ('*sieve*'). Η στοχαστική ανέλιξη που 'γεννά' τα δεδομένα προσεγγίζεται από την λεγόμενη '*Variable Length Markov Chain*' ($VLMC$), που είναι μια ευέλικτη αλλα και 'οικονομική', από πλευράς παραμέτρων, κλάση Μαρκοβιανών μοντέλων. Η επαναδειγματοληψία πραγματοποιείται μέσω προσομοίωσης στο προσαρμοσμένο μοντέλο. Για τα πλεονεκτήματα της δεδομένου μεθόδου, καθώς και για περισσότερες πληροφορίες, ο ενδιαφερόμενος αναγνώστης παραπέμπεται στον *Buhlmann* (1999).

7.2.4 Ανάλυση των Κατηγορικών Χρονοσειρών μέσω της κυματοειδούς (*wavelet*) ανάλυσης

Για την μοντελοποίηση δεδομένων που προέρχονται από Μαρκοβιανές αλυσίδες σχετικά πρόσφατα έχουν αναπτυχθεί κυματοειδείς τεχνικές. Η δεδομένη προσέγγιση επιτρέποντας την παρουσία συμμεταβλητών ουσιαστικά αποτελεί επέκταση του γενικού μοντέλου (4.11). Αναλυτικότερα οι πιθανότητες μετάβασης, μιας μη στάσιμης Μαρκοβιανής αλυσίδας εκφράζονται μέσω των αναμενόμενων τιμών υποδειγμάτων που περιλαμβανουν κυματοειδείς εκφράσεις και εν συνεχεία, δοθέντος της δειγματοληπτικής διαδρομής, εκτιμώνται οι πιθανότητες αλλα και οι συντελεστές των προαναφερθέντων (κυματοειδών) όρων. Μέσω κατάλληλης επιλογής του αριθμού αλλα και της μορφής των κυματοειδών παραγόντων η συγκεκριμένη τεχνική παρέχει ελκυστική μοντελοποίηση διακριτών σημάτων στην μη στάσιμη περίπτωση. Επιπρόσθετα η μέθοδος είναι χρήσιμη για την ανίχνευση αιφνίδιων ή σταθερών αλλαγών στην δομή και τάξη των Μαρκοβιανών αλυσίδων (*Brillinger, Morettin, Irizarry and Chiann* (2000)).

Παράρτημα Α

GLM και Εξαρτημένα Δεδομένα

Α.1 Στοιχεία της θεωρίας των GLM για την Ε.Ο.Κ.

Στην δεδομένη ενότητα παραθέτουμε κάποια γενικά αποτελέσματα της θεωρίας των GLM, τα οποία όμως είναι τροποποιημένα ώστε να αξιοποιούν την πληροφορία του παρελθόντος. Έτσι έχουμε

(i) Τυχαία Συνιστώσα (*Random Component*). Τα δεδομένα μοντέλα αναφέρονται σε τυχαίες μεταβλητές των οποίων η δεσμευμένη κατανομή δοθέντος του παρελθόντος ανήκει στην εκθετική οικογένεια κατανομών (Ε.Ο.Κ) (*exponential family of distributions*) . Έτσι για την δεσμευμένη κατανομή της Y_t δοθέντος του παρελθόντος ($\mathcal{F}_{t-1}$) για κάθε χρονική στιγμή $t = 1, 2, \ldots, N$ θα ισχύει

$$f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}) = exp\Big\{\frac{y_t\theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t; \phi)\Big\} \tag{Α.1}$$

όπου $\alpha_t(\phi) = \phi/\omega_t$, με ϕ να είναι η παράμετρος *διασποράς* και ω_t είναι μια γνωστή παράμετρος που ονομάζεται *βάρος* (*weight*). Η παράμετρος θ_t ονομάζεται *φυσική* (*natural*) παράμετρος της κατανομής.

(ii) Συστηματική Συνιστώσα (*Systematic Component*). Για κάθε $t = 1, 2, \ldots, N$ όπως προαναφέρθηκε υπάρχει η συνάρτηση g για την οποία ισχύει η σχέση (3.1). Η συνάρτηση g είναι γνωστή και ονομάζεται *συνάρτηση σύνδεσης* (*link function*). Ο παράγοντας η_t καλείται γραμμική πρόβλεψη (*linear predictor*) αφού η εκτίμηση του δεδομένου γραμμικού συνδυασμού για κάθε χρονική στιγμή παρέχει την εκτίμηση της δεσμευμένης μέσης τιμής $\mu_t(\boldsymbol{\beta})$.

Αφού η $Y_t \mid \mathcal{F}_{t-1} \in E.O.K$, θα ισχύουν οι ακόλουθες γνωστές σχέσεις που συνδέουν τις δεσμευμένες ροπές για κάθε χρονική στιγμή με την φυσική παράμετρο θ_t,

$$\mu_t = E[Y_t \mid \mathcal{F}_{t-1}] = b'(\theta_t) \tag{A.2}$$

και

$$\sigma_t^2 = Var[Y_t \mid \mathcal{F}_{t-1}] = \alpha_t(\phi) b''(\theta_t) \tag{A.3}$$

όπου $V(\mu_t) = b''(\theta_t)$ ονομάζεται *συνάρτηση διακύμανσης* (*variance function*). Αφού η διακύμανση είναι πάντα θετική συμπεραίνουμε απο την σχέση (A.3) ότι $b''(\theta_t) > 0$ και επομένως η b' είναι μονότονη άρα και αντιστρέψιμη . Επομένως απο την σχέση (A.2) προκύπτει ότι

$$\theta_t = (b')^{-1}(\mu_t). \tag{A.4}$$

Γίνεται λοιπόν φανερό ότι η φυσική παράμετρος θ_t είναι μονότονη συνάρτηση του μ_t [1] και επομένως μπορεί να χρησιμοποιηθεί ως συνάρτηση σύνδεσης. Η συνάρτηση σύνδεσης g για την οποία ισχύει

$$g(\mu_t) = \theta_t(\mu_t) = \eta_t = \boldsymbol{\beta}' \boldsymbol{Z}_{t-1} \tag{A.5}$$

ονομάζεται *κανονικός σύνδεσμος* (*canonical link*). Προφανώς για την δεδομένη συνάρτηση σύνδεσης θα ισχύει ότι

$$g = \mu^{-1} \equiv (b')^{-1}. \tag{A.6}$$

Τέλος, αξίζει να αναφέρουμε ότι για οποιαδήποτε συνάρτηση σύνδεσης ισχύει

$$\theta_t = (b')^{-1}(g^{-1}(\eta_t)) = \mu^{-1}(g^{-1}(\eta_t)) = (g \circ \mu)^{-1}(\eta_t). \tag{A.7}$$

Παρατήρηση: Σύμφωνα με τις σχέσεις (A.2) και (A.3) παρατηρούμε ότι για κάθε χρονική στιγμή η κατάσταση του φαινομένου προσδιορίζεται απο την τιμή της φυσικής παραμέτρου. Βάση όμως της σχέσης (A.7) (σε συνδυασμό με την (A.5)) η εκτίμηση της παραμέτρου θ_t, για κάθε χρονική στιγμή, ανάγεται στην εκτίμηση της γραμμικής πρόβλεψης και ουσιαστικά στην εκτίμηση του σταθερού παραμετρικού διανύσματος $\boldsymbol{\beta}$. Γίνεται λοιπόν φανερό μέσα απο μια διαφορετική οπτική γωνία, με πολύ φυσικό τρόπο, γιατί επιδιώκουμε την καλύτερη δυνατή προσαρμογή του μοντέλου (3.1). Η κατάλληλη επιλογή του διανύσματος των συμμεταβλητών και οι αντίστοιχη εκτίμηση του σχετικού $\boldsymbol{\beta}$ θα προσφέρουν την καλύτερη δυνατή γνώση του μηχανισμού τύχης ($f(y_t \mid \mathcal{F}_{t-1})$) που το διέπει για κάθε χρονική στιγμή και επομένως θα μειώσει την αβεβαιότητα που μας διακατέχει για αυτό.

[1] η αντίστροφη κάθε μονότονης συνάρτησης είναι επίσης μονότονη

Α.2 Συμπερασματολογία Μερικής Πιθανοφάνειας

Στην ενότητα Α'.1 δώσαμε τα βασικά αποτελέσματα της θεωρίας των *GLM*, κατάλληλα τροποποιημένα ώστε να λαμβάνουν υπόψη την διαχρονική εξάρτηση. Έτσι στην δεδομένη παράγραφο θα παρουσιάσουμε την διαδικασία εύρεσης των εκτιμητών των παραμέτρων του μοντέλου (3.1), σύμφωνα με την θεωρία της μερικής πιθανοφάνειας. Η προσέγγιση αυτή αναφέρεται τόσο σε ποσοτικά όσο και σε ποιοτικά εξαρτημένα δεδομένα.

Έστω η χρονοσειρά $\{Y_t\}, t = 1, 2, \ldots, N$ της οποίας η δεσμευμένη κατανομή δοθέντος της ιστορίας $\mathcal{F}_{t-1}$ ανήκει στην E.O.K. Έτσι για κάθε χρονική στιγμή t για την δεσμευμένη συνάρτηση πυκνότητα πιθανότητας (ή συνάρτηση πιθανότητας) της Y_t θα ισχύει η σχέση (Α.1). Θυμίζουμε ότι το $\{\mathbf{Z}_{t-1}\}$ παριστά το p-διάστατο διάνυσμα των τυχαίων χρονοεξαρτώμενων συμμεταβλητών, η g είναι η συνάρτηση σύνδεσης, και ϕ η παράμετρος διασποράς που την θεωρούμε γνωστή.

Για τη συνάρτηση μερικής πιθανοφάνειας της δειγματοληπτικής διαδρομής ισχύει

$$PL(\boldsymbol{\beta}) = \prod_{t=1}^{N} f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}).$$

Έτσι σύμφωνα με την σχέση (Α.1) ο λογάριθμος της μερικής πιθανοφάνειας γράφεται

$$\ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \log f(y_t; \theta_t, \phi \mid \mathcal{F}_{t-1}) = \sum_{t=1}^{N} \Big\{ \frac{y_t\theta_t - b(\theta_t)}{\alpha_t(\phi)} + c(y_t, \phi) \Big\}$$

$$= \sum_{t=1}^{N} \Big\{ \frac{y_t u(\mathbf{z}'_{t-1}\boldsymbol{\beta}) - b(u(\mathbf{z}'_{t-1}\boldsymbol{\beta}))}{\alpha_t(\phi)} + c(y_t, \phi) \Big\} \equiv \sum_{t=1}^{N} \ell_t. \tag{Α.8}$$

Για να δηλώσουμε την εξάρτηση της μερικής πιθανοφάνειας απο το $\boldsymbol{\beta}$ αξιοποιήσαμε το γεγονός ότι η φυσική παράμετρος γράφεται ως σύνθεση των συναρτήσεων g^{-1} και b^{-1} με ανεξάρτητη μεταβλητή την γραμμική πρόβλεψη $\eta_t = \mathbf{z}'_{t-1}\boldsymbol{\beta}$. Την νέα συνάρτηση $(g \circ b')^{-1}(\cdot)$ την συμβολίσαμε με $u(\cdot)$ και επομένως σύμφωνα με τον τύπο (Α.7) θα ισχύει $\theta_t = u(\mathbf{z}'_{t-1}\boldsymbol{\beta})$.

Το *μερικό σκόρ(partial score)* γενικά ορίζεται ως $\nabla\ell(\boldsymbol{\beta}) = (\frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_1}, \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_2}, \ldots, \frac{\partial\ell(\boldsymbol{\beta})}{\partial\beta_p})'$. Για τον υπολογισμό των συνιστωσών του δεδομένου p-διάστατου διανύσματος, σύμφωνα με την σχέση (Α.8), απαιτείται πρώτα να υπολογίσουμε τις ποσότητες $\frac{\partial\ell_t(\boldsymbol{\beta})}{\partial\beta_j}$ για $j = 1, 2, \ldots, p$. Για το σκοπό αυτό χρησιμοποιούμε τον ακόλουθο κανόνα αλυσίδας

$$\frac{\partial\ell_t(\boldsymbol{\beta})}{\partial\beta_j} = \frac{\partial\ell_t}{\partial\theta_t}\frac{\partial\theta_t}{\partial\mu_t}\frac{\partial\mu_t}{\partial\eta_t}\frac{\partial\eta_t}{\partial\beta_j}. \tag{Α.9}$$

Σύμφωνα με τους τύπους (Α.2) και (Α.3) θα έχουμε

$$\frac{\partial \ell_t}{\partial \theta_t} = \frac{(y_t - b'(\theta_t))}{\alpha_t(\phi))} = \frac{(y_t - \mu_t)}{\alpha_t(\phi)} \tag{A.10}$$

και

$$\frac{\partial \theta_t}{\partial \mu_t} = \frac{1}{b''(\theta_t)} = \frac{\alpha_t(\phi)}{Var[Y_t \mid \mathcal{F}_{t-1}]}. \tag{A.11}$$

Ακόμη επειδή $\eta_t = \sum_{j=1}^{p} z_{(t-1)j}\beta_j$ θα ισχύει

$$\frac{\partial \eta_t}{\partial \beta_j} = z_{(t-1)j}. \tag{A.12}$$

Έτσι η εξίσωση (Α.9) γράφεται

$$\frac{\partial \ell_t}{\partial \beta_j} = \frac{(y_t - \mu_t)}{Var[Y_t \mid \mathcal{F}_{t-1}]} \frac{\partial \mu_t}{\partial \eta_t} z_{(t-1)j}$$

για $j = 1, 2, \ldots, p$. Βάσει των προαναφερθέντων, οι εξισώσεις μερικής πιθανοφάνειας είναι

$$\mathbf{S}_N(\boldsymbol{\beta}) = \nabla \ell(\boldsymbol{\beta}) = \mathbf{0}, \tag{A.13}$$

όπου

$$\mathbf{S}_N(\boldsymbol{\beta}) \equiv \nabla \ell(\boldsymbol{\beta}) = \sum_{t=1}^{N} \mathbf{Z}_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{(Y_t - \mu_t(\boldsymbol{\beta}))}{\sigma_t^2(\boldsymbol{\beta})} \tag{A.14}$$

με $\sigma_t^2(\boldsymbol{\beta}) = Var[Y_t \mid \mathcal{F}_{t-1}]$.

Η *διανυσματική στοχαστική ανέλιξη των μερικών σκόρ(partial score vector process)* $\{\mathbf{S}_t(\boldsymbol{\beta})\}, t = 1, 2, \ldots, N$, ορίζεται απο τα μερικά αθροίσματα

$$\mathbf{S}_t(\boldsymbol{\beta}) = \sum_{s=1}^{t} \mathbf{Z}_{s-1} \frac{\partial \mu_s}{\partial \eta_s} \frac{(Y_s - \mu_s(\boldsymbol{\beta}))}{\sigma_s^2(\boldsymbol{\beta})}. \tag{A.15}$$

Το διάνυσμα $\nabla \ell(\boldsymbol{\beta})$ είναι το αντίστοιχο του διανύσματος των σκορ, $\mathbf{U}(\boldsymbol{\beta}) = \frac{\partial \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}}$, που συναντάμε στα κλασσικά γενικευμένα γραμμικά μοντέλα μόνο που τώρα συμβολίζεται με $\mathbf{S}_N(\boldsymbol{\beta})$ και ονομάζεται *διάνυσμα των μερικών σκορ* (*partial score's vector*).

Ακόμη επειδή ισχύει ότι

$$E\Big[\mathbf{Z}_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{(Y_t - \mu_t(\boldsymbol{\beta}))}{\sigma_t^2(\boldsymbol{\beta})} \mid \mathcal{F}_{t-1}\Big] = 0$$

συνεπάγεται ότι $E[\mathbf{S}_N(\boldsymbol{\beta})] = \mathbf{0}$. Με όμοιο τρόπο αποδεικνύεται ότι

$$E\Big[\mathbf{Z}_{s-1} \frac{\partial \mu_s}{\partial \eta_s} \frac{(Y_s - \mu_s(\boldsymbol{\beta}))}{\sigma_s^2(\boldsymbol{\beta})} \mathbf{Z}'_{t-1} \frac{\partial \mu_t}{\partial \eta_t} \frac{(Y_t - \mu_t(\boldsymbol{\beta}))}{\sigma_t^2(\boldsymbol{\beta})}\Big] = \mathbf{0}, \quad s < t.$$

Η λύση του συστήματος των εξισώσεων των εξισώσεων μερικής πιθανοφάνειας (A.13) δηλώνεται με $\hat{\boldsymbol{\beta}}$ και αποτελεί τον εκτιμητή μέγιστης μερικής πιθανοφάνειας του $\boldsymbol{\beta}$. Το σύστημα των εξισώσεων (A.13) είναι μη γραμμικό και συνήθως επιλύεται μέσω της επαναληπτικής διαδικασίας *Fisher scoring*. Με σκοπό να εξηγηθεί πως λειτουργεί ο δεδομένος αλγόριθμος στα πλαίσια του μη τυχαίου πειράματος θα εισάγουμε, για την μονομεταβλητή E.O.K, ορισμένους σημαντικούς πίνακες. Οι δεδομένοι πίνακες θα αξιοποιηθούν, πέρα των κεφαλαίων 3, 4, 5 και στο παράρτημα B.

Σημαντικό ρόλο λοιπόν, για την συμπερασματολογία σύμφωνα με την μερική πιθανοφάνεια διαδραματίζει ο *αθροιστικός κατα συνθήκη πίνακας πληροφορίας (cumulative conditional information matrix)*, $\mathbf{G}_N(\boldsymbol{\beta})$, που ορίζεται ως ακολούθως

$$\begin{aligned}\mathbf{G}_N(\boldsymbol{\beta}) &= \sum_{t=1}^{N} Cov\Big[\mathbf{Z}_{t-1}\frac{\partial \mu_t}{\partial \eta_t}\frac{(Y_t-\mu_t(\boldsymbol{\beta}))}{\sigma_t^2(\boldsymbol{\beta})} \mid \mathcal{F}_{t-1}\Big] \\ &= \sum_{t=1}^{N} \mathbf{Z}_{t-1}\Big(\frac{\partial \mu_t}{\partial \eta_t}\Big)^2\frac{1}{\sigma_t^2(\boldsymbol{\beta})}\mathbf{Z}_{t-1}' \\ &\mathbf{Z}'\mathbf{W}(\boldsymbol{\beta})\mathbf{Z}, \qquad (A.16)\end{aligned}$$

με

$$\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_0' \\ \mathbf{Z}_1' \\ \vdots \\ \mathbf{Z}_{N-1}' \end{bmatrix}$$

ένας $N \times p$ πίνακας και $\mathbf{W}(\boldsymbol{\beta}) = diag(w_1, w_2, \ldots, w_n)$ όπου

$$w_t = \Big(\frac{\partial \mu_t}{\partial \eta_t}\Big)^2\frac{1}{\sigma_t^2(\boldsymbol{\beta})}, \quad t = 1, 2, \ldots, N. \qquad (A.17)$$

Ακόμη, ορίζεται ο *αδέσμευτος πίνακας πληροφορίας (unconditional information matrix)* απο την σχέση

$$Cov(\mathbf{S}_N(\boldsymbol{\beta})) = \mathbf{F}_N(\boldsymbol{\beta}) = E[\mathbf{G}_N(\boldsymbol{\beta})]. \qquad (A.18)$$

Τέλος ορίζεται ο *παρατηρούμενος πίνακας πληροφορίας (observed information matrix)* $\mathbf{H}_N(\boldsymbol{\beta})$, για τον οποίο ισχύει

$$H_N(\boldsymbol{\beta}) \equiv -\nabla\nabla'\ell(\boldsymbol{\beta}).$$

Για τον συγκεκριμένο πίνακα αποδεικνύεται γενικά ο τύπος

$$\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta}) - \mathbf{R}_N(\boldsymbol{\beta}), \qquad (A.19)$$

όπου

$$\mathbf{R}_N(\boldsymbol{\beta}) = \frac{1}{\alpha_t(\phi)} \sum_{t=1}^{N} \mathbf{Z}_{t-1} d_t(\boldsymbol{\beta}) \mathbf{Z}'_{t-1} (Y_t - \mu_t(\boldsymbol{\beta})) \tag{A.20}$$

και $d_t(\boldsymbol{\beta}) = [\partial^2 u(\eta_t)/\partial \eta_t^2]$. Για την απόδειξη της σχέσης (A.20) βλέπε *Fokiano* και *Kedem* (2002) σελίδα 13.

Α.3 Ασυμπτωτική Θεωρία

Ο *MPLE* υπο συγκεκριμένες συνθήκες κανονικότητας έχει ορισμένες επιθυμητές ιδιότητες όμοιες με εκείνες των κλασικών εκτιμητών μέγιστης πιθανοφάνειας. Στην συγκεκριμένη ενότητα τα ασυμπτωτικά αποτελέσματα θα παρουσιαστούν κατα ανάλογο τρόπο με την πλήρως θεμελιωμένη θεωρία της μέγιστης πιθανοφάνειας (*Fahrmeir Kaufmann* (1985)).

Αρχικά λοιπόν αν η f είναι μια συνεχής και φραγμένη συνάρτηση με πεδίο ορισμού το Γ και πεδίο τιμών το R^p (όπου Γ είναι το σύνολο απο το οποίο παίρνει τιμές οι συμμεταβλητή διανυσματική στοχαστική ανέλιξη $\boldsymbol{Z}_{t-1}$) τότε

$$\frac{\sum_{t=1}^{N} f(\boldsymbol{Z}_{t-1})}{N} \to \int_{R^p} f(\boldsymbol{z}) \nu(d\boldsymbol{z}) \tag{A.21}$$

κατα πιθανότητα καθώς το $N \to \infty$. Κατα συνέπεια επειδή το ολοκλήρωμα είναι πραγματικός αριθμός ο πίνακας $\mathbf{G}_N(\boldsymbol{\beta})$, θα έχει ένα μη τυχαίο όριο που είναι ο πίνακας $\mathbf{G}(\boldsymbol{\beta})$, δηλαδή

$$\frac{\mathbf{G}_N(\boldsymbol{\beta})}{N} \to \mathbf{G}(\boldsymbol{\beta}) \tag{A.22}$$

κατα πιθανότητα καθώς το $N \to \infty$. Ο μη τυχαίος πίνακας $\mathbf{G}(\boldsymbol{\beta})$ ονομάζεται *οριακός πίνακας πληροφορίας ανα παρατήρηση*, διάστασης $p \times p$. Αξίζει να σημειώσουμε ότι επειδή ο $\mathbf{G}_N(\boldsymbol{\beta})$ είναι θετικά ορισμένος με πιθανότητα 1, θα ισχύει ότι και ο μη τυχαίος πίνακας $\mathbf{G}(\boldsymbol{\beta})$ είναι θετικά ορισμένος για την πραγματική τιμή $\boldsymbol{\beta}$ και επομένως θα υπάρχει ο αντίστροφος του. Παράλληλα για τον πίνακα $\mathbf{R}_N(\boldsymbol{\beta})$ ισχύει

$$\frac{\mathbf{R}_N(\boldsymbol{\beta})}{N} \to 0 \tag{A.23}$$

κατα πιθανότητα καθώς το $N \to \infty$. Επομένως σύμφωνα με την (A.19) θα έχουμε

$$\frac{\mathbf{H}_N(\boldsymbol{\beta})}{N} \to \mathbf{G}_N(\boldsymbol{\beta}). \tag{A.24}$$

Για την ασυμπτωτική κατανομή του $\mathbf{S}_N(\boldsymbol{\beta})$ υπενθυμίζουμε ότι η διανυσματική στοχαστική ανέλιξη $\{\mathbf{S}_t(\boldsymbol{\beta})\}, t = 1, 2, \ldots$ έχει την ιδιότητα *Martingale*. Επομένως χρησιμοποιώντας το σχετικό Κ.Ο.Θ των *Martingales* (*Hall* και *Heyde* (1980)) προκύπτει

$$\frac{\mathbf{S}_N(\boldsymbol{\beta})}{\sqrt{N}} \rightarrow N_p(\mathbf{0}, \mathbf{G}(\boldsymbol{\beta})) \tag{A.25}$$

κατα κατανομή καθώς το $N \rightarrow \infty$.

Βάση των προαναφερθέντων ασυμπτωτικών αποτελεσμάτων μπορούμε να προχωρήσουμε στο ακόλουθο θεώρημα το οποίο εξασφαλίζει την ύπαρξη, την μοναδικότητα και προσδιορίζει την ασυμπτωτική κατανομή των εκτιμητών μέγιστης μερικής πιθανοφάνειας (*MPLE*).

Θεώρημα A.3.1 *Κάτω απο κατάλληλες συνθήκες ομαλότητας και κανονικότητας η πιθανότητα ότι τοπικά υπάρχει μοναδικός εκτιμητής μέγιστης μερικής πιθανοφάνειας συγκλίνει στο 1. Επιπλέον υπάρχει μια ακολουθία εκτιμητών μέγιστης μερικής πιθανοφάνειας που είναι συνεπείς και ασυμπτωτικά κανονικοί. Δηλαδή θα ισχύει*

$$\sqrt{N}(\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}) \xrightarrow{d} N_p(0, \mathbf{G}^{-1}(\boldsymbol{\beta}))$$

καθώς το $N \rightarrow \infty$.

Με τα συγκεκριμένα ζητήματα έχουν ασχοληθεί αρκετοί συγγραφείς μεταξύ των οποίων οι *Albert* και *Anderson* (1984) καθώς και ο *Kaufmann* (1989).

Παράρτημα Β

Επαναλαμβανόμενα Επανασταθμιζόμενα Ελάχιστα Τετράγωνα *(Iterative Reweighted Least Squares-IRLS)*

Β.1 Εισαγωγή

Όπως έχουμε ήδη αναφέρει στο Παράρτημα Α΄.2 η επίλυση των εξισώσεων

$$\mathbf{S}_N(\boldsymbol{\beta}) = \nabla \log PL(\boldsymbol{\beta}) = \mathbf{0} \tag{Β.1}$$

οδηγεί στην εύρεση των εκτιμητών μέγιστης μερικής πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ για τα μοντέλα παλινδρόμησης των κατηγορικών χρονοσειρών. Επειδή το σύστημα εξισώσεων (Β.1) είναι μη γραμμικό η λύση του επιτυγχάνεται επαναληπτικά μέσω της μεθόδου *Fisher scoring (Fsc)* που αποτελεί τροποποίηση του αλγορίθμου *Newton-Raphson (NR)*. Τα γνωστά αποτελέσματα των μεθοδολογικών προσεγγίσεων (*NR*) και (*Fsc*) (βλέπε *Agresti* (2002), κεφάλαιο 4), για ανεξάρτητες παρατηρήσεις, δεν μπορούν να χρησιμοποιηθούν αμιγώς στις ποιοτικές χρονοσειρές όπου συναντάμε την διαχρονική εξάρτηση και την ετερογένεια. Στις ενότητες που θα ακολουθήσουν θα παρουσιαστούν οι απαραίτητες τροποποιήσεις ώστε να καταλήξουμε στον αλγόριθμο *Fsc*, στα πλαίσια του μη τυχαίου πειράματος, αποδεικνύοντας και σε αυτή την περίπτωση ότι η δεδομένη διαδικασία ισοδυναμεί με την μέθοδο *IRLS*.

Β.2 Μεθοδος *NR* και *Fsc* για Κατηγορικές Χρονοσειρές.

Στην περίπτωση λοιπόν διαχρονικής εξάρτησης και ετερογένειας η μέθοδος NR στην $(k+1)$-οστή επανάληψη της δίνει την ακόλουθη λύση για το $\hat{\boldsymbol{\beta}}$

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + [\mathbf{H}_N(\hat{\boldsymbol{\beta}}^{(k)})]^{-1}\mathbf{S}_N(\hat{\boldsymbol{\beta}}^{(k)}) \tag{B.2}$$

με $\mathbf{H}_{N,ij}(\boldsymbol{\beta}) = -\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}_i \partial \boldsymbol{\beta}_j}$.

Παρατήρηση Β.2.1 Βλέπουμε πως πλέον δεν χρησιμοποιούμε τον *Hessian* πίνακα αλλά τον *παρατηρούμενο πίνακα πληροφορίας (observed information matrix)* για τον οποίο ισχύει $\mathbf{H}_N(\boldsymbol{\beta}) = (Hessian) \times (-1)$ (*McCullagh and Nelder* (1989)).

Η μέθοδος *Fsc* αναλογικά με την περίπτωση των ανεξάρτητων παρατηρήσεων θα χρησιμοποιεί αντί του $\mathbf{H}_N(\boldsymbol{\beta})$ τον πίνακα $E[\mathbf{H}_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}]$ που θα ονομάζεται *αναμενόμενος πίνακας πληροφορίας*(*expected information matrix*).

Πρόταση Β.2.1 *Για τον αναμενόμενο πίνακα πληροφορίας ισχύει*

$$E[\mathbf{H}_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}] = \mathbf{G}_N(\boldsymbol{\beta}). \tag{B.3}$$

Απόδειξη Β.2.1 *Για την περίπτωση των ανεξάρτητων δεδομένων (στην E.O.K) γνωρίζουμε ότι ισχύει*

$$E\Big[\frac{\partial \ell_i}{\partial \beta_i} \cdot \frac{\partial \ell_i}{\partial \beta_j}\Big] = E\Big[-\frac{\partial^2 \ell_i}{\partial \beta_i \partial \beta_j}\Big]$$

*(*Cox *και* Hinkley *(1974), ενότητα 4.8). Αναλόγως στις ποιοτικές χρονοσειρές θα έχουμε*

$$E\Big[\frac{\partial \ell_t}{\partial \beta_i} \cdot \frac{\partial \ell_t}{\partial \beta_j} \mid \mathcal{F}_{t-1}\Big] = E\Big[-\frac{\partial^2 \ell_t}{\partial \beta_i \partial \beta_j} \mid \mathcal{F}_{t-1}\Big]. \tag{B.4}$$

Για το ij στοιχείο του $p \times p$ πίνακα $E[\mathbf{H}_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}]$ ισχύει

$$E[\mathbf{H}_{N,ij}(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}] = E\Big[-\frac{\partial^2 \ell(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \mid \mathcal{F}_{t-1}\Big] \overset{\ell=\sum_{t=1}^N \ell_t}{=} E\Big[-\sum_{t=1}^N \frac{\partial^2 \ell_t(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \mid \mathcal{F}_{t-1}\Big]$$

$$= \sum_{t=1}^N \Big\{E\Big[-\frac{\partial^2 \ell_t(\boldsymbol{\beta})}{\partial \beta_i \partial \beta_j} \mid \mathcal{F}_{t-1}\Big]\Big\} \overset{(4)}{=} \sum_{t=1}^N \Big\{E\Big[\frac{\partial \ell_t(\boldsymbol{\beta})}{\partial \beta_i} \cdot \frac{\partial \ell_t(\boldsymbol{\beta})}{\partial \beta_j} \mid \mathcal{F}_{t-1}\Big]\Big\}$$

$$= \sum_{t=1}^N \Big\{E\Big[\frac{(y_t - \mu_t)}{\sigma_t^2(\boldsymbol{\beta})}\frac{\partial \mu_t}{\partial \eta_t} Z_{(t-1)i} \frac{(y_t - \mu_t)}{\sigma_t^2(\boldsymbol{\beta})}\frac{\partial \mu_t}{\partial \eta_t} Z_{(t-1)j} \mid \mathcal{F}_{t-1}\Big]\Big\}$$

$$= \sum_{t=1}^{N} \Big\{ E\Big[\frac{(y_t - \mu_t)^2}{(\sigma_t^2(\boldsymbol{\beta}))^2} \Big(\frac{\partial \mu_t}{\partial \eta_t} \Big)^2 Z_{(t-1)i} Z_{(t-1)j} \mid \mathcal{F}_{t-1} \Big] \Big\}$$

$$= \sum_{t=1}^{N} \Big\{ \Big[\frac{E[(y_t - \mu_t)^2 \mid \mathcal{F}_{t-1}]}{(\sigma_t^2(\boldsymbol{\beta}))^2} \Big(\frac{\partial \mu_t}{\partial \eta_t} \Big)^2 Z_{(t-1)i} Z_{(t-1)j} \Big] \Big\}$$

$$= \sum_{t=1}^{N} \Big\{ Z_{(t-1)i} Z_{(t-1)j} \Big(\frac{\partial \mu_t}{\partial \eta_t} \Big)^2 \frac{1}{\sigma_t^2(\boldsymbol{\beta})} \Big\} \equiv \mathbf{G}_{Nij}(\boldsymbol{\beta}).$$

Έτσι, σύμφωνα με την σχέση (B.2), η μέθοδος *Fsc* δίνει λύση για το $\hat{\boldsymbol{\beta}}$ κατα την $(k+1)$-οστή επανάληψη

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \Big\{ E[\mathbf{H}_N(\hat{\boldsymbol{\beta}}^{(k)}) \mid \mathcal{F}_{t-1}] \Big\}^{-1} \mathbf{S}_N(\hat{\boldsymbol{\beta}}^{(k)}) \overset{(B.3)}{\Rightarrow}$$

$$\hat{\boldsymbol{\beta}}^{(k+1)} = \hat{\boldsymbol{\beta}}^{(k)} + \mathbf{G}_N^{-1}(\hat{\boldsymbol{\beta}}^{(k)}) \mathbf{S}_N(\hat{\boldsymbol{\beta}}^{(k)}) \tag{B.5}$$

υποθέτοντας ότι υπάρχει ο πίνακας $\mathbf{G}_N^{-1}$.

Παρατήρηση B.2.2 Στην περίπτωση που έχω κανονική συνάρτηση σύνδεσης ισχύει $\mathbf{H}_N(\boldsymbol{\beta}) = \mathbf{G}_N(\boldsymbol{\beta})$ και επομένως, όπως και στις ανεξάρτητες παρατηρήσεις, τα αποτελέσματα της μεθόδου *NR* θα ταυτίζονται με εκείνα του αλγορίθμου *Fsc*.

Σε θεωρητικό πλαίσιο ο εκτιμητής μέγιστης μερικής πιθανοφάνειας $\hat{\boldsymbol{\beta}}$ δίνεται απο την σχέση

$$\hat{\boldsymbol{\beta}} = \lim_{k \to \infty} \hat{\boldsymbol{\beta}}^{(k)}$$

με $\hat{\boldsymbol{\beta}}^{(k)}$ να παρέχεται είτε απο την σχέση (B.2) είτε απο την σχέση (B.5). Στην πράξη το k λαμβάνει μικρές τιμές αφού η σύγκλιση των δυο αλγορίθμων είναι σχετικά γρήγορη. Αναλυτικότερα η επαναληπτική διαδικασία ολοκληρώνεται όταν για μεγάλο k και για κάθε j, ισχύει

$$|\beta_j^{(k+1)} - \hat{\beta}_j| \leq c|\beta_j^{(k)} - \hat{\beta}_j|^2 \quad \text{για κάποιο} \quad c > 0,$$

όπου με $\hat{\beta}_j$, $j = 1, 2, \ldots, p$ συμβολίζουμε την j συνιστώσα του $MPLE$. Το προαναφερθέν κριτήριο σύγκλισης ονομάζεται δεύτερης τάξης (*second-order*). Τέλος αξίζει να τονίσουμε ότι για μη κανονικές συναρτήσεις σύνδεσης η μέθοδος *Fsc* πλεονεκτεί συγκριτικά με την *NR* για τους ακόλουθους λόγους (βλέπε *Agresti* (2002), κεφάλαιο 4)

1) Παράγει τον ασυμπτωτικό πίνακα συνδιακύμανσης ως ένα *by-product*.

2) Ο πίνακας $E[\mathbf{H}_N(\boldsymbol{\beta}) \mid \mathcal{F}_{t-1}]$ είναι αναγκαστικά μη αρνητικά ορισμένος.

3) Ο αλγόριθμος *Fsc* συνδέεται άμεσα με την μέθοδο των ελαχίστων τετραγώνων.

Β.3 Παρουσίαση της Μέγιστης Μερικής Πιθανοφάνειας μέσω της διαδικασίας *IRLS*.

Όπως είναι γνωστό για ανεξάρτητα δεδομένα, στα πλαίσια της θεωρίας των *GLM*, ο αλγόριθμος *Fsc* μπορεί να αντιμετωπιστεί ως μέθοδος εκτίμησης επαναλαμβανόμενων επαναστάθμιζόμενων ελαχίστων τετραγώνων. Το δεδομένο αποτέλεσμα ισχύει και στις κατηγορικές χρονοσειρές.

Η σχέση (5) υποθέτωντας ότι υπάρχει ο $\mathbf{G}_N^{-1}(\boldsymbol{\beta})$ γράφεται

$$\underbrace{\mathbf{G}_N(\hat{\boldsymbol{\beta}}^{(k)})\hat{\boldsymbol{\beta}}^{(k+1)}}_{p\times 1} = \underbrace{\mathbf{G}_N(\hat{\boldsymbol{\beta}}^{(k)})}_{p\times p}\underbrace{\hat{\boldsymbol{\beta}}^{(k)}}_{p\times 1} + \underbrace{\mathbf{S}_N(\hat{\boldsymbol{\beta}}^{(k)})}_{p\times 1} \tag{B.6}$$

Για λόγους απλότητας των συμβολισμών προσωρινά θέτουμε όπου $\hat{\boldsymbol{\beta}}^{(k)}$ το $\boldsymbol{\beta}$ και το δεύτερο μέλος της (B.6) γράφεται

$$\mathbf{G}_N(\boldsymbol{\beta})\boldsymbol{\beta} + \mathbf{S}_N(\boldsymbol{\beta}) = \begin{bmatrix} \sum_{j=1}^{p} G_{N,1j}(\boldsymbol{\beta})\cdot\beta_j \\ \sum_{j=1}^{p} G_{N,2j}(\boldsymbol{\beta})\cdot\beta_j \\ \vdots \\ \sum_{j=1}^{p} G_{N,pj}(\boldsymbol{\beta})\cdot\beta_j \end{bmatrix} + \begin{bmatrix} \frac{\partial \ell}{\partial \beta_1} \\ \frac{\partial \ell}{\partial \beta_2} \\ \vdots \\ \frac{\partial \ell}{\partial \beta_p} \end{bmatrix}.$$

Άρα το ℓ-οστό στοιχείο του p-διάστατου διανύσματος $\mathbf{G}_N(\boldsymbol{\beta})\boldsymbol{\beta} + \mathbf{S}_N(\boldsymbol{\beta})$ θα ισούται με

$$\sum_{j=1}^{p}\Big\{\Big(\sum_{t=1}^{N} Z_{(t-1)\ell}Z_{(t-1)j}\frac{1}{\sigma_t^2(\boldsymbol{\beta})}\Big(\frac{\partial\mu_t}{\partial\eta_t}\Big)^2\beta_j\Big\} + \sum_{t=1}^{N} Z_{(t-1)\ell}\frac{\partial\mu_t}{\partial\eta_t}\frac{(Y_t-\mu_t(\boldsymbol{\beta}))}{\sigma_t^2(\boldsymbol{\beta})}$$

$$= \sum_{t=1}^{N} Z_{(t-1)\ell}\cdot\omega_t\Big\{\eta_t + (Y_t-\mu_t)\frac{\partial\eta_t}{\partial\mu_t}\Big\}, \quad \ell = 1,2,\ldots,p$$

όπου $\mu_t, \eta_t, (\partial\mu_t/\partial\eta_t), \omega_t$ είναι υπολογισμένα για $\boldsymbol{\beta} = \hat{\boldsymbol{\beta}}^{(k)}$ και $\omega_t = \Big(\frac{\partial\mu_t}{\partial\eta_t}\Big)^2\frac{1}{\sigma_t^2(\boldsymbol{\beta})}$, $\quad t = 1,2,\ldots,N$. Για $t = 1,2,\ldots,N$ ορίζουμε

$$q_t^{(k)} = \sum_{j=1}^{p} Z_{(t-1)j}\boldsymbol{\beta} + (Y_t-\mu_t)\frac{\partial\eta_t}{\partial\mu_t} =$$

$$= \eta_t(\boldsymbol{\beta}) + (Y_t-\mu_t)\frac{\partial\eta_t}{\partial\mu_t}.$$

Θεωρώντας λοιπόν το διάνυσμα $\boldsymbol{q}^{(k)} = (q_1^{(k)}, q_2^{(k)}, \ldots, q_2^{(k)})'_{N\times 1}$ το δεξιό μέλος της (6) γράφεται $\underbrace{\mathbf{Z}'}_{p\times N}\underbrace{\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})}_{N\times N}\underbrace{\boldsymbol{q}^{(k)}}_{N\times 1}$ με $\mathbf{Z} = \begin{bmatrix} \mathbf{Z}_0 \\ \mathbf{Z}_1 \\ \vdots \\ \mathbf{Z}_{N-1} \end{bmatrix}$. Έτσι λαμβάνοντας υπόψη ότι $\mathbf{G}_N(\boldsymbol{\beta}) =$

$\mathbf{Z}'\mathbf{W}\mathbf{Z}'$ (βλέπε παράρτημα A.2 σχέση A.16) η (B.6) γράφεται

$$[\mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\mathbf{Z}]\hat{\boldsymbol{\beta}}^{(k+1)} = \mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{q}^{(k)}.$$

Επομένως η μέθοδος *Fsc* απλοποιείται στην ακόλουθη έκφραση

$$\hat{\boldsymbol{\beta}}^{(k+1)} = [\mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})\boldsymbol{q}^{(k)} \tag{B.7}$$

όπου το $\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})$ και $\boldsymbol{q}^{(k)}$ είναι υπολογισμένα στο $\hat{\boldsymbol{\beta}}^{(k)}$.

Η σχέση (B.6) ουσιαστικά δίνει τις γενικές κανονικές εξισώσεις της διαδικασίας εκτίμησης των *Σταθμιζόμενων Ελαχίστων Τετραγώνων(Weighted Least Squares-WLS)* που η λύση τους παρέχεται απο τον τύπο (B.7). Πράγματι αντιμετωπίζοντας το διάνυσμα $\boldsymbol{q} = (q_1(\boldsymbol{\beta}), q_2(\boldsymbol{\beta}), \ldots, q_N(\boldsymbol{\beta}))'$ ως ένα γραμμικό μετασχηματισμό του διανύσματος $\boldsymbol{y} = (y_1, y_2, \ldots, y_N)'$ σύμφωνα με την σχέση

$$g(y_t) \approx g(\mu_t) + (y_t - \mu_t)g'(\mu_t) = \eta_t + (y_t - \mu_t)\cdot(\frac{\partial \eta_t}{\partial \mu_t}) \equiv q_t, \quad t = 1, 2, \ldots, N, \tag{B.8}$$

όπου g είναι η συνάρτηση σύνδεσης, τότε μπορούμε να θεωρήσουμε το γενικό γραμμικό μοντέλο

$$\boldsymbol{q} = \mathbf{Z}\boldsymbol{\beta} + \boldsymbol{\epsilon}. \tag{B.9}$$

Στο δεδομένο υπόδειγμα ο πίνακας σχεδιασμού είναι ο $\underbrace{\mathbf{Z}}_{N\times p}$. Στην περίπτωση που ο πίνακας διακυμάνσεων -συνδιακυμάνσεων του διανύσματος των σφαλμάτων $\boldsymbol{\epsilon}$ είναι ο $\mathbf{V}$, τότε όπως γνωρίζουμε ο εκτιμητής *WLS* του $\boldsymbol{\beta}$ είναι

$$[\mathbf{Z}'\mathbf{V}^{-1}\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{V}^{-1}\boldsymbol{q}. \tag{B.10}$$

Θεωρώντας ότι ο $\mathbf{V}^{-1}$ ισούται με τον πίνακα $\mathbf{W}$ τότε η σχέση (B.10) ταυτίζεται με την (B.7) για την k-οστή επανάληψη και ουσιαστικά αποτελεί την λύση της (B.6). Σύμφωνα με την δεδομένη προσέγγιση η *προσαρμοσμένη ή 'βοηθητική' αποκριτική* (*adjusted* or *"working" response*) μεταβλητή $\boldsymbol{q}$ κατα τον k-οστό κύκλο του επαναληπτικού σχήματος για την t-οστή συνιστώσα της έχει τιμή που δίνεται απο

$$q_t^{(k)} = \eta_t(\hat{\boldsymbol{\beta}}^{(k)}) + (Y_t - \mu_t)(\frac{\partial \eta_t}{\partial \mu_t}).$$

Στην δεδομένη επανάληψη το διάνυσμα $\boldsymbol{q}^{(k)}$ παλινδρομεί πάνω στο $\mathbf{Z}$ με βάρος τον πίνακα $\mathbf{W}(\hat{\boldsymbol{\beta}}^{(k)})$ με σκοπό να ληφθεί μια νέα εκτίμηση $\hat{\boldsymbol{\beta}}^{(k+1)}$ σύμφωνα με την σχέση (B.7). Αυτή η εκτίμηση δίνει μια νέα τιμή για την γραμμική πρόβλεψη που είναι η

$\boldsymbol{\eta}^{(k+1)} = \mathbf{Z}\hat{\boldsymbol{\beta}}^{(k+1)}$ και κατα συνέπεια μια νέα *προσαρμοσμένη* τιμή $\boldsymbol{q}^{(k+1)}$ (σύμφωνα με το μοντέλο (Β.9)) για τον επόμενο κύκλο. Με τον τρόπο αυτό ο εκτιμητής μέγιστης μερικής πιθανοφάνειας του $\boldsymbol{\beta}$ προκύπτει απο μια επαναληπτική χρήση της μεθόδου *WLS*, στην οποία ο πίνακας των βαρών καθώς και η *'βοηθητική'* εξαρτημένη μεταβλητή αλλάζουν σε κάθε κύκλο. Για τον λόγο αυτό η διαδικασία (Β.7) ονομάζεται *IRLS*. Η προαναφερθείσα μεθοδολογική προσέγγιση είναι έγγυρη για όλα τα *GLM* ανεξάρτητα απο την επιλογή της συνάρτησης σύνδεσης.

Για να λάβουμε μια αρχική εκτίμηση για το $\boldsymbol{\beta}$ χρησιμοποιούμε τα δεδομένα $\boldsymbol{y}$ ως μια αρχική εκτίμηση του διανύσματος $\boldsymbol{\mu} = (\mu_1, \mu_2, \ldots, \mu_t, \ldots, \mu_N)'$ όπου $\mu_t = E[Y_t \mid F_{t-1}]$. Έτσι προκύπτουν οι πρώτες εκτιμήσεις των $\mathbf{W}$ και $\boldsymbol{q}$. Οι επαναλήψεις συνεχίζονται μέχρι κάποιο κριτήριο σύγκλισης να ικανοποιείται.

Στο σημείο αυτό παρουσιάζουμε τα βήματα του αλγορίθμου εύρεσης του *MPLE* $\hat{\boldsymbol{\beta}}$ που μπορεί να υλοποιηθεί μέσω οποιοδήποτε υπολογιστικού προγράμματος.

Βήμα 1. Για να πάρουμε αρχική λύση για το $\boldsymbol{\beta}$, έστω την $\hat{\boldsymbol{\beta}}^{(0)}$, επιλύουμε το γραμμικό σύστημα

$$\mathbf{Z}'\mathbf{Z}\hat{\boldsymbol{\beta}}^{(0)} = \mathbf{Z}' \cdot \boldsymbol{g}(\boldsymbol{y}),$$

όπου $\boldsymbol{g}(\boldsymbol{y}) = [g(y_1), g(y_2), \ldots, g(y_N)]'$.
Βήμα 2. Θέτουμε $\boldsymbol{n}^{(0)} = \boldsymbol{g}(\boldsymbol{y})$ και υπολογίζουμε το $(\partial\mu_t)/(\partial\eta_t)$ καθώς και τα στοιχεία του $\mathbf{W}$ για $\hat{\boldsymbol{\beta}}^{(0)}$.
Βήμα 3. Βρίσκουμε το $\hat{\boldsymbol{\beta}}^{(1)}$ σύμφωνα με την σχέση (Β.7)

$$\hat{\boldsymbol{\beta}}^{(1)} = [\mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(0)})\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(0)})\boldsymbol{q}^{(0)}.$$

Βήμα 4. Για το $\hat{\boldsymbol{\beta}}^{(2)}$ ομοίως θα έχουμε

$$\hat{\boldsymbol{\beta}}^{(2)} = [\mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(1)})\mathbf{Z}]^{-1}\mathbf{Z}'\mathbf{W}(\hat{\boldsymbol{\beta}}^{(1)})\boldsymbol{q}^{(1)}$$

κ.ο.κ.
Βήμα 4. Σταματάμε όταν κάνουμε m βήματα. Συνήθως $m \leq 50$.

Παράρτημα Γ

Τεχνικό Παράρτημα

Γ.1 Απόδειξη της (4.34)

$$\begin{aligned}
\mathbf{U}_t(\boldsymbol{\beta}) = \mathbf{D}_t(\boldsymbol{\beta})\boldsymbol{\Sigma}_t^{-1}(\boldsymbol{\beta}) &= \begin{bmatrix} \frac{\partial \pi_{t1}}{\partial \eta_{t1}} & \frac{\partial \pi_{t1}}{\partial \eta_{t2}} \\ \frac{\partial \pi_{t2}}{\partial \eta_{t1}} & \frac{\partial \pi_{t2}}{\partial \eta_{t2}} \end{bmatrix}' \begin{bmatrix} \frac{\partial \theta_{t1}}{\partial \pi_{t1}} & \frac{\partial \theta_{t1}}{\partial \pi_{t2}} \\ \frac{\partial \theta_{t2}}{\partial \pi_{t1}} & \frac{\partial \theta_{t2}}{\partial \pi_{t2}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial \pi_{t1}}{\partial \eta_{t1}} & \frac{\partial \pi_{t2}}{\partial \eta_{t1}} \\ \frac{\partial \pi_{t1}}{\partial \eta_{t2}} & \frac{\partial \pi_{t2}}{\partial \eta_{t2}} \end{bmatrix} \begin{bmatrix} \frac{\partial \theta_{t1}}{\partial \pi_{t1}} & \frac{\partial \theta_{t1}}{\partial \pi_{t2}} \\ \frac{\partial \theta_{t2}}{\partial \pi_{t1}} & \frac{\partial \theta_{t2}}{\partial \pi_{t2}} \end{bmatrix} \\
&= \begin{bmatrix} \frac{\partial \pi_{t1}}{\partial \eta_{t1}}\frac{\partial \theta_{t1}}{\partial \pi_{t1}} + \frac{\partial \pi_{t2}}{\partial \eta_{t1}}\frac{\partial \theta_{t2}}{\partial \pi_{t1}} & \frac{\partial \pi_{t1}}{\partial \eta_{t1}}\frac{\partial \theta_{t1}}{\partial \pi_{t2}} + \frac{\partial \pi_{t2}}{\partial \eta_{t1}}\frac{\partial \theta_{t2}}{\partial \pi_{t2}} \\ \frac{\partial \pi_{t1}}{\partial \eta_{t2}}\frac{\partial \theta_{t1}}{\partial \pi_{t1}} + \frac{\partial \pi_{t2}}{\partial \eta_{t2}}\frac{\partial \theta_{t2}}{\partial \pi_{t1}} & \frac{\partial \pi_{t1}}{\partial \eta_{t2}}\frac{\partial \theta_{t1}}{\partial \pi_{t2}} + \frac{\partial \pi_{t2}}{\partial \eta_{t2}}\frac{\partial \theta_{t2}}{\partial \pi_{t2}} \end{bmatrix} \equiv \mathbf{A} \qquad (\Gamma.1)
\end{aligned}$$

απο την άλλη, σύμφωνα με την (4.35), θα έχουμε

$$\frac{\partial \boldsymbol{u}(\boldsymbol{\eta}_t)}{\partial \boldsymbol{\eta}_t} = \begin{bmatrix} \frac{\partial u_1(\boldsymbol{\eta}_t)}{\partial \eta_1} & \frac{\partial u_2(\boldsymbol{\eta}_t)}{\partial \eta_1} \\ \frac{\partial u_1(\boldsymbol{\eta}_t)}{\partial \eta_2} & \frac{\partial u_2(\boldsymbol{\eta}_t)}{\partial \eta_2} \end{bmatrix} \equiv \mathbf{B}.$$

Για να δειχθεί λοιπόν η σχέση (4.34) αρκεί να αποδείξουμε ότι $\mathbf{A} \equiv \mathbf{B}$. Ενδεικτικά θα δείξουμε ότι $A_{11} = B_{11}$.

$$\begin{aligned}
B_{11} &= \frac{1}{\frac{h_1}{1-h_1-h_2}} \frac{\frac{\partial h_1}{\partial \eta_1}(1-h_1-h_2) - h_1(-\frac{\partial h_1}{\partial \eta_1} - \frac{\partial h_2}{\partial \eta_1})}{(1-h_1-h_2)^2} \\
&= \frac{\frac{\partial h_1}{\partial \eta_1}(1-h_1-h_2) + h_1 + \frac{\partial h_1}{\partial \eta_1} + h_1\frac{\partial h_2}{\partial \eta_1}}{h_1(1-h_1-h_2)} \\
&\overset{(4.11)}{=} \frac{\frac{\partial \pi_{t1}}{\partial \eta_1}(1-\pi_{t1}-\pi_{t2}) + \pi_{t1}\frac{\partial \pi_{t1}}{\partial \eta_1} + \pi_{t1}\frac{\partial \pi_{t2}}{\partial \eta_1}}{\pi_{t1}(1-\pi_{t1}-\pi_{t2})} \\
&= \frac{\frac{\partial \pi_{t1}}{\partial \eta_1}(1-\pi_{t2}) + \pi_{t1}\frac{\partial \pi_{t2}}{\partial \eta_1}}{\pi_{t1}(1-\pi_{t1}-\pi_{t2})}
\end{aligned}$$

$$= \frac{\partial \pi_{t1}}{\partial \eta_1} \frac{1 - \pi_{t2}}{\pi_{t1}(1 - \pi_{t1} - \pi_{t2})} + \frac{\partial \pi_{t2}}{\partial \eta_1} \frac{\pi_{t1}}{\pi_{t1}(1 - \pi_{t1} - \pi_{t2})}$$

$$\overset{(4.27)}{\Longrightarrow} B_{11} = \frac{\partial \pi_{t1}}{\partial \eta_1} \frac{\partial \theta_{t1}}{\partial \pi_{t1}} + \frac{\partial \pi_{t2}}{\partial \eta_1} \frac{\partial \theta_{t2}}{\partial \pi_{t1}} \equiv A_{11}.$$

Συνεχίζοντας με όμοιο τρόπο και για τα υπόλοιπα στοιχεία των πινάκων **A** και **B** αποδεικνύεται το ζητούμενο.

Παράρτημα Δ

Μοντελοποίηση των δίτιμων σειρών μέσω του *S-PLUS*

Δ.1 Εισαγωγικά

Οι εκτιμητές $MPLE$ για τις κατηγορικές χρονοσειρές μπορούν να υπολογιστούν μέσω του στατιστικού προγράμματος $S-PLUS$. Αυτό προκύπτει απο το γεγονός ότι οι εξισώσεις των μερικών σκόρ (A.14) έχουν την ίδια μορφή με αυτή που παρατηρείται όταν έχουμε ανεξάρτητα δεδομένα. Αυτό σημαίνει ότι τα τυπικά σφάλματα των παραμέτρων των υποδειγμάτων προκύπτουν προσεγγιστικά απο την αντιστροφή του πίνακα $\mathbf{G}_N(\boldsymbol{\beta})$ (A.16) και την λήψη τετραγωνικών ριζών των διαγωνίων στοιχείων του. Βέβαια δεν πρέπει να ξεχνάμε πως όλα τα αποτελέσματα είναι δεσμευμένα.

Στο $S-PLUS$ η προσαρμογή των λογιστικών μοντέλων για δίτιμες σειρές υλοποιείται μέσω της συνάρτησης $glm()$ και της βιβλιοθήκης $MASS$ (βλέπε *Venables and Ripley* (1999)). Για την επιλογή βέλτιστου μοντέλου από ένα σύνολο συμμεταβλητών χρησιμοποιείται το κριτήριο πληροφορίας AIC. Στο $S-PLUS$ δεν χρειάζεται να προσαρμόσουμε όλα τα δυνατά υποδείγματα, ώστε να επιλέξουμε εκείνο με το μικρότερο AIC αφού αυτό γίνεται αυτόματα μέσω της συνάρτησης $stepAIC$.

Για την προσαρμογή του εκάστοτε μοντέλου χρησιμοποιείται το ακόλουθο $script file$ του $S-PLUS$

```
attach(data)
model.glm ← glm(yt∼ ..., family=binomial, data=data)
summary(model.glm)
anova(model.glm).
```

Για την εύρεση του βέλτιστου μοντέλου μέσω του AIC έχουμε

```
modelx.step ← stepAIC(model.glm, trace=F)
model.step$anova.
```

Ο πίνακας $\mathbf{G}_N(\boldsymbol{\beta})$ δίδεται απο

```
vcov.glm ← function(model.glm){
so ← summary(model.glm, corr=F)
so$dispersion*so$cov.unscaled
}
.vcov(model.glm)
```

Για τα *fitted values*, τα υπόλοιπα του *Pearson*, το $\mathcal{X}^2$ του *Pearson* καθώς και το MSE έχουμε

```
residpearson ← residuals(model.glm,type=‘‘pearson’’)
residpearson2 ← residpearson^2
f1 ←fitted(model.glm)
PearsonStatistic1 ← sum(residpearson2)
endiameso ← ((yt-f1)^2)/(f1*(1-f1))
PearsonStatistic2 ← sum(endiameso)
PearsonStatistic1
PearsonStatistic2
MSE ← 1/(length(yt))*(sum((yt-f1)^2))
MSE.
```

ΒΙΒΛΙΟΓΡΑΦΙΑ

Ελληνική

Ξενάκης, Α. Σ. (1998). *Ανάλυση Χρονολογικών Σειρών και Προβλέψεις*, Εκδόσεις Οικονομικού Πανεπιστημίου Αθηνών, Αθήνα.

Πιττής, Ν. (2003). *Σημειώσεις Χρονοσειρών*, Πανεπιστήμιο Πειραιώς.

Ξένη

Adali, T. and Ni, H. (2003). Partial Likelihood for Signal Processing, *IEEE Transactions on Signal Processing*, **51, 1**, 204-212.

Akaike, H. (1973). Information theory and an extension of the maximum likelihood principle. In B.N. Petroc and F. Kaski, editors, *Second International Symposium in Information Theory*, 267-281. Akademiai Kiado, Budapest.

Akaike, H. (1974). A new look at the statistical model identification. *IEEE Transactions on Automatic Control*, **AC-19**, 716-723.

Albert, A. and Anderson, J. A. (1984). On the existence of maximum likelihood estimates in logistic regression models, *Biometrica*, **71**, 1-10.

Andersen, E. B. (1973). *Conditional Inference and Models for Measuring.* Mentalhygiejnisk Forlag, Copenhagen.

Agresti, A. (2002). *Categorical Data Analysis*, 2^{nd} ed., Wiley, New York.

Bartlett, M. S. (1950). Periodogram analysis and continuous spectra. *Biometrica*, **37**, 1-16.

Basawa, I. V. and Prakasa Rao, R. L. S. (1980). *Statistical Inference for Stochastic Processes.* Academic Press, London.

Bhat, B. R. (1974). On the method of maximum likelihood for dependent observations. *Journal of the Royal Statistical Society Series B*, **36**, 48-53.

Birch, M. W. (1964a). A new proof of the Pearson-Fisher theorem, *Annals of Mathematical Statistics*, **35**, 817-824.

Bonney, E. G. (1987). Logistic regression for dependent binary observations, *Biometrics*, **43**, 951-973.

Box, G. E. P. and Jenkins, G. M. (1976). *Time Series Analysis, Forecasting and Control,* 2^{nd} ed., Holden-Day, San Francisco.

Brillinger, D. R. (1996). An analysis of ordinal-valued time series. In *Athens Conference on Applied Probability and Time Series Analysis*, volume II: Time Series Analysis of *Lecture Notes in Statistics*, 73-87, Springer, New York.

Brillinger, D. R., Morettin, P. A., Irizarry, R. A. and Chiann, C. (2000). Some wavelet-based analyses of Markov chain data, *Signal Processing*, **80**, 1607-1627.

Brown, B. M. (1971). Martingale central limit theorems, *Annals of Mathematical Statistics*, **42**, 59-66.

Buhlmann, P. (1999). Sieve bootstrap with variable length Markov chains for stationary categorical times series, *Research Report, Eidgenossische Technische Hochschule (ETH)*, **89**, Zurich, Switzerland.

Chandler, R. E. (2003). On the use of generalized linear models for interpreting climate variability,

Chandler, R. E. and Wheater, H. S. (1998). Climate change detection using Generalized Linear Models for rainfall data - a case study from the West of Ireland. I. Preliminary analysis and modelling of rainfall occurence,

Chatfield, C. (1996). *The analysis of Time Series*, Chapman & Hall, Lonon.

Choi, B. (1992). *ARMA Model Identification*, Springer, New York.

Coe, R. and Stern, R. D. (1984). A model fitting analysis of daily rainfall data, *Journal of Royal Statistical Society*, **A47**, 1-34.

Conaway, M. R. (1989) A random effects model for binary data. *Biometrics*, **46**, 317-328.

Cox, D. (1970). *The Analysis of Binary Data*, Chapman & Hall, London.

Cox, D. (1975). Partial likelihood. *Biometrica*, **62**, 69-76.

Cox, D. R. and Hinkley, D. V. (1974). *Theoretical Statistics*, Chapman & Hall, London.

Cox, D. R. and Snell, E. J. (1989). *The Analysis of Binary Data*, 2^{nd} ed., Ghapman and Hall, London.

Cressie, N. A. C. and Read, T. R. C. (1984). Multinomial goodness of fit tests. *Journal of the Royal Statistical Society, Series B*, **46**, 440-464.

Crouchley, R. (1995) A random-effects model for ordered categorical data. *Journal of the American Statistical Association*, **90**, 489-498.

Crowder, M. J. (1976). Maximum likelihood estimation for dependent observations. *Journal of the Royal Statistical Society Series B*, **38**, 45-53.

Durbin, J. and Koopman, S. J. (2001). *Time series analysis by state space methods*, Oxford University Press, Oxford.

Fahrmeir, L. (1987). Asymptotic testing theory for generalized linear models, *Statistics*, **18**, 65-76.

Fahrmeir, L. (1992a). Posterior mode estimation by extended Kalman filtering for multivariate dynamic generalized linear models, *Journal of the American Statistical Association*, **87**, 501-509.

Fahrmeir, L. (1992b). State space modelling and conditional mode estimation for categorical time series. In: D. Brillinger, P. Caines, & J. Geweke (Eds.), New directions in time series analysis, **1**, Springer, New York, 87-109.

Fahrmeir, L. and Kaufmann, H. (1985). Consistency and asymptotic normality of the maximum likelihood estimates in generalized linear models, *Annals of Statistics*, **13**, 342-368.

Fahrmeir, L. and Kaufmann, H. (1987). Regression models for nonstationary categorical time series, *Journal of Time Series Analysis*, **8**, 147-160.

Fahrmeir, L. and Lang, S. (2000). Bayesin semiparametric regression analysis of multicategorical time-space data, *Annals of the Institute of Statistical Mathematics*, **53**, 11-30.

Fahrmeir, L. and Tutz, G. (2001). *Multivariate Statistical Modelling Based on Generalized Linear Models*, 2^{nd} ed., Springer, New York.

Finney, D. (1971). *Probit Analysis*, 3^{nd} ed., Cambridge University Press, Cambridge.

Fokianos, K. and Kedem, B. (2001). Partial likelihood inference for time series following generalized linear models, *Journal of Time Series Analysis*, **25**, issue 2, 173-197.

Fokianos, K. and Kedem, B. (2002). *Regression Models for Time Series Analysis*, Wiley, New Jersey.

Fokianos, K. and Kedem, B. (2003). Regression Theory for Categorical Time Series, *Statistical Science*, **0**, 1-20.

Glonek, G. F. V. and McCullagh, P. (1995) Multivariate logistic models. *Journal of the Royal Statistical Society B*, **57**, 533-546.

Haggan, V. and Ozaki, T. (1979). Amplitude-dependent AR model fitting for nonlinear random vibrations. Paper presented at the International Time Series Meeting, University of Nottingham, UK, March 1979.

Hall, P. and Heyde, C. C. (1980). *Martingale Limit Theory and Its Applications*, Academic Press, New York.

Hamilton, J. D. (1994). *Time Series Analysis*, Princeton University Press, New Jersey.

Harrison, P. J. and Stevens, C. F. (1976). Bayesian forecasting, *Journal of the Royal Statistical Society, Series B*, **38**, 205-247, with discussion.

Johnson, V. E. and Albert, J. H. (1999). *Ordinal Data Modelling*, Springer, New York.

Kalbfleisch, J. D. and Sprott, D. A. (1970). Application of likelihood methods to models involving large number of parameters. *Journal of the Royal Statistical Society Series B*, **32**, 175-208.

Kalman, R. E. (1960). A new approach to linear filtering and predictions problems. *Trans. ASME J. Basic Eng., Ser. D,* **82**, 35-45.

Kalman, R. E. and Bucy, R. S. (1961). New results in linear filtering and prediction problems. *Trans. ASME J. Basic Eng., Ser. D,* **83**, 95-108.

Karlin, S. and Taylor, H. M. (1975). *A First Course in Stochastic Processes*, 2^{nd} ed., Academic Press, New York.

Kaufmann, H. (1987). Regression models for nonstationary categorical time series: Asymptotic estimation theory. *Annals of Statistics*, **15**, 79-98.

Kaufmann, H. (1989). On existence and uniqueness of maximum likelihood estimates in quantal and ordinal response models, *Metrica*, **13**, 291-313.

Keenan, D. M. (1982). A time series analysis of binary data. *Journal of American Statistical Association*, **77**, 816-821.

Kendall, M. G. (1954). Note on bias in the estimation of autocorellation. *Biometrica*, **41**, 403-4.

Kolmogorov, A. (1941). Interpolation und extrapolation von stationaren Zufalligen Folgen. *Bull. Acad. Sci. (Nauk)*, USSR, Ser. Math., **5**, 3-14.

Lang, J. B. and Agresti, A. (1994). Simultaneously modeling joint and marginal distributions of multivariate categorical responses, *Journal of the American Statistical Association*, **89**, 625-632.

Li, W. K. (1991). Testing model adequacy for some Markov regression models for time series, *Biometrica*, **78**, 83-89.

Liang, K.-Y. and Zeger, S. (1986). Longitudinal Data Analysis Using Generalized Linear Models. *Biometrica*, **73**, 13-22.

Luce, R. D. (1959). *Individual Choice Behavior*, Wiley, New York.

MacDonald, I. L. and Zucchini, W. (1997). *Hidden Marcov and Other Models for Discrete-Valued Time Series,* Chapman and Hall, London.

McCullagh, P. (1980). Regression models for ordinal data. *Journal of Royal Statistical Society, Series B*, **42**, 109-142, with discussion.

McCullagh, P. and Nelder, J. A. (1989). *Generalized Linear Models*, 2nd ed., Chapman & Hall, London.

McFadden, D. (1973). Conditional logit analysis of qualitative choice behavior. In P. Zarembka, editor, *Frontiers in Econometrics*, 105-142, Academic Press, New York.

McKenzie, E. (1985). Some simple models for discrete variate time series, *Water Resources Bulletin*, **21**, 645-650.

McKenzie, E. (1986). Autoregressive moving-average processes with negative-binomial and geometric marginal distributions, *Advances in Applied Probability*, **18**, 679-705.

Nelder, J. A. and Wedderburn, R. W. M. (1972). Generalized Linear Models. *Journal of the Royal Statistical Society, Series A*, **135**, 370-384.

Pegram, G. G. S. (1980). An autoregressive model for multilag Marcov chains, *Journal of Applied Probability*, **17**, 350-362.

Pierce, D. A. and Schafer, D. W. (1986). Residuals in generalized linear models, *Journal of the American Statistical Association*, **81**, 977-983.

Pratt, W. J. (1981). Concavity of the log-likelihood. *Journal of the American Statistical Association*, **76**, 103-106.

Priestley, M. B. (1980). State-dependent models: a general approach to non-linear time series analysis. *Journal of Time Series Analysis*, **1**, 47-71.

Priestley, M. B. (1981). *Spectral analysis and Time Series*, vols. I, II. Academic Press, London.

Pruscha, H. (1993). Categorical time series with a recursive scheme and with covariates, *Statistics*, **24**, 43-57.

Raftery, A. E. (1985a). A model for high-order Marcov chains, *Journal of the Royal Statistical Society Series B*, **47**, 528-539.

Rao, C. R. (1973). *Linear Statistical Inference and Its Applications*, 2^{nd} ed., Wiley, New York.

Rijn, P. W. and Molenaar, P. C. M. (2003). State space analysis of categorical time series, ;

Schwarz, G. (1978). Estimating the dimension of a model, *Annals of Statistics.* **6**, 461-464.

Silvapulle, M. J. (1981). On the existence of maximum likelihood estimates for the binomial response models. *Journal of Royal Statistical Society, Series B*, **43**, 310-313.

Sindosi, O. A., Katsoulis, B. D. and Bartzokas, A. (2003). An objective definition of air mass types affecting Athens, Greece; the corresponding atmospheric pressure patterns and air pollution levels, *Enviromental Technology*, **24**, 947-962.

Slud, E. V. (1982). Consistency and efficiency of inferences with the partial likelihood, *Biometrica*, **69**, 547-552.

Snell, E. J. (1964). A scaling procedure for ordered categorical data, *Biometrics*, **20**, 592-607.

Stoffer, D. and Tyler, D. E. (1998). Matcing sequences: Cross-spectral analysis of categorical time series, *Biometrica*, **85**, 201-13.

Stoffer, D., Tyler, D. E. and McDougall, A. J. (1993). Spectral analysis for categorical time series: Scaling and the spectral envelope, *Biometrica*, **80**, 611-22.

Subba Rao, T. and Gabr, M. M. (1984). *An Introduction to Bispectral Analysis and Bilinear Time Series Models*. Springer-Verlag, Berlin.

Subba Rao, T., Priestley, M. B. and Lessi, O. (1997). *Applications of Time Series Analysis in Astronomy and Meteorology*, Chapman & Hall, London.

Ten Have, T. R. (1996) A mixed effects model for multivariate ordinal response data including correlated discrete failure times with ordinal responses, *Biometrics*, **52**, 473-491.

Tong, H. (2001). A personal journey through time series in Biometrica. *Biometrica*, **88**, 1, 195-218.

Tong, H. and Lim, K. S. (1980). Threshold autoregression, limit cycles, and cyclical data. *Journal of the Royal Statistical Society. (B)*, **42**,245-92.

Venables, W. N. and Ripley, B. D. (1999). *Modern Applied Statistics with S-PLUS*, 2^{nd} ed., Springer, New York.

Wedderburn, R. W. M. (1976). On the existence and uniqueness of the maximum likelihood estimates, *Biometrica*, **63**, 27-32.

West, M. and Harrison, P. J. (1997). *Bayesian Forecasting and Dynamic Models*, 2^{nd} ed., Springer, New York.

Wiener, N. (1949). *The extrapolation, Interpolation and Smoothing of Stationary Time Series with Engineering Applications.* Wiley, New York.

Wong, W. H. (1986). Theory of partial likelihood. *Annals of Statistics*, **14**, 88-123.

Yule, G. U. (1927). On a method of investigating periodicities in disturbed series with special reference to Wolfer's sunspot numbers. *Phil. Trans. Roy. Soc. London, Ser. A,* **226**, 267-98.