ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ
ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ
ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ
ΔΠΜΣ ΕΦΑΡΜΟΣΜΕΝΕΣ ΜΑΘΗΜΑΤΙΚΕΣ ΕΠΙΣΤΗΜΕΣ

ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΜΕ ΤΙΣ ΚΑΤΑΝΟΜΕΣ ΤΥΠΟΥ ΦΑΣΕΩΝ

Μεταπτυχιακή Διπλωματική εργασία

ΡΑΠΑΙ ΕΙΡΗΝΗ-ΣΟΝΙΛΑ

ΕΠΙΒΛΕΠΟΥΣΑ ΚΑΘΗΓΗΤΡΙΑ: ΧΡΥΣΗΙΣ ΚΑΡΩΝΗ

ΑΘΗΝΑ, 2011

Αφιέρωνεται...

στον Άγγελο & στο Αγγελούδι μου

Ευχαριστίες

Η διπλωματική αυτή εργασία εκπονήθηκε στα πλαίσια μεταπτυχιακών σπουδών στη Σχολή των Εφαρμοσμένων Μαθηματικών και Φυσικών Επιστημών του Εθνικού Μετσόβιου Πολυτεχνείου στον τομέα της Στατιστικής κατα τη διάρκεια 2009-2011. Με την παρούσα εργασία περατώνονται οι μεταπτυχιακές σπουδές μου.

Έτσι λοιπόν στο σημείο αυτό θα ήθελα να ευχαριστήσω θερμά όλους όσους βοήθησαν και συνετέλεσαν στη διεξαγωγή και στην ολοκλήρωη αυτής της εργασίας. Καταρχήν, θα ήθελα να ευχαριστήσω την επιβλέπουσα καθηγήτρια της διπλωματικής μου εργασίας κ. Καρώνη Χρυσηϊδα, τόσο για την καθοδήγησή της και την πολύτιμη συμβολή της σε κάθε φάση της εκπόνησής της, όσο και για την ανάθεση και την εμπιστοσύνη της εργασίας στο πρόσωπο μου .

Επιπλέον, θα ήθελα να ευχαριστήσω τον Άγγελο, ο οποίος μου συμπαραστάθηκε σε μια πολύ δύσκολη φάση της ζωής μου. Τέλος, θα ήθελα να ευχαριστήσω τους γονείς μου οι οποίοι με στερήσεις, στηρίζουν τις προσπάθειές μου καθ' όλη τη διάρκεια των μαθητικών αλλά και ακαδημαϊκών μου σπουδών.

Περίληψη

Στόχος αυτής της διπλωματικής εργασίας είναι να μελετήσουμε την προσαρμογή των phase type distirbutions σε πραγματικά δεδομένα. Η μοντελοποίηση με αυτές τις κατανομές θεωρείται ευρέως χρήσιμη σε εφαρμογές που σχετίζονται με στοχαστικές διαδικασίες όπως: βιοστατιστική, τηλεπικοινωνίες, ουρές αναμονής, ανάλυση επιβίωσης και αξιοπιστίας. Υπάρχουν πολλοί λόγοι από στατιστικής απόψεως για να χρησιμοποιήσει κανείς τις λεγόμενες phase type distributions στις στοχαστικές διαδικασίες, τις οποίες θα αναλύσουμε διεξοδικά παρακάτω.

Έτσι λοιπόν στο κεφάλαιο 3 προχωράμε σε μια διεξοδική έρευνα για να καταγράψουμε τους τρόπους με τους οποίους μπορούμε να εκτιμήσουμε τις παραμέτρους των κατανομών αυτών. Το συμπέρασμα ύστερα απο τη μελέτη αυτή είναι ότι τελικά σχεδόν όλοι οι αλγόριθμοι βασίζονται σε δύο πολύ δημοφιλείς μεθόδους για την εκτίμηση των παραμέτρων, τη μέθοδο των ροπών και τη μέθοδο μέγιστης πιθανοφάνειας. Επιπλέον κάνουμε μια εκτενής αναφορά στο λεγόμενο EM αλγόριθμο, ο οποίος χρησιμοποιείται για την εκτίμηση των παραμέτρων της κατανομής που προσαρμόζουμε. Ωστόσο ο αλγόριθμος αυτός σε γενικές γραμμές είναι αρκετά περίπλοκος και δεν μπορεί να χρησιμοποιηθεί εύκολα στην πράξη.

Τέλος, η τελευταία ενότητα της εργασίας περιλαμβάνει δύο εφαρμογές, κατά τις οποίες προσαρμόζουμε σε δύο σύνολα δεδομένων την Coxian κατανομή (είναι μια από τις γνωστές κατανομές που ανήκουν στην κατηγορία των phase type distributions). Η εκτίμηση των παραμέτρων θα γίνει εφαρμόζοντας την κλασική μέθοδο μέγιστης πιθανοφάνειας με το στατιστικό πακέτο R.

Abstract

The aim of this thesis is to study the use of phase type distributions in modelling data. These distributions have been used in a wide range of stochastic modelling applications in areas as diverse as telecommunications, finance, biostatistics, queuing theory, drug kinetics and survival analysis. There are many reasons for using phase type distributions in describing stochastic processes, which we analyse in detail.

In Chapter 3 we investigate thoroughly the various methods for estimating the parameters of these distributions. The conclusion from this investigation is that almost all algorithms are based on two very popular methods of estimation, the method of maximum likelihood and the method of moments. In addition, we refer extensively to the so-called EM algorithm, which can be used to estimate the parameters of the phase type distributions. However this algorithm is generally quite complex and can not be applied easily in practice.

Finally, the last section of the work includes two applications, in which we fit a special case of the phase type distributions, the Coxian distribution, to two sets of data. Parameters are estimated by maximum likelihood using the statistical package R.

Περιεχόμενα Σελίδα

Κεφάλαιο 1ο

Εισαγωγή

Σκοπός αυτής της διπλωματικής εργασίας είναι να παρουσιάσουμε τους αλγορίθμους για την εκτίμηση των παραμέτρων μιας ειδικής κατηγορίας κατανομών, των phase type distributions (PHD), με απώτερο στόχο την μοντελοποίηση και την προσαρμογή τους σε περίπλοκα προβλήματα στοχαστικών διαδικασιών. Ειδικότερα, οι phase type distributions περιγράφουν το χρόνο απορρόφησης t μιας Μαρκοβιανής αλυσίδας (συνεχούς ή διακριτού χρόνου), στην οποία υπάρχει μια μόνο κατάσταση απορρόφησης ενώ οι υπόλοιπες καταστάσεις $\rho < \infty$ είναι μεταβατικές (θεωρούμε ότι οι καταστάσεις της Μαρκοβιανής αλυσίδας είναι οι λεγόμενες φάσεις-στάδια της διαδικασίας). Η κατηγορία αυτή είναι μια γενικότερη κατηγορία κατανομών στην οποία ανήκουν γνωστές κατανομές όπως εκθετική, Γάμμα, Erlang, Coxian και άλλες οι οποίες θα αναφερθούν αναλυτικά παρακάτω.

Ύστερα από την εισαγωγή του Neuts (1975), οι phase type κατανομές έχουν χρησιμοποιηθεί ευρέως στον τομέα των στοχαστικών διαδικασιών βρίσκοντας εφαρμογές σε τομείς των εφαρμοσμένων μαθηματικών. Χαρακτηριστικά οι Asmussen, Nerman και Olsson (1996) αναφέρουν ότι:

''... there has been a rapidly growing realization of PH (phase type) distribution as a main computational vehicle of applied probability.''

Η μοντελοποίηση με αυτές τις κατανομές θεωρείται ευρέως χρήσιμη σε εφαρμογές που σχετίζονται με στοχαστικές διαδικασίες. Υπάρχουν πολλοί λόγοι από στατιστικής απόψεως για να χρησιμοποιήσει κανείς τις λεγόμενες PH κατανομές. Πρώτα απ' όλα από υπολογιστική άποψη οι PH κατανομές είναι σαφέστατα πιο ευέλικτες κατανομές, καθώς μπορούν να προσεγγιστούν από κάθε κατανομή η οποία είναι ορισμένη στο $[0,\infty)$, καθώς επίσης η μοντελοποίηση τους οδηγεί σε μοντέλα τα οποία είναι αλγοριθμικά εύκολα υπολογίσιμα.

Η δομή της διπλωματικής εργασίας είναι η ακόλουθη:

- Στο κεφάλαιο 1, παρουσιάζουμε μια μικρή εισαγωγή των phase type distributions.

- Στο κεφάλαιο 2, εισάγουμε τον ορισμό των phase type κατανομών τόσο για τις συνεχής όσο και για τη διακριτή περίπτωση. Υπολογίζουμε την συνάρτηση πυκνότητας πιθανότητας, συνάρτηση κατανομής, ροπές και άλλα. Επιπλέον αναφέρουμε χαρακτηριστικά παραδείγματα δηλαδή κατανομές όπως είναι η εκθετική, αρνητική διωνυμική κατανομή, Erlang κατανομή, Coxian κατανομή οι οποίες έχουν τη μορφή των PH κατανομών.

- Στο κεφάλαιο 3, το οποίο είναι και το ποιο σημαντικό κομμάτι αυτής της εργασίας, αναφέρονται μέθοδοι που χρησιμοποιούμε για την προσαρμογή των κατανομών αυτών σε εμπειρικά δεδομένα. Ειδικότερα αναλύουμε τις

εξής μεθόδους: μέθοδο των ροπών, μέθοθο μέγιστης πιθανοφάνειας όπου σε αυτή τη περίπτωση χρησιμοποιούμε τον αλγόρθμο ΕΜ.

- Στο κεφάλαιο 4, το οποίο αποτελεί την τελευταία ενότητα της εργασίας, υπολογίζουμε με την μέθοδο μέγιστης πιθανοφάνειας τις παραμέτρους της Coxian κατανομής. Οι εκτιμήτριες μέγιστης πιθανοφάνειας θα υπολογιστούν με τη χρήση του στατιστικού πακέτου R.

Κεφάλαιο 2ο

Phase type distributions

Εισαγωγή

Η κλάση των κατανομών τύπου φάσεων (phase type κατανομές) είναι καλά καθιερωμένη στο πεδίο των εφαρμοσμένων μαθηματικών. Μερικά απο τα πεδία δράσης της όπως αναφέραμε στην περίληψη είναι τα δίκτυα επικοινωνίας, η θεωρία κινδύνου, η θεωρία ουρών αναμονής, η θεωρία αξιοπιστίας, η βιοστατιστική κλπ.

Στις τελευταίες δεκαετίες, πολλή έρευνα έχει γίνει σχετικά με τις στοχαστικές διαδικασίες όπου η διάρκεια ακολουθεί phase type κατανομή. Οι phase type κατανομές έχουν εξεταστεί και μελετηθεί πρώτα απο τον Neuts (1975; βλ. και Neuts,1981). Στη συνέχεια ο O'Cinneide (1989, 1990, 1999) παρουσιάζει κάποιες χαρακτηριστικές ιδιότητες αυτών των κατανομών και εκ τότε συνέχεια γίνονται μελέτες πάνω σε αυτές τις κατανομές.

Οι phase type κατανομές περιγράφουν το χρόνο απορρόφησης σε μια Μαρκοβιανή αλυσίδα, η οποία αποτελείται από πεπερασμένο το πλήθος

μεταβατικές καταστάσεις (τις οποίες εμείς πολλές φορές θα τις αναφέρουμε ως φάσεις ή στάδια της Μαρκοβιανή διαδικασίας) και μια μοναδική κατάσταση απορρόφησης. Μερικά παραδείγματα είναι η εκθετική κατανομή, η Erlang κατανομή, η υπερεκθετική κατανομή, η Coxian κατανομή, η Γάμμα κατανομή και άλλες.

Η πρώτη λοιπόν αυτή ενότητα οργανώνεται ως εξής. Άρχικά στην παράγραφο 2.1 αναφέρουμε κάποιες βασικές έννοιες σχετικά με τις Μαρκοβιανές αλυσίδες, με τη βοήθεια των οποίων στην παράγραφο 2.2 θα εισάγουμε τον ορισμό των συνεχών phase type κατανομών. Στη συνέχεια στην επόμενη παράγραφο 2.3 μελετούμε διεξοδικά τις ιδιότητες των PHD δηλαδή υπολογίζουμε συνάρτηση πυκνότητας πιθανότητας, συνάρτηση κατανομής, ροπογεννήτρια συνάρτηση, συνάρτηση Laplace-Stieltjes. Στην ενότητα 2.4 περιλαμβάνονται αρκετά παραδείγματα στις συνεχείς PHD, όπου κάθε παράδειγμα συνοδεύται απο το αντίστοιχο διάγραμμα φάσεως για καλύτερη εποπτεία της κάθε κατανομής. Τέλος αυτό το κεφάλαιο ολοκληρώνεται με τις παραγράφους 2.5 και 2.6 στις οποίες αναλύουμε ακόμη την περίπτωση όπου έχουμε διακριτή PHD και κάποιες ενδεικτικές εφαρμογές που έχουν σημειωθεί μέχρι τώρα στην βιβλιογραφία αντίστοιχα.

2.1 Μαρκοβιανή αλυσίδα

Καταρχήν προτού ορίσουμε τις phase type distributions θεωρώ απαραίτητο σε αυτό το σημείο να κάνουμε μια μικρή εισαγωγή σε κάποιες βασικές ιδιότητες των Markov διαδικασιών με πεπερασμένο χώρο καταστάσεων (πολλές φορές μπορεί να τις συναντούμε στη βιβλιογραφία ως συνεχούς χρόνου Μαρκοβιανή αλυσίδα).

Ορισμός

Μαρκοβιανή αλυσίδα $\{X(t)\}_{t\geq 0}$, με τιμές στον διακριτό χώρο καταστάσεων E, είναι μια στοχαστική διαδικασία η οποία πληρεί την παρακάτω ιδιότητα:

$$P\left(X(t_n)=i_n \middle| X(t_{n-1})=i_{n-1},...,X(t_0)=i_0\right)=P(X(t_n)=i_n \middle| X(t_{n-1})=i_{n-1})$$

Μάλιστα η στοχαστική διαδικασία λέγεται ομοιογενής εαν $P\left(X(t+h)=j \middle| X(t)=i\right)$ εξαρτάται μόνο από το h, όπου σ᾽αυτή την περίπτωση το συμβολίζουμε με p_{ij}^{h}. Ονομάζουμε τον αντίστοιχο πίνακα μετάβασης $P(h)=\left\{P_{ij}^{h}\right\}_{i,j\in E}$

Έστω λοιπόν ότι $\{X(t)\}_{t\geq 0}$ είναι μια Μαρκοβιανή αλυσίδα με πεπερασμένο χώρο καταστάσεων $E=\{1,2,...,p,p+1\}$. Θεωρούμε επιπλέον ότι οι μεταβλητές T_1, T_2….συμβολίζουν το χρόνο που χρειάζεται η Μαρκοβιανή αλυσίδα να μεταβεί από την μια κατάσταση στην άλλη, όπου $T_0=0$. Τότε η διακριτού χρόνου διαδικασία, $\{Y_n\}_{n\in N}$ όπου $Y_n = X(T_n)$, είναι μια Μαρκοβιανή αλυσίδα η οποία παρακολουθεί ποιες καταστάσεις έχουν επισκεφθεί. Συμβολίζουμε με $Q=\{q_{ij}\}$ με $i, j \in E$ τον πίνακα μετάβασης.

Εάν $Y_n=i$, τότε η διαφορά $T_{n+1}-T_n$ ακολουθεί εκθετική κατανομή με συγκεκριμένη παράμετρο λ_i. Επιπρόσθετα εάν συμβολίσουμε με $Y_0=i_0, Y_1=i_1,..., Y_n=i_n$ τότε οι χρόνοι T_1-T_0, $T_2-T_1,...,T_{n+1}-T_n$ είναι ανεξάρτητοι μεταξύ τους. Γνωρίζοντας ότι $T_{n+1}-T_n$ ακολουθεί εκθετική κατανομή με παράμετρο λ_i δοθέντος ότι $Y_n=i$,τότε η δεσμευμένη πιθανότητα όπου θα μεταπηδήσει η Μαρκοβιανή αλυσίδα $\{X(t)\}_{t\geq 0}$ κατά τη διάρκεια του χρονικού διαστήματος

[t, t+dt] είναι $\lambda_i dt$. Συνεπώς έχουμε ότι η πιθανότητα μετάβασης από την κατάσταση i στην κατάσταση j είναι q_{ij}. Ενώ για $i \neq j$ η πιθανότητα μετάβασης από την i στην j στο χρονικό διάστημα [t, t+dt] είναι $\lambda_i dt q_{ij}$. Επομένως για $i \neq j$ έχω $\lambda_{ij} = \lambda_i q_{ij}$. Ορίζουμε $\lambda_{ii} = -\sum_{i \neq j} \lambda_{ii}$ συνεπώς καταλήγουμε στην τελική μορφή του πίνακα $\Lambda = \{\lambda_{ij}\}_{i,j \in E}$, ο οποίος ονομάζεται γεννήτορας της αλυσίδας.

Έστω ότι έχουμε $p_{ij}^t = P(X(t) = j / X(0) = i)$ και ο αντίστοιχος πίνακας μετάβασης $P_u = \{p_{ij}\}_{ij} \in E$. Τότε προκύπτει η εξής σημαντική σχέση ανάμεσα στους P_u και Λ:

$$P_u = \exp(\Lambda u) = \sum_{n=0}^{\infty} \frac{\Lambda^n u^n}{n!}$$

διότι γενικά ισχύει $\exp(A) = \sum_{n=0}^{\infty} \frac{A^n}{n!}$.

όπου μέσω αυτής της μορφής του πίνακα P_u θα υπολογίσουμε στην επόμενη παράγραφο τη συνάρτηση πυκνότητας πιθανότητας.

Οι καταστάσεις της Μαρκοβιανής αλυσίδας μπορούν να ταξινομηθούν ως εξής: Μια κατάσταση i θα την χαρακτηρίζουμε ως περιοδική ή επαναληπτική εάν η i έχει επισκεφθεί από την $\{X(t)\}_{t \geq 0}$ ξανά. Ενώ η i θα ονομάζεται απορρόφησης εάν είναι αδύνατον να μεταπηδήσει έξω από αυτή την κατάσταση ξανά, και για αυτό εάν $q_{ij} = 0$ για όλα τα $i \neq j$ υπονοώντας ότι $\lambda_{ij} = 0$ για όλα τα j.

2.2 Συνεχής phase type κατανομές (Continuous phase type distributions)

Θεωρούμε μια Μαρκοβιανή στοχαστική ανέλιξη $\{X(t)\}_{t\geq 0}$ με πεπερασμένο χώρο καταστάσεων $E=\{1,2,...,p,p+1\}$ όπου οι καταστάσεις 1,2,...,ρ είναι μεταβατικές ενώ η ρ+1 είναι κατάσταση απορρόφησης. Τότε η στοχαστική ανέλιξη $\{X(t)\}_{t\geq 0}$ θα έχει πίνακα μετάβασης πιθανοτήτων της μορφής:

$$Q=\begin{pmatrix} T & t \\ O & 0 \end{pmatrix}$$

όπου Τ είναι ένας πίνακας $\rho\times\rho$ διαστάσεων, ενώ ο t είναι ένας πίνακας στήλη διαστάσεως $\rho\times 1$ και το O είναι ένα πίνακας γραμμή με μηδενικά στοιχεία διαστάσεως $1\times\rho$.

Ας μελετήσουμε λοιπόν τα στοιχεία των υποπινάκων Τ και t. Καταρχήν ο πίνακας Τ μας δείχνει τις πιθανότητες μεταπήδησης μεταξύ των μεταβατικών καταστάσεων 1,2,.....ρ όπου τα μη διαγώνια στοιχεία του είναι μη αρνητικά δηλ. $t_{ij}\geq 0$ για i, j = 1,2,.....ρ ενώ τα διαγώνια στοιχεία του είναι αρνητικές ποσότητες $t_{ii}=-\sum_{\substack{j=0\\ i\neq j}}^{p} t_{ij}$. Το διάνυσμα $t=(t_{10},t_{20},...,t_{p0})$, το οποίο εκφράζει την πιθανότητα μεταπήδησης από τις μεταβατικές καταστάσεις στην κατάσταση απορρόφησης ρ+1, παρατηρούμε, ότι ικανοποιεί μια σχέση της μορφής t = -Te όπου $e=(1,1,...,1)'$.

Ύστερα από την ανάλυση των στοιχείων του πίνακα Q, προχωράμε τώρα στη ανάλυση ιδιοτήτων των phase type distribution δηλαδή στον υπολογισμό της συνάρτησης πυκνότητας-πιθανότητας και συνάρτησης κατανομής, καθώς και στον υπολογισμό της συνάρτησης Laplace-Stieltjes.

Καταρχήν θεωρούμε $\pi = (\pi_1, \pi_2, ..., \pi_p, \pi_{p+1})$ την αρχική κατανομή της Μαρκοβιανής στοχαστικής αλυσίδας $\{ X(t),\ t \geq 0\}$ όπου $\pi_i = P(X_0 = i)$ για i=1,2,…p ενώ $P(X_0 = p+1) = 0$. Έτσι λοιπόν με βάση την παραπάνω εισαγωγική θεωρία των PH κατανομών διατυπώνουμε τον εξής ορισμό :

Ορισμός

Ο χρόνος απορρόφησης $t = \inf\{ t \geq 0 / X_t = p+1\}$ λέμε ότι ακολουθεί phase type distribution και γράφουμε $t \sim CPH(\pi,T)$ (Continuous phase type distribution). Λέμε ότι το ζεύγος παραμέτρων (π, T) αναπαριστά την κατανομή των phase type distribution. Επίσης είναι προφανές ότι, η διάσταση των PH κατανομών συμπίπτει με τη διάσταση των παραμέτρων π,T.

Με βάση αυτά που αναφέραμε στην προηγούμενη παράγραφο ο πίνακας μετάβασης Μαρκοβιανής αλυσίδας ορίζεται ως εξής:

$$P_s = \exp(Qs) = \sum_{n=0}^{\infty} \frac{Q^n s^n}{n!}$$

Θα αποδείξουμε ότι ο πίνακας P_s μπορεί να γραφτεί στη μορφή:

$$P_s = \begin{pmatrix} \exp(Ts) & e - \exp(Ts)e \\ 0 & 1 \end{pmatrix}$$

Απόδειξη

$$\exp(Qs) = I + \sum_{n=1}^{\infty} \frac{Q^n s^n}{n!} = I + \sum_{n=1}^{\infty} \frac{s^n}{n!} \begin{pmatrix} T^n & -T^n e \\ 0 & 0 \end{pmatrix}$$

$$= I + \begin{pmatrix} \sum_{n=1}^{\infty} \frac{s^n}{n!} T^n & -\sum_{n=1}^{\infty} \frac{s^n}{n!} T^n e \\ 0 & 0 \end{pmatrix}$$

$$= \begin{pmatrix} I + \sum_{n=1}^{\infty} \frac{s^n}{n!} T^n & -\sum_{n=1}^{\infty} \frac{s^n}{n!} T^n e \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \exp(Ts) & -\{\exp(Ts)e - Ie\} \\ 0 & 1 \end{pmatrix}$$

$$= \begin{pmatrix} \exp(Ts) & e - \exp(Ts)e \\ 0 & 1 \end{pmatrix}$$

2.3 Ιδιότητες των phase type distributions

Σ' αυτή την ενότητα θα παραθέσουμε μερικές βασικές ιδιότητες των phase type distributions χρησιμοποιώντας την θεωρία πιθανοτήτων.

Έστω ότι με f συμβολίζουμε την συνάρτηση πυκνότητας πιθανότητας η οποία ακολουθεί την PH(π,T). Τότε η ποσότητα f(s)ds μπορεί να ερμηνευτεί ως η πιθανότητα $P\left(t \in [s, s+ds)\right)$. Εαν $t \in [s, s+ds)$ η Μαρκοβιανή αλυσίδα υποθέτουμε ότι θα βρίσκεται στην κατάσταση j σε κάποιο χρόνο s. Εαν υποθέσουμε ότι η αλυσίδα ξεκινάνει απο την κατάσταση i, τότε η πιθανότητα να βρεθεί την χρονική στιγμή s στην κατάσταση j δηλαδή $X(s) = j$ θα είναι $p_{ij}^s = \exp(Ts)_{ij}$. Ενώ εαν έχω $X(s) = j$, τότε η πιθανότητα να μεταπηδήσει στην μοναδική κατάσταση

απορρόφησης $p+1$ στο χρονικό διάστημα $[s, s+ds)$ θα είναι $t_j ds$. Επομένως αποδεικνύουμε την εξής πρόταση:

2.3.1 *Πρόταση*

Έστω ότι $t \sim PH(\pi, T)$ τότε η συνάρτηση πυκνότητας-πιθανότητας f δίνεται από την σχέση:

$$f(s) = \pi \exp(Ts)t, \quad \text{όπου } t = -Te$$

Απόδειξη

$$f(s)ds = P\left(t \in [s, s+ds)\right)$$

$$= \sum_{j=1}^{p} P(t \in [s, s+ds) \,|\, X(s) = j) P\left(X(s) = j\right)$$

$$= \sum_{j=1}^{p} P(\in [s, s+ds) \,|\, X(s) = j) \sum_{i=1}^{p} P\left(X(s) = j \,|\, X(0) = i\right) P\left(X(0) = i\right)$$

$$= \sum_{j=1}^{p} t_j ds \sum_{i=1}^{p} \pi_i \exp(Ts)$$

$$= \sum_{j=1}^{p} \sum_{i=1}^{p} \pi_i \exp(Ts)_{ij} t_j ds$$

$$= \pi \exp(Ts) t ds$$

Τώρα θα προχωρήσουμε στην εύρεση της συνάρτησης κατανομής F(t) της τυχαίας μεταβλητής t. Τότε $1 - F(s)$ είναι η πιθανότητα κατα την οποία η Μαρκοβιανή αλυσίδα $\{ X(t), t \geq 0\}$ δεν έχει φτάσει στην κατάσταση απορρόφησης μέχρι την χρονική στιγμή s, δηλαδή ισχύει ότι $t \succ s$. Άρα επειδή $t \succ s$ σημαίνει ότι $\{X(s) \in \{1, 2, ..., p\}\}$. Επομένως αποδεικνύουμε την εξής πρόταση:

2.3.2 *Πρόταση*

Έστω ότι t ~ PH (π, T) τότε η συνάρτηση κατανομής F δίνεται από τη σχέση:

$$F(t) = 1 - \pi \exp(Tt)e, \; t \succ s$$

Απόδειξη

$$1 - F(s) = P(t \succ s)$$

$$= P(X(s) \in \{1, 2, ..., p\})$$

$$= P\left(\bigcup_{j=1}^{p} (X(s) = j)\right)$$

$$= \sum_{j=1}^{p} P(X(s) = j)$$

$$= \sum_{j=1}^{p} P(X(s) = j \,|\, X(0) = i) P(X(0) = i)$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} p_{ij}^{s} \pi_i$$

$$= \sum_{i=1}^{p} \sum_{j=1}^{p} \pi_i \exp(Ts)_{ij}$$

$$= \pi \exp(Ts) e$$

2.3.3 *Πρόταση*

Έστω t ~ PH(π,T). Τότε:

i) Η νιοστή ροπή της t δίνεται από την σχέση $E(t^n) = (-1)^n \, n! \, \pi T^{-n} e$

ii) Η ροπογεννήτρια δίνεται από την σχέση $E(e^{st}) = \pi(-sI - T)^{-1} t$

Απόδειξη

i) Θα αποδείξουμε το πρώτο μέρος της πρότασης με επαγωγή.

Άρα αρχικά για n=1 έχουμε:

$$E(t) = \int_0^\infty s f(s) ds$$

$$= \int_0^\infty s \pi e^{Ts} t ds$$

$$= -\int_0^\infty \pi e^{Ts} T^{-1} t ds$$

$$= \pi \Gamma^{-2} t$$

$$= \pi T^{-2} t(-Te)$$

$$= -\pi T^{-1} e$$

Έτσι λοιπόν υποθέτουμε ότι $E(t^k) = (-1)^k k! \pi T^{-k} e$ ισχύει για $n = k$ και θα αποδείξουμε ότι ισχύει και για $n = k+1$. Δηλαδή έχουμε τα εξής αποτελέσματα:

$$E(t^{k+1}) = \int_0^\infty s^{k+1} f(s) ds$$

$$= \int_0^\infty s^{k+1} \pi e^{Ts} t ds$$

$$= -\int_0^\infty (k+1) s^k \pi e^{Ts} T^{-1} t ds$$

$$= -(k+1) T^{-1} \int_0^\infty s^k \pi e^{Ts} t ds$$

$$= -(k+1) T^{-1} (-1)^k k! \pi \Gamma^{-\kappa} e$$

$$= (-1)^k (k+1)! \pi \Gamma^{-(k+1)} e$$

Επιπλέον, η μέση τιμή των PH κατανομών για κ=1 με παραμέτρους (π, T) είναι:

$$E(t) = m_1 = -\pi T^{-1} e$$

ενώ η διασπορά του είναι:

$$V(t) = \mathrm{E}\left(t^2\right) - \left[\mathrm{E}(t)\right]^2 = \; m_2 - {m_1}^2 = 2\pi\Gamma^{-2}e - \left(\pi T^{-1}e\right)^2$$

ii) Προχωράμε τώρα στο δεύτερο μέρος της απόδειξης, στη λεγόμενη συνάρτηση ροπογεννήτρια.

$$\mathrm{E}\left(\mathrm{e}^{\mathrm{st}}\right) = \int_0^\infty \mathrm{e}^{\mathrm{sx}} \mathrm{f}(\mathrm{x})\mathrm{dx}$$

$$= \int_0^\infty \mathrm{e}^{\mathrm{sx}} \pi \mathrm{e}^{\mathrm{Tx}} \mathrm{tdx}$$

$$= \int_0^\infty \pi \mathrm{e}^{\mathrm{sx}} \mathrm{Ie}^{\mathrm{Tx}} \mathrm{tdx}$$

$$= \int_0^\infty \pi \mathrm{e}^{\mathrm{sxI}} \mathrm{e}^{\mathrm{Tx}} \mathrm{tdx}$$

$$= \int_0^\infty \pi \mathrm{e}^{(\mathrm{sxI}+\mathrm{T})\mathrm{x}} \mathrm{tdx}$$

$$= \pi\left(-\mathrm{sI} - \mathrm{T}\right)^{-1} \mathrm{t}$$

2.3.4 *Πρόταση*

Έστω οτι t ~ PH (π, T) τότε η συνάρτηση Laplace-Stieltjes που δίνεται γενικά από τη σχέση $\mathrm{L}(\mathrm{s}) = \mathrm{E}\left(\mathrm{e}^{-\mathrm{st}}\right) = \pi\left(\mathrm{sI} - \mathrm{T}\right)^{-1} \mathrm{t}$.

Απόδειξη

Έστω $\mathrm{k} \geq 0$, τότε:

$$L(s) = \int_0^\infty \exp(Ts)\mathrm{ds}$$

$$= \int_0^k \sum_{\mathrm{i}=0}^{\infty} \frac{(\mathrm{Ts})^{\mathrm{i}}}{\mathrm{i}!}\mathrm{ds}$$

$$= \sum_{i=0}^{\infty} T^i \frac{k^{i+1}}{(i+1)!}$$

$$= T^{-1}\left(e^{Tk} - I\right) \xrightarrow[k\to\infty]{} (-T)^{-1}$$

2.4 Παραδείγματα των συνεχών PH κατανομών

Σ' αυτή την ενότητα θα παραθέσουμε μερικά παραδείγματα των PH-κατανομών για την καλύτερη κατανόηση της παραπάνω εισαγωγικής θεωρίας. Μια εύχρηστη γραφική απεικόνιση μιας κατανομής τύπου φάσεων είναι το λεγόμενο διάγραμμα τύπου φάσεων, το οποίο δίνεται σε σχέση με τις πιθανότητες εισόδου, τις πιθανότητες εξόδου και τους ρυθμούς μετάβασης. Παραδείγματα κατανομών τα οποία μπορούν να γραφούν ως τύπου φάσεων κατανομές είναι:

Παράδειγμα 2.4.1 Εκθετική κατανομή

Το πιο απλό παράδειγμα των PH-κατανομών είναι η εκθετική κατανομή.

Όπως γνωρίζουμε η εκθετική κατανομή έχει συνάρτηση πυκνότητας πιθανότητας:

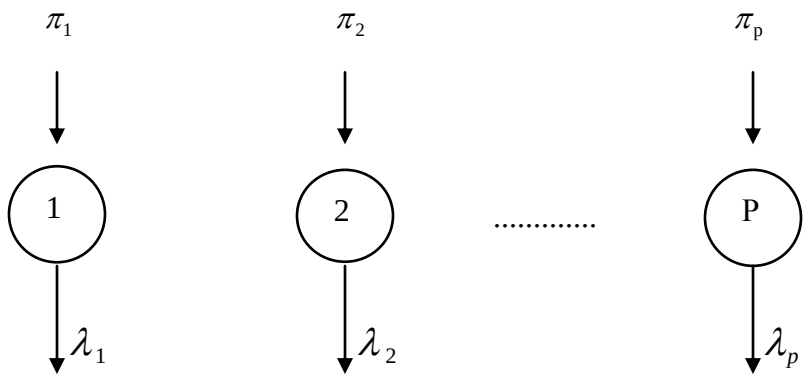
$$f(t) = \lambda e^{(-\lambda t)} \quad \text{με} \quad t > 0$$

Άρα, από τη μορφή της συνάρτησης πυκνότητας πιθανότητας της εκθετικής κατανομής έχουμε ότι $\pi = (1)$ ενώ το $T = (-\lambda)$

Παράδειγμα 2.4.2 Υπερεκθετική κατανομή

Η υπερεκθετική κατανομή αποτελείται από ρ καταστάσεις παράλληλες μεταξύ τους και ορίζεται ως ενας γραμμικός συνδυασμός p εκθετικών κατανομών με ρυθμούς $\lambda_1, \lambda_2, ..., \lambda_p$ όπου χωρίς βλάβη της γενικότητας θεωρούμε ότι $0 < \lambda_1 < \lambda_2 < \cdots < \lambda_\rho$.

Η κατανομή αυτή μπορεί να κατανοηθεί καλύτερα χρησιμοποιώντας το παρακάτω διάγραμμα φάσεων:



Τότε οι αντίστοιχες παράμετροι θα είναι:

$$\pi = (\pi_1, \pi_2, ..., \pi_p) \quad \text{και} \quad T = \begin{pmatrix} -\lambda_1 & 0 & \cdots & 0 \\ 0 & -\lambda_2 & \ddots & 0 \\ \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & \cdots & -\lambda_\rho \end{pmatrix}$$

PDF	$f(u) = \sum_{i=1}^{p} p_i \lambda_i e^{-\lambda_i u}$
CDF	$F(u, p, \lambda) = 1 - \sum_{i=1}^{p} p_i \lambda_i e^{-\lambda_i u}$
GF	$H(u, p, \lambda) = \sum_{i=1}^{p} \frac{p_i \lambda_i}{s + \lambda_i}$
Ροπές	$\mu_i(p, \lambda) = i! \sum_{i=1}^{p} \frac{p_i}{\lambda_i^i}$

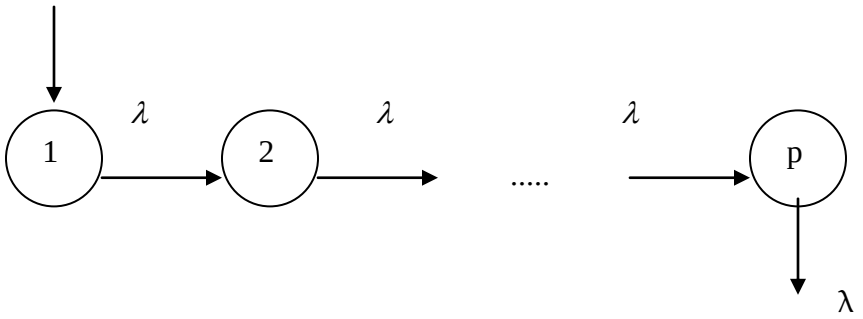
Πίνακας 2.4.1 Συνάρτηση πυκνότητας πιθανότητας (PDF), Συνάρτηση κατανομής (CDF), Ροπογεννήτρια συνάρτηση (GF), Ροπές (μ_i)

Παράδειγμα 2.4.3 Erlang κατανομή με ρ-φάσεις

Η Erlang κατανομή με ρ-φάσεις ορίζεται σαν γάμμα κατανομή με ακέραια παράμετρο p και πυκνότητα

$$f(u,p,\lambda)=\frac{\lambda^{p}u^{p-1}e^{-\lambda u}}{p!}$$

Ο παραπάνω τύπος αντιστοιχεί στη συνέληξη p εκθετικών κατανομών με τον ίδιο ρυθμό λ. Επομένως η κατανομή αυτή μπορεί να κατανοηθεί καλύτερα εάν χρησιμοποιήσουμε το παρακάτω διάγραμμα το οποίο αποτελείται από ρ καταστάσεις συνδεδεμένες στη σειρά.



συνεπώς λαμβάνοντας υπόψη το παραπάνω διάγραμμα θα έχει παραμέτρους:

$$\pi=(1,0,\ldots,0) \qquad \text{και} \qquad T=\begin{pmatrix} -\lambda & \lambda & 0 & \cdots & 0 \\ 0 & -\lambda & \lambda & \cdots & 0 \\ 0 & 0 & -\lambda & \ddots & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda \end{pmatrix}$$

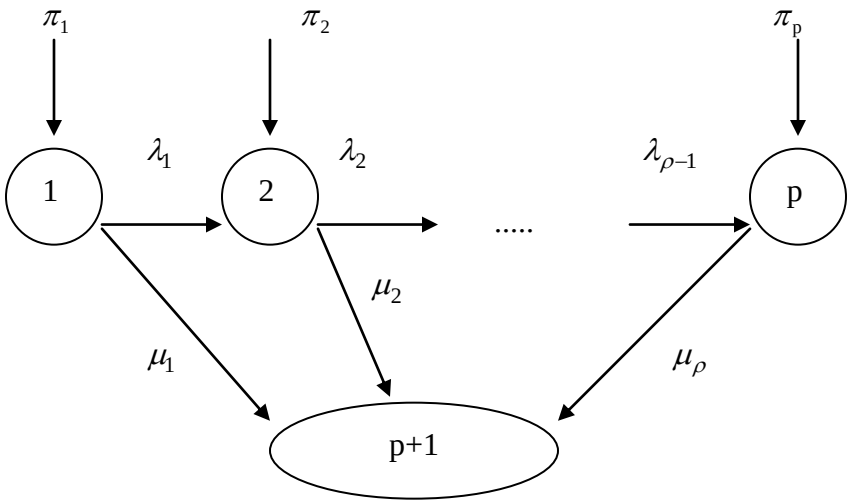
Η Erlang κατανομή έχει τα εξής χαρακτηριστικά:

PDF	$f(u,p,\lambda)=\frac{\lambda^p u^{p-1} e^{-\lambda u}}{p!}$
CDF	$F(u,p,\lambda)=\sum_{i=p}^{\infty}\frac{(\lambda u)^i}{i!}e^{-\lambda u}$
GF	$H(u,p,\lambda)=\left(\frac{\lambda}{u+\lambda}\right)^p$
Ροπές	$\mu_i(p,\lambda)=\frac{(i+p-1)!}{(p-1)!\lambda^i}$

Πίνακας 2.4.2 Συνάρτηση πυκνότητας πιθανότητας (PDF), Συνάρτηση κατανομής (CDF), Ροπογεννήτρια συνάρτηση (GF), Ροπές (μ_i)

Παράδειγμα 2.4.4 Coxian κατανομή

Η κλάση των Coxian κατανομών είναι δημοφιλής στη βιβλιογραφία των εφαρμοσμένων επιστημών. Η Coxian κατανομή ανήκει στη κλάση των κατανομών τύπου φάσεων με διάγραμμα φάσης της μορφής:



Απεικόνιση της Coxian κατανομής

Στον πίνακα αυτών όπου λ_i είναι οι διαδοχικές πιθανότητες μετάβασης ανάμεσα στις καταστάσεις από τις $1,...,p$ ενω με μ_i συμβολίζονται οι πιθανότητες απο τις μεταβατικές καταστάσεις $1,...,p$ στην μοναδική κατάσταση απορρόφησης $p+1$.
Έτσι παρατηρώντας προσεκτικά το παραπάνω διάγραμμα λέμε ότι η Coxian κατανομή είναι μια ειδική περίπτωση των PH κατανομών όπου οι παράμετροι θα έχουν την μορφή:

$$\pi = \left(\pi_1, \pi_2, ..., \pi_p\right)$$

$$T = \begin{pmatrix} -(\lambda_1+\mu_1) & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -(\lambda_2+\mu_2) & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\lambda_p+\mu_p) & \lambda_p \\ 0 & 0 & 0 & \cdots & 0 & -\mu_{p+1} \end{pmatrix}$$

Παράδειγμα 2.4.5 Η μη-κυκλική PH κατανομή με ρ-φάσεις
Η μη-κυκλική κατανομή θα έχει παραμέτρους:

$$\pi = \left(\pi_1, \pi_2, ..., \pi_p\right) \quad \text{και} \quad T = \begin{pmatrix} -\lambda_1 & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -\lambda_2 & \lambda_2 & \cdots & 0 & 0 \\ 0 & 0 & -\lambda_3 & \ddots & 0 & 0 \\ \vdots & \vdots & \ddots & \ddots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -\lambda_{\rho-1} & \lambda_{\rho-1} \\ \mu_1 & \mu_2 & \mu_3 & \cdots & \mu_{\rho-1} & -\lambda_\rho \end{pmatrix}$$

όπου για $i = 1,2,\ldots p\text{-}1$, $\mu_i \geq 0$, $0 < \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_\rho$ και $\lambda_p > \sum_{i=1}^{p-1} \mu_i$.

Γενικά, θα πρέπει να τονίσουμε ότι η μορφή των PH κατανομών δεν είναι μοναδική. Συγκεκριμένα ας παρατηρήσουμε προσεκτικά το παρακάτω παράδειγμα το οποίο αναφέρεται στους Botta, Harris και Marchal (1987). Έστω ότι έχουμε μια τυχαία μεταβλητή η οποία ακολουθεί την PH κατανομή με συνάρτηση πυκνότητας πιθανότητας:

$$f(u) = \frac{2}{3}e^{-2t} + \frac{1}{3}e^{-5t}$$

τότε το ζεύγος των παραμέτρων (α, T) μπορεί να πάρει τις εξής 3 μορφές:

$$\alpha = \begin{pmatrix} \frac{1}{3} & \frac{2}{3} \end{pmatrix} \qquad T = \begin{pmatrix} -5 & 0 \\ 0 & -2 \end{pmatrix}$$

$$b = \begin{pmatrix} \frac{1}{5} & \frac{4}{5} \end{pmatrix} \qquad S = \begin{pmatrix} -2 & 2 \\ 0 & -5 \end{pmatrix}$$

$$g = \begin{pmatrix} 0 & \frac{1}{2} & \frac{1}{2} \end{pmatrix} \qquad R = \begin{pmatrix} -3 & 1 & 1 \\ 1 & -4 & 2 \\ 1 & 0 & -6 \end{pmatrix}$$

2.5 Διακριτές phase type κατανομές (Discrete phase type distributions)

Οι διακριτές phase type κατανομές περιγράφουν το χρόνο απορρόφησης μιας Μαρκοβιανής αλυσίδας διακριτού χρόνου. Οι DPH (Discrete phase type distributions) κατανομές ορίζονται θεωρώντας μια Μαρκοβιανή αλυσίδα με πίνακα μετάβασης πιθανοτήτων της μορφής:

$$P = \begin{pmatrix} T & t \\ 0 & 1 \end{pmatrix}$$

Όπου ο πίνακας Τ είναι ένας στοχαστικός πίνακας, τέτοιο ώστε $I-T$ να μην είναι μοναδιαίος. Ειδικότερα, θεωρούμε μια Μαρκοβιανή αλυσίδα $\{X(n)\}_{n\geq 0}$ με αντίστοιχο χώρο καταστάσεων $E=\{1,...,p,p+1\}$, όπου οι καταστάσεις $1,2,...,p$ είναι μεταβατικές ενω η p+1 είναι κατάσταση απορρόφησης. Έστω ότι $\pi_i = P(X(0)=i)$ δηλώνει την αρχική κατανομή και t_{ij} δηλώνουν τις πιθανότητες μετάβασης $P(X(n+1)=j|X(n)=i)$ για $i,j=1,...,p$. Έστω λοιπόν ότι $\pi=(\pi_1,\pi_2,...,\pi_p)$ είναι το αρχικό διάνυσμα, $\mathbf{T}=\{\mathbf{t}_{ij}\}_{i,j=1,...,p}$ είναι ο πίνακας μετάβασης ανάμεσα στις καταστάσεις και τέλος το διάνυσμα $t=e-Te$ περιλαμβάνει τις πιθανότητες μεταβάσεις στην κατάσταση απορρόφησης.

Ορισμός

Ο χρόνος απορρόφησης $t=\inf\{n\geq 1|X(n)=p+1\}$ λέμε ότι ακολουθεί phase type distribution και γράφουμε $t \sim DPH(\pi,T)$.

2.5.1 *Πρόταση*

Η συνάρτηση πυκνότητας πιθανότητας δίνεται από τη σχέση:

$$f(x)=\pi T^{x-1}t, \quad x\geq 1$$

Απόδειξη

Η πιθανότητα ότι η Μαρκοβιανή αλυσίδα βρίσκεται σε μια από τις μεταβατικές καταστάσεις $i\in\{1,...,p\}$ μετά απο ν-βήματα θα δίνεται από τη σχέση:

$$p_i^{(n)} = P(X(n)=i)\sum_{k=1}^{p}\pi_k\left(T^n\right)_{(k,i)}$$

$$f(n)=P(t=n)=\sum_{i=1}^{p}p_i^{(n-1)}t_i=\pi T^{n-1}t, \quad n\in N$$

2.5.2 *Πρόταση*

Η συνάρτηση κατανομής των διακριτών phase type τυχαίας μεταβλητής δίνεται απο τη σχέση:

$$F(n) = 1 - \pi T^{n} e$$

Απόδειξη

Εξετάζουμε τώρα την πιθανότητα κατα την οποία η απορρόφηση δεν έχει πραγματοποιηθεί και συνεπώς η Μαρκοβιανή αλυσίδα είναι σε μια απο τις μεταβατικές καταστάσεις. Προκύπτει λοιπόν ότι:

$$\begin{aligned} 1 - F(n) &= P(t > n) \\ &= \sum_{i=1}^{p} p_i^{(n)} \\ &= \pi T^{n} e \end{aligned}$$

Στη συνέχεια θα αποδείξουμε ποια είναι η μορφή της πιθανογεννήτριας της t, η οποία ορίζεται ως εξής:

$$G_t(z) = E(z^t) = \sum_{k=0}^{\infty} z^k f(k)$$

2.5.3 *Πρόταση*

Η πιθανογεννήτρια συνάρτηση θα είναι της μορφής:

$$G_t(z) = z\pi(1 - zT)^{-1} t$$

Απόδειξη

$$\begin{aligned} E(z^t) &= \sum_{k=0}^{\infty} z^k f(k) \\ &= \sum_{k=0}^{\infty} z^k \pi T^{k-1} t \\ &= \pi T^{-1} \sum_{k=0}^{\infty} (zT)^k t \end{aligned}$$

$$= \pi T^{-1}\left(\frac{zT}{1-zT}\right)t$$

$$= z\pi\left(1-zT\right)^{-1}t$$

2.5.4 Πρόταση

Η παραγοντική ροπή δίνονται από τη σχέση:

$$G_t^{(k)}(1) = k!\pi T^{k-1}\left(I-T\right)^{-k}e$$

Απόδειξη

$$G_t^{(k)}(1) = \frac{d^k}{dz^k}\Big|_{z=1} G_t(z)$$

$$= k!\pi T^{k-1}\left(I-T\right)^{-k}e$$

Αναφέρουμε χαρακτηριστικά κάποια παραδείγματα στην περίπτωση που έχουμε διακριτή τυχαία μεταβλητή.

Παράδειγμα 2.5.1 Γεωμετρική κατανομή

Έστω $X \sim geo(p)$ με $p \in (0,1)$

Έτσι εαν $P(X=x) = (1-p)^{1-x}p$ είναι η συνάρτηση πυκνότητας πιθανότητας της γεωμετρικής κατανομής, τότε ανήκει στην κλάση των DPH κατανομών μιας με αντίστοιχες παραμέτρους:

$$\pi = (1),\ T = (1-p),\ t = (p)$$

Παράδειγμα 2.5.2 Αρνητική διωνυμική κατανομή

Έστω $X \sim NB(k,p)$, με $p \in (0,1)$ και $k \geq 0$

Η τυχαία μεταβλητή X εκφράζει το άθροισμα των k τυχαίων γεωμετρικών κατανομών με παράμετρο p. Επομένως $P(X=x)=\binom{x+k-1}{k-1}(1-p)^k p^k$ για $x=0,1,...$ Τότε η τυχαία μεταβλητή X έχει τη μορφή των DPH με αντίστοιχες παραμέτρους:

$$\pi=(1,0,...,0),\quad T=\begin{pmatrix}1-p & p & & \\ & 1-p & p & \\ & & \ddots & \\ & & 1-p & p \\ & & & 1-p\end{pmatrix},\quad t=\begin{pmatrix}0\\0\\\vdots\\0\\p\end{pmatrix}$$

2.6 Ανασκόπηση της βιβλιογραφίας:

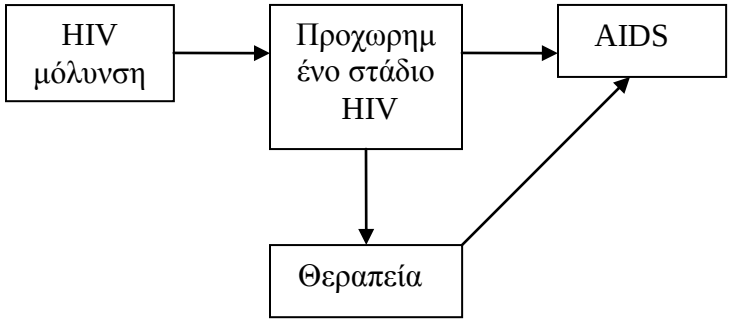
Εφαρμογές των phase type distributions

Οι phase type κατανομές έχουν αποτελέσει σημαντικό κομμάτι της εφαρμοσμένης θεωρίας πιθανοτήτων. Σ' αυτή την παράγραφο θα παραθέσουμε μια σειρά από άρθρα στα οποία αναφέρονται μελέτες σχετικές που έχουν γίνει όσον αφορά την εφαρμογή των phase type distributions. Έτσι παραθέτουμε τα παρακάτω άρθρα. Για άλλα παραδείγματα εφαρμογών, βλέπετε Bladt (2005), Fackrell (2009), Faddy & McClean(1999, 2005).

2.6.1 Η χρήση των phase type distributions στην ανάλυση επιβίωσης από τον Olsson Aalen

Στο άρθρο του ο Aalen (1995) αναφέρει χαρακτηριστικά ότι οι phase type distributions βρίσκουν ευρεία εφαρμογή στη βιοστατιστική. Η δικιά του έρευνα αναφέρει ποικίλες μορφές των phase type μοντέλων τα οποία τα συνδέει με προβλήματα της ανάλυσης επιβίωσης. Γενικά αναφέρει ότι τα phase type μοντέλα έχουν χρησιμοποιηθεί κυρίως στην βιοστατιστική και όχι μόνο. Χαρακτηριστικό παράδειγμα αποτελεί η μελέτη της επώασης του AIDS, η διάρκεια των γεννητικών αλλοιώσεων λόγω του έρπητα και άλλα. Συγκεκριμένα η περιγραφή του μοντέλου για το χρόνο επώασης του AIDS φαίνεται στην εικόνα 2.5.1. Με μια πρώτη ματιά παρατηρούμε ότι το μοντέλο έχει δύο φάσεις για την ανάπτυξη της προχωρημένης νόσου HIV. Σ' αυτό το στάδιο θεραπείας όπως βλέπουμε και στην εικόνα περιγράφεται με έναν ρυθμό γ. Το αποτέλεσμα της θεραπείας είναι να επιβραδύνει την περαιτέρω εξέλιξη του AIDS από τον παράγοντα θ. Ο χρόνος επώασης σε αυτό το μοντέλο είναι ο χρόνος για να πάμε από την κατάσταση 1 (HIV μόλυνση) στην κατάσταση 5 (AIDS).

Το μοντέλο για την περιγραφή του χρόνου επώασης του AIDS είναι ένα παράδειγμα των phase type distributions, στην οποία καμία κατάσταση δεν μπορεί να επισκεφθεί περισσότερο από μια φορά. Ενδεικτικά η αναπαράσταση του παραπάνω προβλήματος σε ένα απλό διάγραμμα φάσης είναι η ακόλουθη:



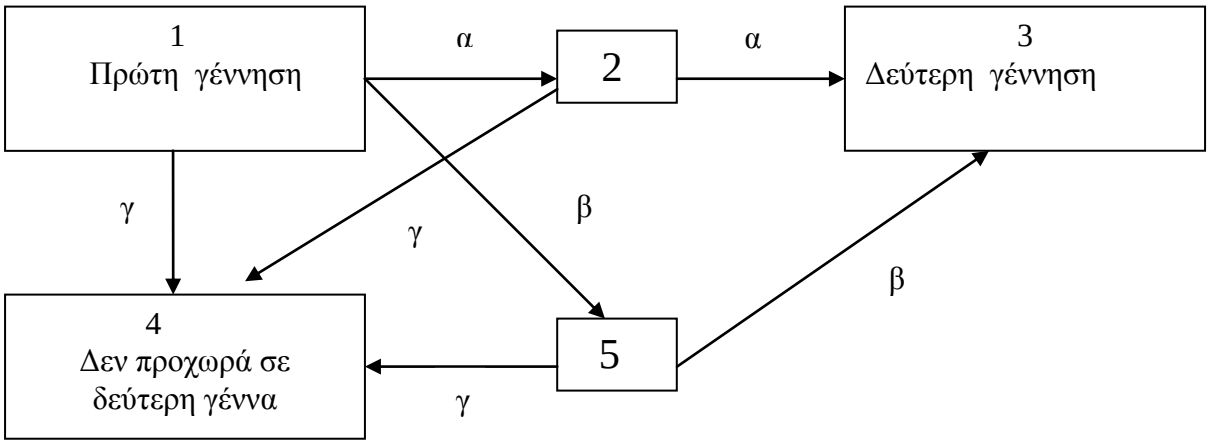
Εικόνα 2.6.1 Το μοντέλο phase type για τον χρόνο επώασης του AIDS

Συχνά, στις εφαρμογές της βιοστατιστικής αυτή η αδυσώπητη εξέλιξη σε μια μόνο κατεύθυνση θα ήταν εξωπραγματική, και θα ήταν φυσικό να υποθέσουμε ότι η διαδικασία κινείται πίσω-μπρος ανάμεσα στις καταστάσεις ακόμα και όταν υπάρχει η κατάσταση απορρόφησης. Τέτοια μοντέλα ονομάζονται models with feedback. Είναι πάντα δυνατό να φτάσουμε από ένα μοντέλο με ανάδραση σ' ένα συγκεκριμένο είδος του λεγόμενου acyclic model ή series model, όπου όλες οι καταστάσεις είναι διατεταγμένες και η μεταφορά μπορεί να γίνει μόνο στην επόμενη κατάσταση έχοντας την τελευταία κατάσταση ως κατάσταση απορρόφησης.

2.6.2 Η χρήση των phase type κατανομών στη μελέτη του χρόνου γεννήσεων

Επιπλέον θα αναφέρουμε ένα άλλο παράδειγμα στο οποίο μπορούμε να προσαρμόσουμε τις λεγόμενες phase type distributions σε πραγματικά δεδομένα. Το παράδειγμα αυτό αφορά γυναίκες που έχουν γεννήσει το πρώτο τους παιδί και ο σκοπός της ανάλυσης είναι να μελετήσουμε το χρόνο μέχρι τη γέννηση του επόμενου παιδιού, εάν τυχόν συμβεί. Η στατιστική ανάλυση και η μελέτη του χρόνου των γεννήσεων είναι σημαντική καθώς βρίσκει εφαρμογή στην έρευνα του δημογραφικού προβλήματος. Έτσι λοιπόν πάρθηκε ένα τυχαίο δείγμα από γυναίκες που ήταν παντρεμένες, οι οποίες είχαν ένα παιδί κατά την περίοδο 1967-1971. Ο συνολικός αριθμός των γυναικών που πήραν μέρος στο δείγμα ήταν 1779 και ανήλθε στις 1982. Ο αριθμός των γυναικών που είχαν γεννήσει και ο αριθμός των γυναικών που δεν ήξεραν εάν είχαν άλλο παιδί ή δεν είχαν, σημαίνει ότι αυτοί οι αριθμοί ήταν αποκομμένοι, και είχαν καταχωρηθεί σ'ένα διάστημα έξι μηνών ξεκινώντας εννέα μήνες ύστερα από την πρώτη τους γέννα. Εάν προσαρμόσουμε το λεγόμενο phase type model στο παραπάνω πρόβλημα θα χρειαστεί να κάνουμε

κάποιες υποθέσεις. Καταρχήν μια πρώτη υπόθεση είναι ότι οι γυναίκες θα πρέπει να περιμένουν λίγο προτού προσπαθήσουν να αποκτήσου ένα άλλο παιδί. Για αυτό το λόγω το μοντέλο μας θα περιέχει το λιγότερο τρεις φάσεις, όπου η μεταφορά από την πρώτη μετάβαση στην δεύτερη κατάσταση πιθανών να σημαίνει ότι το ζευγάρι είναι έτοιμο να κάνει ένα δεύτερο παιδί, ενώ η μετάβαση από την δεύτερη κατάσταση στην τρίτη αναπαριστά την γέννηση του επόμενου παιδιού. Άλλη μια υπόθεση θα ήταν η ιδιότητα της ετερογένειας για κάθε γυναίκα ξεχωριστά. Μερικές γυναίκες μπορεί να γεννήσουν σχετικά γρήγορα, ενώ για άλλες μπορεί να πάρει πολλά χρόνια. Τέλος, θα πρέπει να ενσωματώσουμε στο μοντέλο μας την δυνατότητα κάποιες γυναίκες να μην μπορούν να κάνουν παιδιά για προσωπικού λόγους η λόγω κάποιων προβλημάτων υγείας. Άρα ενσωματώσουμε μια κατάσταση απορρόφησης στο χώρο καταστάσεων. Έτσι λοιπόν το μοντέλο είναι:



Εικόνα 2.6.2 Το μοντέλο phase type για το διάστημα των γεννήσεων

2.6.3 Η χρήση των phase type κατανομών στον τομέα της υγειονομικής περίθαλψης

Πραγματικά έχουν γίνει παρα πολλές εργασίες ως προς την εφαρμογή των phase type distributions στη βιβλιογραφία της υγειονομικής περίθαλψης. Όμως μελετώντας προσεκτικά αυτές τις έρευνες, διαπιστώνουμε ότι οι τομείς στους οποίους οι επιστήμονες έχουν ασχοληθεί θα λέγαμε ότι είναι πολύ περιορισμένος. Οι περισσότερες έρευνες αφορούν τη μοντελοποίηση του χρόνου παραμονής των ασθενών σε ένα γηριατρικό νοσοκομείο, και όλα αυτά τα άρθρα έχουν γραφτεί από ένα σχετικά ''φτωχό'' ερευνητικό επίπεδο. Σε αυτό το σημείο σας παρουσιάζω ύστερα από μια ανασκόπηση της βιβλιογραφίας όλες τις έρευνες που έχουν γίνει στον τομέα της υγείας χρησιμοποιώντας τις phase type distributions.

Ο Faddy (1990) χρησιμοποιήσε δύο μοντέλα θαλάμου για να μοντελοποιήση της εκροής των ερυθρών αιμοσφαιρίων μέσω μιας ένεσης σ'έναν αρουραίο. Ο κάθε θάλαμος αντιπροσώπευε ένα όργανο του σώματος, και ο χρόνος παραμονής των ερυθρών αιμοσφαιρίων στο σώμα προτού γίνει η εκρροή ακολουθούσε την Erlang κατανονή.

Στη συνέχεια στο άρθρο του Faddy (1993) εισήγαγε πιο περίπλοκα δυδιάστα συστήματα θαλάμων, τα οποία χρησιμοποιήθηκαν για να μοντελοποιήσουν τον χρόνο παρακράτησης ενος φαρμάκου μεσω ένεσης σ'ενα όργανο. Στον πρώτο θάλαμο το φάρμακο διαδίδονταν κυκλικά, ενω στο δεύτερο θάλαμο πραγματοποιούνταν κάθαρση του φαρμάκου από το σώμα. Το μοντέλο αυτό εφαρμόστηκε για ένα αντιβιοτικό σε τέσσερα πρόβατα για την καταπολέμηση της νεφρικής ανεπάρκειας, όπου βέβαια δόθηκαν διαφορετικές δώσεις σε κάθε πρόβατο.

Κεφάλαιο 3ο

Εκτίμηση των παραμέτρων των phase type distributions

Εισαγωγή

Οι ενδιαφέρουσες ιδιότητες και η ευρέως χρήση σε εφαρμογές των phase type distributions έχουν εμπνεύσει πολλούς ερευνητές-μελετητές να συνεχίσουν να αναζητούν μεθόδους για να επιλύσουν το πρόβλημα της εκτίμησης των παραμέτρων. Έτσι λοιπόν σε αυτή την ενότητα βασικός στόχος και σκοπός είναι να αναφέρουμε τους τρόπους για την εκτίμηση των παραμέτρων των phase type distributions. Η εκτίμηση των παραμέτρων των phase type distributions δεν είναι μια τετριμμένη διαδικασία. Με μια ανασκόπηση στη βιβλιογραφία αναφέρεται ότι υπάρχουν αρκετά προβλήματα για την προσαρμογή των πραγματικών δεδομένων

μέσω των phase type distributions. Οι δυσκολίες υπάρχουν λόγω της μη γραμμικότητας του προβλήματος, καθώς την ταυτόχρονη βελτίωση του αριθμού των παραμέτρων και της μη μοναδικότητας της μορφής των PH-κατανομών. Αποτέλεσμα όλων αυτών είναι να μην μπορούμε να βρούμε μια ακριβής λύση επομένως έγκειται απαραίτητη η χρήση των αλγορίθμων (αριθμητικών μεθόδων). Βέβαια η εφαρμογή αυτών των αριθμητικών αλγορίθμων για την εκτίμηση των παραμέτρων αφήνουν μερικές φορές ένα ανοιχτό πρόβλημα σύγκλισης όσον αφορά τη χρήση των PH-κατανομών σε εφαρμογές.

Έτσι λοιπόν για να ξεπεράσουμε τέτοιες δυσκολίες, πολλοί ερευνητές έχουν θέσει κάποιους περιορισμούς οι οποίοι θα πρέπει να ικανοποιούνται (Asmussen et al. 1996, Augustin and Buscher 1982, Faddy 1994) όταν θέλουμε να εφαρμόσουμε αυτούς τους αλγορίθμους σε πραγματικά δεδομένα. Με την πάροδο των χρόνων έχουν αναπτυχθεί πολλοί τρόποι για την επίτευξη μιας ιδανικής λύσης. Γενικά μπορούμε να αναφέρουμε ότι υπάρχουν τρεις βασικές τεχνικές οι οποίες εξελίσσονται συνεχώς με την πάροδο των χρόνων. Οι μέθοδοι εκτίμησης των παραμέτρων βασίζονται στις εξής τεχνικές: μέθοδος μέγιστης πιθανοφάνειας (**maximum likelihood**), μέθοδος των ροπών (**moment matching**) καθώς και η μέθοδος ελαχίστων τετραγώνων (**method of least squares**). Προχωράμε τώρα στην περαιτέρω ανάλυση των τεχνικών αυτών.

Όπως ανέφερα και παραπάνω το πρώτο σετ των μεθόδων βασίζεται στην μεγιστοποίηση της λογαριθμικής συνάρτησης

$$-\sum_{i=1}^{n}\log[\,p\exp\{Qt_i\}q] \qquad (1)$$

Σε αυτή την τεχνική βασίστηκε ο Asmussen το 1996 ο οποίος πρότεινε μια διαδικασία προσαρμογής για ολόκληρη την οικογένεια των PH κατανομών, στηριζόμενος στον EM αλγόριθμο των Dempster et al. (1977) για ελλιπή δεδομένα, ανέπτυξε το EMPht πρόγραμμα στη γλώσσα προγραμματισμού C για την εκτίμηση

των παραμέτρων. Ο Faddy (1994, 1998) παρουσίασε τον Nelder και Mead αλγόριθμο χρησιμοποιώντας τη γλώσσα προγραμματισμού Matlab για να επιλύσει την εξίσωση (1) δηλαδή να ελαχιστοποιήσει τη συνάρτηση πιθανοφάνειας. Έπιτα οι Bobbio και Cumani (1992) ανέπτυξαν ένα αλγόριθμο στη γλώσσα προγραμματισμού FORTRAN για να επιλύσει τη σχέση (1) μέσω μιας γραμμικοποιημένης μεθόδου. Πιο συγκεκριμένα πρότειναν μια τεχνική βασιζόμενοι σε μια κανονική παρουσίαση των acyclic phase type distributions (ACP), η οποία προσαρμόζεται για την ερμηνεία των εφαρμογών με αποκομμένα δεδομένα (censored data).

Το δεύτερο σετ των αλγορίθμων βασίζεται στο moment matching. Σε μια σειρά από άρθρα οι Johnson και Taaffe (1989, 1990a, 1990b, 1991) έδειξε πως οι πρώτες k ροπές μιας μη εκφυλισμένης κατανομής μπορεί να συνδυαστεί μέσω ενος συνδιασμού των Erlang κατανομών. Το πακέτο MEDA (χρησιμοποιώντας τη γλώσσα προγραμματισμού PASCAL) αναπτύχθηκε από τον Schmickler (1992), στο οποίο εργαζόμαστε μέσω μιας mixture Erlang κατανομής, υλοποιώντας ένα κριτήριο που βασίζεται στο ταίριασμα της πρώτης ροπής σε συνδυασμό με την ελαχιστοποίηση της διασποράς. Εναλλακτικά ο Johnson (1993) διατύπωσε τον αλγόριθμο MEFIT στην FORTRAN για να ταιριάξει τις τρεις πρώτες ροπές μιας μη εκφυλισμένης κατανομής με τις τρεις πρώτες ροπές της κατανομής Erlang. Έπειτα το 1996 οι Lang και Arthur συγκρίνουν τη μέθοδο των ροπών με την μέθοδο μέγιστης πιθανοφάνειας.

Τέλος το τρίτο σετ των μεθόδων βασίζεται στην ελαχιστοποίηση των αποστάσεων MD (minimum distance). Οι Parr και Schucany (1980) πρότειναν τη διαδικασία MD για την εκτίμηση των παραμέτρων περιοριζόμενοι στις λεγόμενες ACP κατανομές. Η επιλεγείσα απόσταση πραγματοποιείτο μέσω του δείκτη καλής προσαρμογής Kolmogorov - Smirnov. Παρόλα αυτά, η ποιο πολύ χρησιμοιημένη μέθοδος απ᾽όλες τις τεχνικές βασίζεται στη μέθοδο ελαχίστων τετραγώνων.

Όλες οι προηγούμενες μεθόδους που αναφέραμε έχουν περιορισμούς. Οι Riska et al. (2002) ανέφερε ότι, ειδικά στην περίπτωση όπου η αρχική κατανομή είναι άγνωστη τότε η μέθοδος των ροπών είναι περισσότερο αποτελεσματική. Πιο πρόσφατα, οι Marshall και Zenga (2009a) μελέτησαν τη διαδικασία προσαρμογής εξετάζοντας την επιλογή των αρχικών εκτιμήσεων των παραμέτρων από συγκεκριμένα δεδομένα που προέρχονται από ασθενείς οι οποίοι νοσηλεύονται και ο στόχος είναι να εξετάσουμε το χρόνο παραμονής τους στο νοσοκομείο. Η αρχική εκτίμηση παραμέτρων πραγματοποιήθηκε μέσω εξομοίωσης (simulate) της Coxian κατανομής χρησιμοποιώντας την με τέσσερεις φάσεις και με πέντε φάσεις και η ελαχιστοποίηση της συνάρτησης πιθανοφάνειας πραγματοποιείται με τη χρήση του Sequential Quadratic Programming (SQP).

Η πρώτη λοιπόν αυτή ενότητα οργανώνεται ως εξής. Άρχικά στην 3.1 ενότητα κάνουμε μια μικρή εισαγωγή στην μέθοδο μέγιστης πιθανοφάνειας. Η μέθοδος αυτή είναι ιδιαίτερα δημοφιλής λόγω των καλών της ιδιοτήτων. Έπειτα στην 3.2 ενότητα παρουσιάζουμε εναν επαναληπτικό αριθμητικό αλγόριθμο, τον EM. Αναλύουμε διεξοδικά τον αλγόριθμο EM, αναφέρουμε τα πλεονεκτήματα του, θέτουμε τα κριτήρια τερματισμού αυτού του αλγορίθμου και στο τέλος αναφέρουμε και κάποιες παραλλαγές του. Στην ενότητα 3.3 αναφέρουμε και τη μέθοδο των ροπών προσαρμόζοντας την Erlang κατανομή με 2 φάσεις και στη συνέχεια με 3 φάσεις. Στη συνέχεια η ενότητα 3.4 παρουσιάζει κάποια μέτρα απόδοσης, με τα οποία μπορούμε να συγκρίνουμε τις εκτιμήτριες των προσαρμοζμένων μοντέλων με τους αλγορίθμους Quasi-Newton και Nealder-Mead. Τέλος στην ενότητα 3.5 παρουσιάζουμε το PH-plot για τη γραφική απεικόνιση των προσαρμοσμένων phase type distributions.

3.1 Μέθοδος μέγιστης πιθανοφάνειας

Η μέθοδος μέγιστης πιαθανοφάνειας αποτελεί αναμφισβήτητα την ποιο δημοφιλή μέθοδο για τον υπολογισμό εκτμητριών. Γενικά λοιπόν έστω ότι έχουμε ένα τυχαίο δείγμα $X_1, X_2, ..., X_n$ με συνάρτηση πυκνότητας πιθανότητας $f(x; \theta_1, \theta_2, ..., \theta_k)$,η συνάρτηση μέγιστης πιθανοφάνειας ορίζεται ως εξής:

$$L(\theta; x) = L(\theta_1, ..., \theta_k; x_1, ..., x_n) = \prod_{i=1}^{n} f(x; \theta_1, \theta_2, ..., \theta_k)$$

Εαν η συνάρτηση πιθανοφάνειας είναι παραγωγίσημη σε κάθε θ_i, τότε οι εκτημήτριες μέγιστης πιθανοφάνειας δίνονται από την σχέση:

$$\frac{\partial}{\partial \theta_i} L(\theta, x) = 0 \text{ για } i = 1, 2, ..., k$$

Αναφέρουμε κάποιες χαρακτηριστικές ιδιότητες για την μέθοδο μέγιστης πιθανοφάνειας:

- Η εκτιμήτρια μέγιστης πιθανοφάνειας είναι ασυμπτωτικά αμερόληπτη, και γενικότερα η μεροληψία της τείνει στο 0 όσο το μέγεθος του δείγματος αυξάνεται.
- Η εκτιμήτρια μέγιστης πιθανοφάνειας είναι ασυμπτωτικά αποτελεσματική, δηλαδή επιτυγχάνει το κάτω φράγμα του Cramer-Rao όταν ο αριθμός του δείγματος τείνει στο άπειρο. Αυτό σημαίνει ότι ασυμπτωτικά κανένας άλλος αμερόληπτος εκτιμητής δεν έχει μικρότερο μέσω τετραγωνικό σφάλμα απο την εκτιμήτρια μέγιστης πιαθνοφάνειας.
- Η εκτιμήτρια μέγιστης πιθανοφάνειας είναι ασυμπτωτικά κανονική. Καθώς ο αριθμός του δείγματος αυξάνεται η κατανομή της εκτιμήτριας μέγιστης

πιθανοφάνειας τείνει στη Gaussian κατανομή με πίνακα συνδιακύμανσης ισο με τη αντίστροφο πίνακα της πληροφορίας του Fisher.

Βέβαια οι συνθήκες κανονικότητας που απαιτούνται για την εξασφάλιση αυτής της συμπεριφοράς της εκτιμήτριας μέγιστης πιθανοφάνειας είναι:

1. Η πρώτη και δεύτερη παράγωγος της λογαριθμικής συνάρτησης πρέπει να είναι καλώς ορισμένες.
2. Ο πίνακας της πληροφορίας του Fisher πρέπει να είναι διάφορος του μηδένος.

Έστω $\mathrm{I}(\theta;\mathrm{y}) = -\dfrac{\partial^2 \log \mathrm{L}(\theta)}{\partial\theta\partial\theta'}$ είναι ο πίνακας των αρνητικών δεύτερων-μερικών παραγώγων της λογαριθμικής συνάρτησης. Υπό τις συνθήκες κανονικότητας, τότε ο αναμενόμενος πίνακας πληροφορίας Fisher $\mathrm{I}(\theta)$ δίνεται απο τη σχέση:

$$\begin{aligned}\mathrm{I}(\theta) &= \mathrm{E}\left\{\mathrm{S}(\mathrm{Y};\theta)\mathrm{S}'(\mathrm{Y};\theta)\right\} \\ &= -\mathrm{E}_\theta\left\{\mathrm{I}(\theta;\mathrm{Y})\right\}\end{aligned}$$

όπου $\mathrm{S}(\mathrm{Y};\theta) = \dfrac{\partial \log \mathrm{L}(\theta)}{\partial\theta}$

Ο ασυμπτωτικός πίνακας συνδιασποράς της εκτιμήτριας μέγιστης πιθανοφάνειας $\hat{\theta}$ είναι ίσο με τον αντίστροφο του πίνακα πληροφορίας $\mathrm{I}(\theta)$, το οποίο μπορεί να προσεγγιστεί απο $\mathrm{I}\left(\hat{\theta}\right)$, τότε το τυπικό σφάλμα δίνεται απο τη σχέση:

$$\mathrm{SE}\left(\hat{\theta}_\mathrm{i}\right) \approx \left(\mathrm{I}^{-1}\left(\hat{\theta}\right)\right)_{\mathrm{ii}}^{1/2}$$

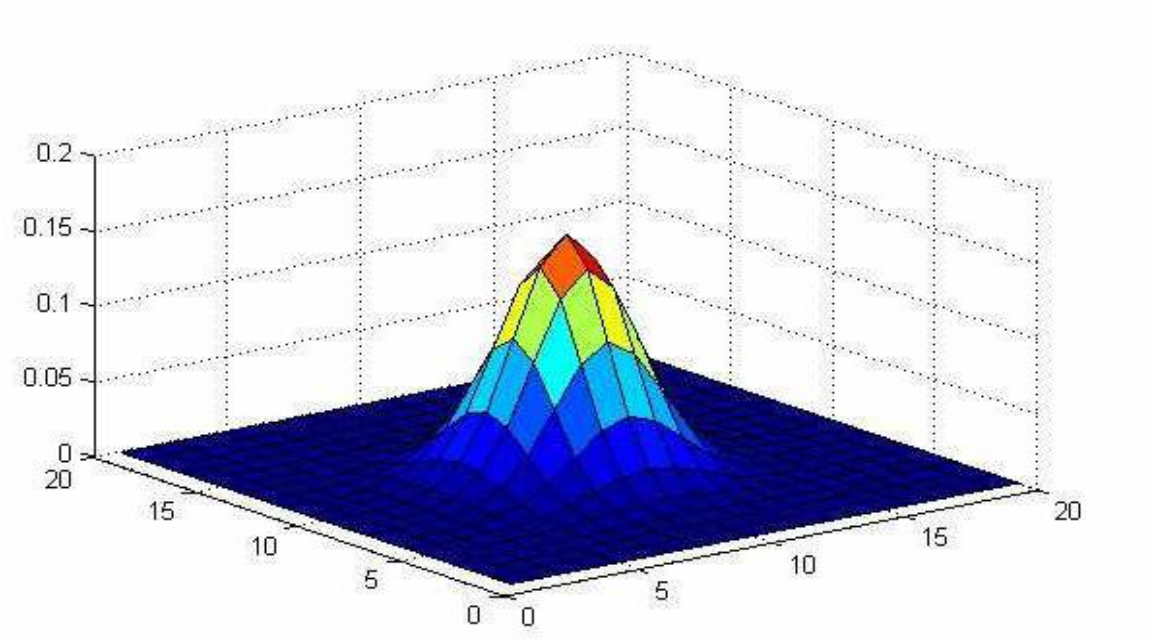
Είναι σύνηθες στην πράξη να εκτιμήσουμε τον αντίστροφο του πίνακα συνδιασποράς απο τον πίνακα παρατηρούμενης πληροφορίας $\mathrm{I}\left(\hat{\theta};\mathrm{y}\right)$, παρά να

υπολογίσουμε τον αναμενόμενο πίνακα πληροφορίας $\mathrm{I}(\theta)$ όταν $\theta = \hat{\theta}$. Αυτή η προσέγγιση δίνεται απο τη σχέση:

$$\mathrm{SE}\left(\hat{\theta}_{\mathrm{i}}\right) \approx \left(\mathrm{I}^{-1}\left(\hat{\theta}; \mathrm{y}\right)\right)_{\mathrm{ii}}^{1/2}$$

3.2 Εκτίμηση των παραμέτρων των PH κατανομών χρησιμοποιώντας τον αλγόριθμο EM

Η μέθοδος μέγιστης πιθανοφάνειας αποτελεί την ποιο δημοφιλή μέθοδο για την εκτίμηση των παραμέτρων του μοντέλου καθως και προσαρμογή των δεδομένων μεσω των PH κατανομών. Είδαμε ότι η εκτίμητρια μέγιστης πιθανοφάνειας απαιτεί να μεγιστοποιήσουμε μια έκφραση $L(\theta)$. Αυτό όμως μπορεί να είναι από τετριμμένο, έως εξαιρετικά πολύ δύσκολο. Η απλή περίπτωση είναι η πιθανοφάνεια L να εξαρτάται από μία βαθμωτή παράμετρο, να είναι παντού συνεχής και δις παραγωγίσιμη, και η πρώτη παράγωγος της να έχει ρίζες που μπορούν να βρεθούν με κλειστό τύπο. Όταν τα πράγματα δεν είναι τόσο απλά, πρέπει να καταφύγουμε σε πιο προχωρημένες μεθόδους βελτιστοποίησης. Αυτές οι αριθμητικές μέθοδοι λειτουργούν διορθώνοντας επαναληπτικά μια αρχική εκτίμηση της λύσης, μέχρι αυτή να φτάσει αρκετά κοντά στην πραγματική λύση.



Μια γνωστή αριθμητική μέθοδος είναι ο αλγόριθμος ΕΜ (Expectation Maximization). Ο αλγόριθμος ΕΜ είναι μια γενική μέθοδος εύρεσης εκτιμητών μέγιστης πιθανοφάνειας των παραμέτρων μιας δοθείσας κατανομής, σε προβλήματα όπου κάποιες μεταβλητές δεν έχουν παρατηρηθεί (μη παρατηρήσημες ή κρυμμένες μεταβλητές). Συνεπώς το πεδίο εφαρμογής του ΕΜ αλγορίθμου συνίσταται για την επίλυση δύο βασικών προβλημάτων. Το πρώτο υφίσταται όταν έχουμε δεδομένα απο τα οποία ορισμένες τιμές λειπούν είναι τα λεγόμενα missing data, εξαιτίας κάποιων λαθών ή περιορισμών που υπάρχουν κατα τη διάρκεια της διαδικασίας παρατήρησης του πειράματος. Το δεύτερο, αφορά κυρίως εφαρμογές μικτών μοντέλων, στα οποία η μεγιστοποιήση της πιθανοφάνειας είναι αναλυτικά αδύνατη. Γι' αυτό υποθέτουμε την ύπαρξη κάποιων επιπρόσθρτων, αλλά κρυμμένων μεταβλητών, που προσδιορίζουν την ομάδα στην οποία ανήκει το κάθε πρότυπο. Όπως προαναφέρθηκε οι Asmussen, Nerman και Olsson (1996) ανέπτυξαν τον αλγόριθμο ΕΜ (expectation maximization) για να εκτιμήσει τις παραμέτρους υπολογίζοντας πρώτα τη συνάρτηση μέγιστης πιθανοφάνειας όταν θα την εφάρμοζε σε πραγματικά δεδομένα. Επίσης ενέκριναν ότι ο αλγόριθμος

αυτός μπορεί να χρησιμοποιηθεί ακόμα και για την προσαρμογή κατανομής μέσω των PH κατανομών. Μάλιστα ο Olsson (1996) επέκτεινε τον αλγόριθμο αυτό στο να μπορεί να χρησιμοποιηθεί ακόμα και στην περίπτωση που έχουμε δεξιά αποκομμένα δεδομένα (right censored data) ή ακόμα και δεδομένα που είναι σε μορφή διαστήματος (interval-censored data).

Η βασική ιδέα του αλγορίθμου ΕΜ είναι πως εκτός απο τα δεδομένα που έχουμε παρατηρήσει υπάρχουν και κάποια άλλα δεδομένα τα οποία αφενος μεν δεν έχουμε παρατηρήσει αλλά αν τα είχαμε παρατηρήσει το πρόβλημα της εκτιμήτριας μέγιστης πιθανοφάνειας θα ήταν πολύ πιο απλό. Επομένως υποθέτουμε πως υπάρχουν χαμένα δεδομένα (missing data). Είναι σημαντικό σε αυτό το σημείο να πούμε ότι η ιδέα των missing data είναι μια τεχνική που χρησιμοποιείται στη στατιστική. Η τεχνική αυτή δηλαδή να συμπληρώνουμε τα παρατηρούμενα δεδομένα (observed data) με μη παρατηρούμενα δεδομένα (missing data) ονομάζεται συμπλήρωση δεδομένων (data augumentation). Τα δεδομένα που προκύπτουν απο το συνδυασμό των παρατηρούμενων και μη παρατηρούμενων ονομάζονται πλήρη δεδομένα (complete data). Θα πρέπει να αναφέρουμε οτι η τεχνική της συμπλήρωσης δεδομένων αποτελεί τη βάση και για άλλες πολύ δημοφιλείς τεχνικές όπως οι τεχνικές Markov Chain Monte Carlo (MCMC) .

Ιστορικά ο αλγόριθμος ΕΜ πρωτοπαρουσιάστηκε στη γενική του μορφή το 1977 (Dempster et al., 1977). Διάφορες παραλλαγές του αλγορίθμου σε συγκεκριμένες εφαρμογές έχουν χρησιμοποιηθεί ήδη στη δεκαετία του 1960 και προηγουμένως. Στην πραγματικότητα ερευνητές σε διάφορα αντικείμενα είχαν χρησιμοποιήσει επαναληπτικούς αλγορίθμους για την επίλυση των προβλημάτων τους οι οποίοι αργότερα αποδείχθηκε ότι μπορούν να ιδωθούν σαν εφαρμογές του αλγορίθμου ΕΜ. Ο αλγόριθμος έγινε ιδιαίτερα δημοφιλής τα τελευταία χρόνια λόγω της χρήσης υπολογιστών για την επίλυση ολοένα και μεγαλύτερων στατιστικών προβλημάτων.

Η γενική φιλοσοφία του ΕΜ διατυπώνεται ακολούθως. Ξεκινάμε με μια αρχική εκτίμηση των παραμέτρων του μεικτού μοντέλου, που πρέπει να εκτιμηθούν. Κάθε επανάληψη αποτελείται απο δύο βήματα. Το πρώτο είναι το Ε-βήμα (expectation step) στο οποίο προσπαθούμε να υπολογίσουμε ένα κάτω φράγμα της λογαριθμηκής πιθανοφάνειας και να το μεγιστοποιήσουμε στην κατανομή των κρυμμένων μεταβλητών. Το δεύτερο βήμα είναι το Μ-βήμα (maximization step) στο οποίο μεγιστοποιείται το κάτω φράγμα ως προς τις παραμέτρους της μικτής κατανομής. Αυτά τα δύο βήματα επαναλαμβάνονται μέχρι να υπάρξει σύγκλιση στην ακολουθία των παραμέτρων, δηλαδή όταν φτάσουμε σε κάποιο τοπικό μέγιστο.

3.2.1 Κατασκευή του ΕΜ αλγορίθμου για ένα πλήρες δείγμα

Θεωρούμε τυχαίες μεταβλητές $(y_1, y_2, ..., y_M)$ απο $PH_p(\pi,T)$. Είμαστε στην περίπτωση όπου έχουμε μόνο μη πλήρη δεδομένα, διοτι έχουμε μόνο τους χρόνους απορρόφησης και καμία αλλη πληροφορία δεν είναι διαθέσιμη.

Έστω λοιπόν $(y_1, y_2, ..., y_M)$ και $\theta = (\pi, T, t)$, όπου $t = -Te$. Η συνάρτηση πιθανοφάνειας για τα μη πλήρη δεδομένα (incomplete data) δίνεται απο τη σχέση:

$$L(\theta; y) = \prod_{k=1}^{M} \pi e^{Ty_k} t$$

και η λογαριθμική συνάρτηση είναι:

$$l(\theta; y) = \sum_{k=1}^{M} \log f(y_k)$$

όπου $f\left(y_k\right)=\pi e^{Ty_k}t$. Αντικαθιστώντας όπου $\pi=\sum_{j=1}^{p-1}\pi_j e_j^{'}+\left(1-\sum_{j=1}^{p-1}\pi_j\right)e_p^{'}$ τότε:

$$f\left(y_k\right)=\sum_{j=1}^{p-1}\pi_j e_j^{'}e^{Ty_k}t+\left(1-\sum_{j=1}^{p-1}\pi_j\right)e_p^{'}e^{Ty_k}t$$

Υποθέτουμε λοιπόν ότι έχουμε μια ολοκληρωμένη παρατήρηση της Μαρκοβιανής αλυσίδας $\left\{X\left(t\right)\right\}_{t\geq 0}$ με p καταστάσεις. Επιπλέον, θεωρούμε οτι ο χρόνος απορρόφησης είναι $y\in\left(y_1,y_2,...,y_M\right)$, με n βήματα που θα έχει πραγματοποιήσει νωρίτερα , οι καταστάσεις που θα έχει επισκευθεί με την σειρά θα είναι $i_0,i_1,...,i_n$, ενω ο χρόνος που θα διαρκέσει αναμέσα στις μεταβάσεις απο τη μια κατάσταση στην άλλη, είναι αντίστοιχα $s_0,s_1,...,s_n$. Άρα εύκολα αντιλαμβανόμαστε ότι $s_0+s_1+...+s_n=y.$

Η συνάρτηση πυκνότητας-πιθανότητας για ένα πλήρες δείγμα θα είναι της μορφής:

$$L_f\left(\theta,x\right)=\prod_{i=1}^{p}\pi_i^{B_i}\prod_{i=1}^{p}\prod_{j\neq i}^{p}t_{ij}^{N_{ij}}e^{-t_iZ_i}\prod_{i=1}^{p}t_i^{N_i}\ e^{-t_{ii}Z_i}\ ,$$

όπου

$B_i=\sum_{\nu=1}^{p}1_{\left\{I_0^{[\nu]}=i\right\}}$ το πλήθος των Μαρκοβιανών διαδικασίων που ξεκινούν απο την κατάσταση i, i=1,2,…,p

$Z_i=\sum_{\nu=1}^{p}\prod_{k=0}^{m(\nu)-1}1_{\left\{I_0^{[\nu]}=i\right\}}S_k^{[\nu]}$ ο συνολικός χρόνος παραμονής στην κατάσταση i, i=1,2,…,p

$N_{ij}=\sum_{\nu=1}^{p}\sum_{k=0}^{m(\nu)-1}1_{\left\{I_k^{[\nu]}=i,I_{k+1}^{[\nu]}=j\right\}}$ το συνολικό πλήθος αλμάτων από την κατάσταση i στην κατάσταση j για $i\neq j$ για i=1,2,…,p και για j=1, 2,…, p, Δ

όπου B_i είναι ο αριθμός των διαδικασιών στη καταάσταση i, N_i είναι ο αριθμός των διαδικασιών που εξέρχονται απο την κατάσταση i στην κατάσταση απορρόφησης, N_{ij} είναι ο αριθμός των μεταπηδήσεων ανάμεσα σε όλες τις διαδικασίες, Z_i είναι ο συνολικός χρόνος στην κατάσταση i πριν την απορρόφηση για όλες τις διαδικασίες.

Η λογαριθμική συνάρτηση πιθανοφάνειας για τα πλήρη δεδομένα είναι:

$$l_f(\theta; x) = \sum_{i=1}^{p} B_i \log(\pi_i) + \sum_{i=1}^{p}\sum_{j\neq i}^{p} N_{ij} \log(t_{ij}) - \sum_{i=1}^{p}\sum_{j\neq i}^{p} t_{ij} Z_i + \sum_{i=1}^{p} N_i \log(t_i) - \sum_{i=1}^{p} t_i Z_i$$

Είναι σαφέστατα ξεκάθαρο ότι οι εκτιμητές μέγιστης πιθανοφάνειας δίνονται:

$$\hat{\pi}_i = \frac{B_i}{M}, \quad \hat{t}_{ij} = \frac{N_{ij}}{Z_i}, \quad \hat{t}_i = \frac{N_{i0}}{Z_i}, \quad \hat{t}_{ii} = -\left(\hat{t}_i + \sum_{\substack{j=1 \\ j\neq i}}^{p} \hat{t}_{ij} \right) \text{ για } i, j = 1, 2, ..., p$$

Έστω ότι έχουμε κάποιες αρχικές τιμές των παραμέτρων $\theta_0 = (\pi_0, T_0, t_0)$ τότε ο αλγόριθμος EM με μια συνοπτική σκιαγράφηση δουλεύει ως εξής:

1. Ε-βήμα Υπολογισμός της συνάρτησης

$$h : \theta \to E_{\theta_0}\left(l_f(\theta; x) \middle| Y = y \right)$$

2. Μ-βήμα

$$\theta_0 = \arg\max_{\theta} h(\theta).$$

3. Επιστροφή στο βήμα (1).

Το Ε-βήμα και Μ-βήμα επαναλαμβάνονται μέχρι να επιτευχθεί η σύγκλιση.

Αφού η εξίσωση 3* είναι μια γραμμική εξίσωση των επαρκών στατιστικών συναρτήσεων B_i, Z_i, N_i και N_{ij} είναι αρκετό να υπολογίσουμε τις αντίστοιχες δεσμευμένες τιμές αυτών των συναρτήσεων. Έστω ότι $B_i^k, Z_i^k, N_i^k, N_{ij}^k$ είναι οι αντίστοιχες στατιστικές συναρτήσεις τότε:

$$B_i = \sum_{k=1}^{M} B_i^k \qquad Z_i = \sum_{k=1}^{M} Z_i^k \qquad N_i = \sum_{k=1}^{M} N_i^k \qquad N_{ij} = \sum_{k=1}^{M} N_{ij}^k$$

για $i, j = 1, 2, ..., p$ με $i \neq j$ και για αυτό $E_\theta\left(S \middle| Y = y\right) = \sum_{k=1}^{M} E_\theta\left(S^k \middle| Y_k = y_k\right)$, όπου $S \in \left\{B_i, Z_i, N_i, N_{ij}\right\}$. Το ποιο δύσκολο σημείο που συναντάμε στον EM αλγόριθμο είναι στο υπολογισμό της δεσμευμένης μέσης τιμής στο E-βήμα $E_\theta\left(S^k \middle| Y_k = y_k\right)$.

Χαρακτηριστικά αναφέρουμε το εξής θεώρημα:

Θεώρημα Για $i, j = 1, 2, ..., p, i \neq j$ έχουμε τα εξής αποτελέσματα

$$E_{(\pi,T)}\left[B_i^{[k]} \middle| Y_k = y_k\right] = \frac{\pi_i e_i' \exp(T y_k) t}{\pi \exp(T y_k) t}$$

$$E_{(\pi,T)}\left[Z_i^{[k]} \middle| Y_k = y_k\right] = \frac{\int_o^{y_k} \pi \exp(Tu) e_i e_i' \exp\left(T(y_k - u)\right) t\, du}{\pi \exp(T y_k) t}$$

$$E_{(\pi,T)}\left[N_i^{[k]} \middle| Y_k = y_k\right] = \frac{t_i \pi \exp(T y_k) e}{\pi \exp(T y_k) t}$$

$$E_{(\pi,T)}\left[N_{ij}^{[k]} \middle| Y_k = y_k\right] = \frac{t_{ij} \int_o^{y_k} \pi \exp(Tu) e_i e_j' \exp\left(T(y_k - u)\right) du}{\pi \exp(T y_k) t}$$

EM χρησιμοποιώντας τους Runge-Kutta (EM-RK)

Επιπλέον στο άρθρο του ο Asmussen (1996) κάνει τις εξής υποθέσεις. Έστω ότι οι όροι $a(y|\pi,T)$, $b(y|T)$, $c(y;p|\pi,T)$ είναι διανύσματα p-δίαστασης τα οποία ορίζονται ως εξής:

$$a(y|\pi,T) = \pi \exp\{Ty\}$$

$$b(y|T) = \exp\{Ty\}$$

$c(y;i|\pi,T) = \int_0^y \pi \exp\{Tu\} e_i \exp\{T(y-u)\} tdu$ όπου i=1,2,…,p ενώ e_i είναι το μοναδιαίο διάνυσμα. Τότε αποδεικνύεται ότι:

$$E_{(\pi,T)}\left[B_i^{[\nu]} | Y = y_\nu\right] = \frac{\pi_i b_i(y_\nu|T)}{\pi b(y_\nu|T)} \qquad E_{(\pi,T)}\left[N_{ij}^{[\nu]} | Y = y_\nu\right] = \frac{t_{ij} c_i(y_\nu;i|\pi,T)}{\pi b(y_\nu|T)}$$

$$E_{(\pi,T)}\left[Z_i^{[\nu]} | Y = y_\nu\right] = \frac{c_i(y_\nu|\pi,T)}{\pi b(y_\nu|T)} \qquad E_{(\pi,T)}\left[N_{i0}^{[\nu]} | Y = y_\nu\right] = \frac{t_i \alpha_i(y_\nu|\pi,T)}{\pi b(y_\nu|T)}$$

3.2.2 Πλεονεκτήματα και μειονεκτήματα του αλγορίθμου EM

Πλεονεκτήματα

1. Όπως όλες οι επαναληπτικές μέθοδοι έτσι και ο αλγόριθμος EM βασίζεται σε καλές αρχικές τιμές. Το σημαντικό πλεονέκτημα του αλγορίθμου είναι πως αν οι αρχικές τιμές είναι μέσα στο αποδεκτό πεδίο ορισμού τότε σε κάθε επανάληψη είμαστε σίγουροι πως οι τιμές που θα πάρουμε θα είναι και αυτές αποδεκτές, κάτι που δεν μπορεί να εγγυηθεί απο άλλες μεθόδους. Κάτι τέτοιο είναι πολύ σημαντικό

καθώς μας προφυλάσει από μη αποδεκτές λύσεις, το οποίο συμβαίνει αρκετά συχνά σε προβλήματα με μεγάλο αριθμό παραμέτρων.

2. Ο αλγόριθμος είναι συχνά πολύ εύκολος να προγραμματιστεί ακόμα και σε απλά πακέτα, δεδομένου ότι δεν προυποθέτει την ύπαρξη παραγώγων και αντίστροφων πινάκων, όπως είναι η μέθοδος Newton-Raphson.

3. Επιπλέον ο αλγόριθμος ΕΜ θεμελιώθηκε πάνω σε καθαρά στατιστική επιχειρηματολογία και επομένως προσφέρει πολύ χρήσιμη στατιστική ερμηνεία.

4. Πολλές φορές κάποια ενδιάμεσα αποτελέσματα του αλγορίθμου έχουν την δική τους σημασία και μπορούν να χρησιμοποιήθουν για περαιτέρω στατισική ανάλυση.

Μειονεκτήματα

1. Αργή σύγκλιση. Ο αλγόριθμος ΕΜ συγκλίνει ποιο αργά απο ότι οι άλλοι αλγόριθμοι. Για παράδειγμα η σύγκλιση του είναι γραμμική ενω η σύγκλιση της μεθόδου Newton-Raphson είναι τετραγωνική. Αυτό στην πράξη σημαίνει ότι χρειαζόμαστε περισσότερες επαναλήψεις μέχρι να βρούμε τη λύση.

2. Η βασική ιδιότητα του αλγορίθμου να αυξάνει την πιθανοφάνεια σε κάθε βήμα δε σημαίνει απαραίτητα πως στο τέλος του αλγορίθομου έχει βρεθεί το ολικό μέγιστο και όχι κάποιο τοπικό μέγιστο. Το σημείο στο οποίο θα συγκλίνει ο αλγόριθμος εξαρτάται απο τις αρχικές τιμές. Συνεπώς για να είμαστε σίγουροι πως έχει βρεθεί όντως το όλικο μέγιστο πρέπει να επαναλάβουμε τον αλγόριθμο με διαφορετικές αρχικές τιμές.

3. Η λύση εξαρτάται όπως και σε κάθε αλγόριθμο απο τις αρχικές τιμές. Δεδομένης της πιο αργής σύγκλισης του ΕΜ κακές αρχικές τιμές ίσως οδηγήσουν σε πολύ περρισότερες επαναλήψεις του αλγορίθμου και επομένως η ανάγκη για καλύτερες αρχικές τιμές είναι μεγαλύτερη.

3.2.3 Κριτήρια τερματισμού

Ας δούμε μερικά κριτήρια τερματισμού του ΕΜ που χρησιμοποιούνται στην πράξη. Τα κριτήρια αυτά γενικά χωρίζονται σε δύο κατηγορίες:

1. Κριτήρια βασισμένα στην πρόοδο της πιθανοφάνειας

Σταματάμε τις επαναλήψεις όταν: $\left|\frac{L^{(r+1)} - L^{(r)}}{L^{(r+1)}}\right| \leq \varepsilon$ όπου ε είναι μια μικρή τιμή

και $L^{(r)}$ είναι η λογαριθμική πιθανοφάνειας μετά την r επανάληψη. Ουσιαστικά το κριτήριο λέει να σταματήσουμε όταν η λογαριθμική συνάρτηση δεν αλλάζει πια από επαναλήψη σε επαναλήψη. Υπαρχουν δύο σημαντικά σημεία που πρέπει κανείς να έχει υπόψη του: το οτι δεν αλλάζει η πιθανοφάνεια δε σημαίνει πως οι τιμές των παραμέτρων δεν αλλάζουν, αν η πιθανοφάνεια είναι σχεδόν επίπεδη σε κάποιο σημείο του παραμετρικού χώρου τότε οι παράμετροι μπορεί να αλλάζουν και μάλιστα δραματικά. Επίσης το ότι η πιθανοφάνεια αλλάζει πολύ λίγο, δε σημαίνει πως βρέθηκε απαραίτητα το μέγιστο καθώς μπορεί να σταματήσει αλλά ύστερα απο λίγες επαναλήψεις ο αλγόριθμος να ξέφευγε από την περιοχή αυτή.

2. Κριτήρια βασισμένα στην αλλαγή των τιμών των παραμέτρων

Σταματάμε τις επαναλήψεις όταν $\max_j\left(\left|\theta_j^{(r+1)}-\theta_j^{(r)}\right|\right)\le\varepsilon$ όπου $\theta_j^{(r)}$ είναι η τιμή της παραμέτρου θ_j μετα την r επανάληψη. Δηλαδή το κριτήριο ικανοποιείται όταν όλες οι παράμετροι αλλάζουν απο μια επανάληψη στην επόμενη λιγότερο από μια μικρή ποσότητα ε. Εναλλακτικά μπορεί κανείς να χρησιμοποιήσει κριτήρια που μετράνε τη μέση διαφορά απο επανάληψη σε επανάληψη η κάποια άλλη απόσταση. Για παράδειγμα μπορεί κανείς να σταματήσει τις επαναλήψεις όταν:

$$\sum_{j=1}^{p}\left(\theta_j^{(r+1)}-\theta_j^{(r)}\right)\le\varepsilon$$

δηλαδή το άθροισμα τετραγωνικών αποκλίσεων για όλες τις παραμέτρους σε δύο διαδοχικές επαναλήψεις να είναι μικρότερο του ε, όπου ε είναι μια ποσότητα της τάξης του 10^{-10}.

3.2.4 Πώς επιτυχγάνεται η σύγκλιση του αλγορίθμου ΕΜ

Σε αυτή την ενότητα θα εξετάσουμε γιατί ο αλγόριθμος ΕΜ σε κάθε επανάληψη μεγαλώνει την πιθανοφάνεια, δηλαδή να εξετάσουμε γιατί ο αλγόριθμος συγκλίνει.

Έστω λοιπόν ότι τα παρατηρούμενα δεδομένα τα συμβολίζουμε με X ενω τα πλήρη δεδομένα (complete data) με Y. Οι παράμετροι που θέλουμε να εκτιμήσουμε συμβολίζονται με $\psi=(\psi_1,\psi_2,...,\psi_d)$. Η κατανομή των πραγματικών δεδομένων είναι $g(y;\psi)$ ενώ των πλήρων δεδομένων είναι $g_c(y;\psi)$. Τέλος η ποσότητα που

θέλουμε να μεγιστοποιήσουμε είναι η $L(\psi)$ ενω με $L_c(\psi)$ θα συμβολίσουμε αντίστοιχα την ποσότητα των πλήρων δεδομένων. Ουσιαστικά θέλουμε να δείξουμε ότι:

$$L\left(\psi^{(k+1)}\right) \geq L\left(\psi^{(k)}\right)$$

όπου $\psi^{(k)}$ συμβολίζουμε τις παραμέτρους μετά την επανάληψη.

Η δεσμευμένη κατανομή των πλήρων δεδομένων δοθέντως των παρατηρούμενων δεδομένων θα είναι:

$$g(y|x;\psi) = \frac{g_c(y;\psi)}{g(x;\psi)}$$

$$\log L(\psi) = \log g(x;\psi)$$

επειδή έχουμε μια ένα προς ένα αντιστοίχιση των πλήρων δεδομένων στα πραγματικά και επομένως η από κοινού τους κατανομή θα είναι η ίδια με την κατανομή των πλήρων δεδομένων. Έτσι η λογαριθμική πιθανοφάνεια γράφεται ως:

$$\begin{aligned}\log L(\psi) &= \log g(x;\psi) \\ &= \log g_c(y;\psi) - \log g(x;\psi) \\ &= \log L_c(\psi) - \log g(y|x;\psi)\end{aligned}$$

Αν πάρουμε τις αναμενόμενες τιμές και στα δύο μέρη της ισότητας ως προς τη δεσμευμένη κατανομή των πλήρων δεδομένων δοθέντως των πραγματικών προκύπτει ότι:

$$\begin{aligned}\log L(\psi) &= E_{y|x;\psi^{(k)}}\left\{\log g_c(y;\psi)\right\} - E_{y|x;\psi^{(k)}}\left\{\log g(y|x;\psi)\right\} \\ &= Q\left(\psi;\psi^{(k)}\right) - H\left(\psi;\psi^{(k)}\right)\end{aligned}$$

όπου $Q\left(\psi;\psi^{(k)}\right)$ είναι η ποσότητα που μεγιστοποιηούμε σε κάθε Μ-βήμα. Επομένως η διαφορά της πιθανοφάνειας σε δύο διαδοχικές επαναλήψεις είναι:

$$\log \mathrm{L}\left(\psi^{(k+1)}\right) - \log \mathrm{L}\left(\psi^{(k)}\right)$$
$$= \left\{ \mathrm{Q}\left(\psi^{(k+1)};\psi^{(k)}\right) - \mathrm{Q}\left(\psi^{(k)};\psi^{(k)}\right) \right\} - \left\{ \mathrm{H}\left(\psi^{(k+1)};\psi^{(k)}\right) - \mathrm{H}\left(\psi^{(k)};\psi^{(k)}\right) \right\}$$

Για τον πρώτο όρο γνωρίζουμε πως για την κατασκευή του ΕΜ αλγορίθμου ότι στο Μ-βήμα έχουμε μεγιστοποιήσει τη συνάρτηση Q και αρα ισχύει ότι:

$$\mathrm{Q}\left(\psi^{(k+1)};\psi^{(k)}\right) \geq \mathrm{Q}\left(\psi^{(k)};\psi^{(k)}\right)$$

Επομένως αρκεί να δείξουμε ότι:

$$\mathrm{H}\left(\psi^{(k+1)};\psi^{(k)}\right) - \mathrm{H}\left(\psi^{(k)};\psi^{(k)}\right) \leq 0$$

Από τον ορισμό προηγούμενως της $\mathrm{H}\left(\psi;\psi^{(k)}\right)$ έχουμε ότι:

$$\mathrm{H}\left(\psi;\psi^{(k)}\right) = \mathrm{E}_{\mathrm{y}|\mathrm{x};\psi^{(k)}} \left\{ \log \mathrm{g}\left(\mathrm{y}|\mathrm{x};\psi\right) \right\}$$

και άρα

$$\mathrm{H}\left(\psi^{(k+1)};\psi^{(k)}\right) - \mathrm{H}\left(\psi^{(k)};\psi^{(k)}\right) = \mathrm{E}_{\mathrm{y}|\mathrm{x};\psi^{(k)}} \left\{ \log \mathrm{g}\left(\mathrm{y}|\mathrm{x};\psi\right) \right\}$$

$$= \mathrm{E}_{\mathrm{y}|\mathrm{x};\psi^{(k)}} \left\{ \log \mathrm{g}\left(\mathrm{y}\middle|\mathrm{x};\psi^{(k)}\right) \right\}$$

$$= \mathrm{E}_{\mathrm{y}|\mathrm{x};\psi^{(k)}} \left\{ \log \mathrm{g}\left(\mathrm{y}|\mathrm{x};\psi\right) - \log \mathrm{g}\left(\mathrm{y}\middle|\mathrm{x};\psi^{(k)}\right) \right\}$$

$$= \mathrm{E}_{\mathrm{y}|\mathrm{x};\psi^{(k)}} \left\{ \log \frac{\mathrm{g}\left(\mathrm{y}|\mathrm{x};\psi\right)}{\mathrm{g}\left(\mathrm{y}\middle|\mathrm{x};\psi^{(k)}\right)} \right\}$$

Γνωρίζουμε απο την ανισότητα Jensen ότι ισχύει το εξής:

$$\mathrm{E}\left(\log \mathrm{X}\right) \leq \log \mathrm{E}\left(\mathrm{X}\right) \text{ για κάθε τυχαία μεταβλητή X}$$

Επομένως για την περίπτωση μας θα είναι:

$$E_{y|x;\psi^{(k)}}\left\{\log\frac{g\left(y|x;\psi\right)}{g\left(y|x;\psi^{(k)}\right)}\right\}\le\log E_{y|x;\psi^{(k)}}\left\{\frac{g\left(y|x;\psi\right)}{g\left(y|x;\psi^{(k)}\right)}\right\}=0 \quad \text{διότι}$$

$$E_{y|x;\psi^{(k)}}\left\{\log\frac{g\left(y|x;\psi\right)}{g\left(y|x;\psi^{(k)}\right)}\right\}=\int\frac{g\left(y|x;\psi\right)}{g\left(y|x;\psi^{(k)}\right)}g\left(y|x;\psi^{(k)}\right)dy$$

$$=\int g\left(y|x;\psi\right)dy=1$$

και άρα ο αλγόριθμος θα τείνει στο 0 . Συνεπώς η ποσότητα είναι πάντα 0 και άρα η πιθανοφάνεια αυξάνεται σε κάθε επανάληψη.

3.2.5 Παραλλαγές του αλγορίθμου EM

Σ' αυτή την υποενότητα θα αναφερθούμε σε παραλλάγες του EM αλγορίθμου στην περίπτωση όπου αντιμετωπίσουμε συγκεκριμένα προβλήματα ως προς τον υπολογισμό του E-βήματος καθώς οι αναμενόμενες τιμές που χρειάζονται δεν υπάρχουν σε κλειστή μορφή. Στην περίπτωση αυτή έχουμε δύο επιλογές είτε θα υπολογίσουμε τα ολοκληρώματα που αφορούν τις αναμενόμενες τιμές με κάποια αριθμητική μέθοδο είτε θα προχωρήσουμε σε Monte Carlo ολοκλήρωση δηλαδή να υπολογίσουμε τις αναμενόμενες τιμές με ολοκλήρωση. Ενδεικτικά αναφέρουμε ότι η δεύτερη περίπτωση είναι και ποιο εύκολη στην υλοποιήση της με απλά στατιστικά πακέτα. Υπάρχουν οι εξής εναλλακτικοί αλγόριθμοι:

- Stochastic EM (SEM): αυτός ο αλγόριθμος σε κάθε E-βήμα αντί να υπολογίζει την αναμενόμενη τιμή απλά χρησιμοποιεί μια τυχαία μεταβλητή που γεννάμε απο τη δεσμευμένη κατανομή των missing data για δοθείσες τιμές των παραμέτρων και τα πραγματικά δεδομένα. Ο αλγόριθμος SEM έχει πολύ καλές ιδιότητες. Συγκλίνει στην περιοχή του μεγίστου. Δεν αυξάνει την

πιθανοφάνεια σε κάθε επανάληψη αλλά έχει μια ξεκάθαρη τάση προς το μέγιστο και όταν φτάσει στην περιοχή του απλά κυμαίνεται γύρω από αυτό.

- Monte Carlo EM (MCEM): Ο αλγόριθμος υπολογίζει το Ε-βήμα με Monte Carlo ολοκλήρωση. Με στατιστικούς όρους εκτιμά την αναμενόμενη τιμή παίρνοντας ένα μεγάλο δείγμα από την κατανομή και χρησιμοποιώντας τη δειγματική μέση τιμή ως εκτιμήτρια. Μοιάζει με τον SEM στο ότι και οι δύο προσωμοιώνουν απ' τη δεσμευμένη κατανομή στο Ε-βήμα αλλά ο MCEM προσομειώνει πολλές τιμές από όπου κι εκτιμά τη μέση τιμή ενω ο SEM μόνο μία τιμή. Επιπλέον πρέπει να σημειωθεί ότι στον αλγόριθμο Monte Carlo η πιθανοφάνεια μπορεί να μην αυξάνει σε κάθε επανάληψη λόγω του σφάλματος που εισάγουμε.

- Generalized EM: Κάτω από αυτή την επωνυμία αναφερόμαστε σε αλγόριθμους που επειδή η μεγιστοποιήση που χρειάζεται στο Μ-βήμα δεν είναι εύκολο να γίνει τότε αρκεί κατα τη διάρκεια του Μ-βήματος όχι να μεγιστοποιήσουμε την πιθανοφάνεια των πλήρων δεδομένων αλλά να βρούμε μια τιμή που μας εξασφαλίζει πως μεγαλώσαμε έστω και λίγο την πιθανοφάνεια που είχαμε στην προηγούμενη επανάληψη. Δηλαδή βεβαιωνόμαστε ότι ο αλγόριθμος διατηρεί τη μονοτονία του και πως σε κάθε επανάληψη η πιθανοφάνεια μεγαλώνει άσχετα με το γεγονός πως δε μεγαλώνει όσο πιθανότατα θα ήταν εφικτό.

3.3 Προσέγγιση των PH κατανομών μέσω της μεθόδου των ροπών

Αρχικά θα συνοψίσουμε για μεταγενέστερη χρήση τα κυριότερα χαρακτηριστικά της Erlang κατανομής και της γενικευμένης Erlang κατανομής, τις οποίες θα τις συμβολίζουμε για λόγους συντομίας ως ED_k (Erlang distribution) και GED_k (Generalized Erlang distribution) αντίστοιχα.

3.3.1 Η Erlang κατανομή ED_k

Η κατανομή μιας τυχαίας μεταβλητής y, η οποία εκφράζει το άθροισμα k-σταδίων, τα οποία είναι εκθετικά κατανεμημένα με την ίδια παράμετρο $\lambda > 0$, δίνεται από την Erlang κατανομή με συνάρτηση πυκνότητας- πιθανότητας:

$$ED_k: \quad f_Y(y) = \lambda^k y^k e^{-\lambda y} / \Gamma(k)$$

όπου Γ(k)=(k-1)!, με κ≥1, κ φυσικός αριθμός. Η παράμετρος κ ονομάζεται παράμετρος σχήματος ενώ η παράμετρος λ εκφράζει το ρυθμό της παραμέτρου (rate parameter). Η σωρευτική συνάρτηση κατανομής δίνεται από την σχέση:

$$ED_k: \quad F_Y(y) = \gamma(k, \lambda y)/(k-1)!$$
$$= 1 - \sum_{i=0}^{k-1} e^{-\lambda y} (\lambda y)^i / k!$$

όπου $\gamma(\cdot)$ είναι η συνάρτηση Γάμμα. Οι τέσσερις πρώτες ροπές της Erlang κατανομής είναι:

1η ροπή	$\mu_1=\frac{\kappa}{\lambda}$
Διασπορά	$\sigma^2=\frac{\kappa}{\lambda^2}$
3η ροπή	$\mu_3=\frac{2}{\sqrt{k}}$
4η ροπή	$\mu_4=\frac{6}{\sqrt{k}}$

3.3.2 Η Γενικευμένη Erlang κατανομή GED_k

Όταν οι παράμετροι των σταδίων είναι διαφορετικοί τότε η κατανομή της τυχαίας μεταβλητής Y (υπενθυμίζουμε ότι το Y εκφράζει το χρόνο απορρόφησης) θα δίνεται από τη Γενικευμένη Erlang κατανομή διάστασης k, με συνάρτηση πυκνότητας πιθανότητας:

$$\text{GED}_k: \quad f_Y(y)=(-1)^{k+1}\prod_{i=1}^{k}\lambda_i\sum_{i=0}^{k}\frac{e^{-\lambda y}}{\prod_{\substack{j=1\\ j\neq i}}^{k}(\lambda_i-\lambda_j)}$$

με $\lambda_i>0$ για κάθε i. Οι τέσσερις πρώτες ροπές της Γενικευμένης Erlang κατανομής είναι :

1η ροπή	$\mu_1=\sum_i\frac{1}{\lambda_i}$
Διασπορά	$\sigma^2=\sum_i\frac{1}{\lambda_i^2}$
3η ροπή	$\mu_3=\sum_i\frac{2}{\lambda_i^3}$
4η ροπή	$\mu_4=\sum_i\frac{9}{\lambda_i^4}+\sum_{i\neq j}\frac{6}{\lambda_i^2\lambda_j^2}$

3.3.3 Μέθοδος των ροπών (2 στάδια-φάσεις)

Όπως προαναφέρθηκε αυτή η κλασική μέθοδος χρησιμοποιείται για να προσεγγίσουμε την καμπύλη που θα ταιριάξει καλύτερα (η αλλιώς μπορούμε να πούμε η καμπύλη που θα προσαρμοστεί καλύτερα) στα εμπειρικά δεδομένα τα οποία θα είναι διαθέσιμα από ένα πραγματικό πρόβλημα.

Σ'αυτή τη μέθοδο ουσιαστικά προσεγγίζουμε τη συνάρτηση πυκνότητας πιθανότητας (αναφέρεται πολλές φορές και ως πραγματική κατανομή) με τη Γενικευμένη Erlang κατανομή, όπου εμείς απαιτούμε οι ροπές των δυο κατανομών να είναι ίσες. Όπως θα δούμε παρακάτω η προσέγγιση αυτή είναι εφικτή μέχρι ενός σημείου, ύστερα από το σημείο μπορούμε να εφαρμόσουμε άλλη μέθοδο. Βέβαια να σημειώσουμε ότι ακόμα και στην περίπτωση όπου μπορούμε να εφαρμόσουμε τη μέθοδο των ροπών, επιβάλλονται κάποιοι αυστηροί περιορισμοί σχετικά με την πραγματική κατανομή, οι οποίοι αν δεν ικανοποιηθούν θα οδηγηθούμε σε άτοπο. Τέλος για να αντιμετωπίζουμε το πρόβλημα αυτό στατιστικοί μελετητές προχώρησαν στη διαμόρφωση ενός μαθηματικού προγράμματος, το οποίο μας παρέχει την καλύτερη δυνατή λύση στο πρόβλημα μας.

Προχωράμε τώρα στην παρουσίαση όσων προαναφέρθηκαν. Πρώτα θα ξεκινήσουμε στην περίπτωση που έχουμε δύο φάσεις. Έτσι λοιπόν υποθέτουμε ότι έχουμε μια GED_2 με δύο παραμέτρους $\lambda_1, \lambda_2 > 0$. Η συνάρτηση πυκνότητας πιθανότητας της GED_2 θα είναι της μορφής:

$$f_2(y) = \frac{\lambda_1 \lambda_2}{\lambda_1 - \lambda_2}\left[e^{-\lambda_2 y} - e^{-\lambda_1 y}\right] \text{ με } y \geq 0$$

Υποθέτοντας ότι η πραγματική κατανομή έχει μέση τιμή μ και διασπορά σ^2 αντίστοιχα, θα τα εξισώσουμε με τις δύο πρώτες ροπές της GED2 και έτσι θα έχουμε τις εξής εξισώσεις:

$$\frac{1}{\lambda_1}+\frac{1}{\lambda_2}=\mu \quad \text{και} \quad \frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}=\sigma^2$$

Έτσι αν λύσουμε την πρώτη εξίσωση ως προς $\frac{1}{\lambda_1}=\mu-\frac{1}{\lambda_2}$ και στη συνέχεια αντικαταστήσουμε στην άλλη θα έχουμε:

$$\left(\frac{1}{\lambda_2}\right)^2-\mu\left(\frac{1}{\lambda_2}\right)+\frac{\mu^2-\sigma^2}{2}=0$$

ή

$$\frac{1}{\lambda_2}=\frac{1}{2}\left[\mu\pm\left(2\sigma^2-\mu^2\right)^{0.5}\right]$$

Δεδομένου ότι μας ενδιαφέρει η παράμετρος λ να είναι απολύτως θετικός αριθμός θα πρέπει να υποβληθούν οι ακόλουθες συνθήκες. Για να εξασφαλίσουμε λοιπόν αυστηρά μια θετική τιμή του λ αρκεί να ισχύει:

$$0<\sqrt{2\sigma^2-\mu^2}<\mu \quad \text{ή} \quad 0.5<\frac{\sigma^2}{\mu^2}<1 \quad \text{ή} \quad 0.702\mu=\frac{\mu}{\sqrt{2}}<\sigma<\mu$$

Πράγματι λοιπόν αποδεικνύουμε ότι όταν ικανοποιείται αυτή η υπόθεση τότε οι παράμετροι λ_1, λ_2 είναι θετικές. Βέβαια για να εξασφαλίσουμε την ύπαρξη μιας προσαρμογής χρησιμοποιώντας τη μέθοδο των ροπών με τρεις ή παραπάνω φάσεις η παραπάνω υπόθεση είναι αναγκαία αλλά όχι απαραίτητη (ικανή).

3.3.4 Μέθοδος των ροπών (3 στάδια-φάσεις)

Έστω ότι κάποιος έχει στα χέρια του την πραγματική κατανομή ορισμένη στο σύνολο των θετικών αριθμών με μέση τιμή μ, διασπορά σ^2 και ροπή τρίτης τάξης μ_3. Θα εξισώσουμε με τις πρώτες ροπές της GED3 με άγνωστες παραμέτρους $\lambda_1, \lambda_2, \lambda_3$. Η συνάρτηση πυκνότητας πιθανότητας της GED3 θα είναι της μορφής:

$$f_3(y) = (\lambda_1\lambda_2\lambda_3)\left[\frac{e^{-\lambda_1 y}}{(\lambda_1-\lambda_2)(\lambda_1-\lambda_3)} + \frac{e^{-\lambda_2 y}}{(\lambda_2-\lambda_1)(\lambda_2-\lambda_3)} + \frac{e^{-\lambda_3 y}}{(\lambda_3-\lambda_1)(\lambda_3-\lambda_2)}\right]$$

Εξισώνοντας τις τρεις πρώτες ροπές θα έχουμε:

$$\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3} = \mu$$

$$\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2} = \sigma^2$$

$$\frac{2}{\lambda_1^3}+\frac{2}{\lambda_2^3}+\frac{2}{\lambda_3^3} = \mu_3$$

Το παραπάνω σύστημα αποτελείται από τρεις μη γραμμικές εξισώσεις με άγνωστες τιμές τις ποσότητες $\lambda_1, \lambda_2, \lambda_3$. Γενικά ένα τέτοιο σύστημα θα μας οδηγήσει σε άτοπο όπως θα αποδειχθεί στο παρακάτω παραδείγμα. Η αποτυχία επίτευξης του στόχου μας οδήγησε στο να ερευνήσουμε την απόκλιση από την ακριβή ποσότητα.

Για παράδειγμα, χρησιμοποιώντας ένα ενδεικτικό παράδειγμα της Weibull κατανομής με παραμέτρους (α, β) = (1.2, 2) τότε αυτή ικανοποιεί την υπόθεση. Προσπαθώντας να την προσεγγίσουμε με την GED3 τότε το παραπάνω σύστημα γράφεται στη μορφή:

$$\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3} = 1.88$$

$$\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2}=2.479$$

$$\frac{2}{\lambda_1^3}+\frac{2}{\lambda_2^3}+\frac{2}{\lambda_3^3}=2.968$$

Όπου η λύση του παραπάνω συστήματος θα μας δώσει $\frac{1}{\lambda_1}=1.3707, \frac{1}{\lambda_2}=-0.2293,$

$\frac{1}{\lambda 3}=0.7499$ το οποίο μας οδηγεί σε άτοπο αφού $\lambda_2<0$.

Επιπλέον αν πάρουμε ως παράδειγμα πάλι την Weibull με παραμέτρους (α, β)= (1.4, 4) τότε πάλι η ικανοπειται η συνθήκη αλλά η λύση του 3×3 του μη γραμμικού συστήματος μας δίνει τα εξής αποτελέσματα:

$$\frac{1}{\lambda_1}+\frac{1}{\lambda_2}+\frac{1}{\lambda_3}=3.6457$$

$$\frac{1}{\lambda_1^2}+\frac{1}{\lambda_2^2}+\frac{1}{\lambda_3^2}=6.9621$$

$$\frac{2}{\lambda_1^3}+\frac{2}{\lambda_2^3}+\frac{2}{\lambda_3^3}=11.0077$$

Τότε το σύστημα αυτό έχει έξι λύσεις, όμως κανένα από τις λύσεις της δεν είναι θετικός αριθμός. Το συμπέρασμα είναι ότι τελικά αυτά τα δύο παραδείγματα δικαιολογούν τον ισχυρισμό μας ότι η συνθήκη είναι ικανή αλλά όχι απαραίτητη για να μας εξασφαλίσει την ύπαρξη της προσαρμογής χρησιμοποιώντας τη μέθοδο των ροπών με τρεις ή περισσότερες φάσεις.

3.4 Σύγκριση των παραμέτρων των αλγορίθμων Quasi-Newton και Nealder-Mead

Σ'αυτή την ενότητα παρουσιάζουμε κάποια μεγέθη που χρησιμοποιούνται για να συγκρίνουμε τα αποτελέσματα των εκτιμήσεων των παραμέτρων που προέκυψαν από τους αλγορίθμους Quasi-Newton και Nealder-Mead.

Οι αλγόριθμοι Quasi-Newton (QN) και Nelder-Mead (NM) θα χρησιμοποιήθουν στην ουσία για την εκτίμηση των παραμέτρων της κατανομής όταν ελαχιστοποιήσουμε τη λογαριθμική συνάρτηση. Σε γενικές γραμμές ο αλγόριθμος Quasi-Newton βρίσκει εφαρμογές στις λεγόμενες gradient methods, όπου οι gradient methods είναι περισσότερο αποδοτικές όταν η συνάρτηση που θα ελαχιστοποιηθεί είναι C^1 δηλαδή έχει συνεχής πρώτες παραγώγους. Ενώ αντίθετα ο αλγόριθμος Nelder-Mead είναι περισσότερο κατάλληλος σε προβλήματα που είναι μη γραμμικά ή όταν έχουν σημεία ασυνέχειας.

Για να μπορέσουμε να συγκρίνουμε τα αποτελέσματα των δύο αλγορίθμων ως προς την αποδότικότητα τους θα χρησιμοποιήσουμε τρία μέτρα απόδοσης, εκ των οποίων τα δύο πρώτα έχουν εισαχθεί από τους Okumara και Dohi (2006) και χρησιμοποιούνται από τους Marshall και Zenga (2009b). Αυτά είναι:

Mean relative distance (MRD)

$$MRD = \frac{\sum \frac{\left|\text{MLEs}^{*}\text{-εκτιμήσεις}\right|}{\text{MLEs}}}{\text{αριθμός των επιτυχών εκτιμήσεων}}$$

*maximum likelihood estimates

Rate of convergence (ROC

$$ROC = \frac{\text{αριθμός των επιτυχών εκτιμήσεων}}{\text{συνολικό αριθμό εκτιμήσεων}} \times 100$$

Έτσι λοιπόν αν η ποσότητα ROC είναι μεγαλύτερη από την MRD τότε ο αλγόριθμος είναι κατάλληλος. Όμως μικρή τιμή της ποσότητας MRD σημαίνει ότι η ταχύτητα σύγκλισης είναι γρήγορη. Το τελευταίο μέτρο απόδοσης του αλγορίθμου (Marshall & Zenga, 2010) είναι:

Rate of algorithm's success (RAS):

$$RAS = \frac{\text{αριθμός των αποδεκτών εκτιμήσεων}}{\text{συνολικό αριθμό εκτιμήσεων}} \times 100$$

Το μέτρο αυτό αξιολογεί τον αριθμό των σωστών εκτιμητριών των αλγορίθμων ή με αλλά λόγια τον ρυθμό επιτυχίας του αλγορίθμου. Συνεπώς μεγάλη τιμή της ποσότητας RAS σημαίνει ότι ο αλγόριθμος λειτούργησε αποδοτικά σε σχέση με τα αποτελέσματα.

3.5 Το PH-plot πρόγραμμα

Το PH-plot πρόγραμμα είναι ένα πρόγραμμα στη Matlab για τη γραφική απεικόνιση των προσαρμοσμένων phase type distributions. Έχει τροποποιηθεί για να μπορούμε με βολικούς τρόπους να αποκτήσουμε τις γραφικές απεικονίσεις καθώς επίσης τα αποτελέσματα της προσαρμογής του δείγματος, όταν κάποιος εφαρμόζει τις PH-κατανομές σε ένα δείγμα. Αυτές οι γραφικές απεικονίσεις περιλαμβάνουν την προσαρμοσμένη συνάρτηση κατανομής και την προσαρμοσμένη συνάρτηση επιβίωσης (όπου σε περίπτωση που έχουμε αποκομμένα δεδομένα χρησιμοποιούμε την εκτιμήτρια Kaplan-Meier), την

προσαρμοσμένη συνάρτηση πυκνότητας πιθανότητας. Επιπλέον εμφανίζονται και κάποια χαρακτηριστικά για το δείγμα που έχουμε προσαρμόσει τέτοια είναι: ο αριθμός των παρατηρήσεων του δείγματος, η μέση τιμή του δείγματος, η διάμεσος, η τυπική απόκλιση, ο συντελεστής διακύμανσης (CV). Δίνοντας λοιπόν δεδομένα από ένα δείγμα για ένα χαρακτηριστικό που μας ενδιαφέρει τότε οι κυριότερες στατιστικές ποσότητες είναι:

- Μέση τιμή: $\hat{m} = \sum_{i=1}^{N} \frac{X_i}{N}$
- Τυπική απόκλιση: $\hat{\sigma} = \sqrt{\sum_{i=1}^{N} (X_i - \hat{m})^2 / (N-1)}$
- Συντελεστής διακύμανσης (CV): $\hat{c} = \frac{\hat{\sigma}}{\hat{m}}$

Επιπλέον υπολογίζει τον αριθμό των παροδικών καταστάσεων, τον πίνακα μεταπήδης των πιθανοτήτων, το αρχικό διάνυσμα των πιθανοτήτων μετάβασης από τις καταστάσεις (1,2,……p), το διάνυσμα της πιθανότητας απορρόφησης , το διάνυσμα για το χρόνο που παρέμεινε στις καταστάσεις (1,2,……p) σε δευτερόλεπτα και σε λεπτά. Ενώ ο υπολογισμός των παραπάνω ποσοτήτων απορρέουν από την εκτίμηση του ζεύγους $\left(\hat{\pi}, \hat{\mathrm{T}}\right)$, το οποίο έχει εκτιμηθεί μέσω του αλγορίθμου EMpht. Για παράδειγμα, ο πίνακας μετάβασης πιθανοτήτων υπολογίζεται ως εξής:

$$P_{jk} = \begin{cases} -\frac{T_{jk}}{T_{jj}}, \varepsilon\alpha\nu\, j \neq k, 1 \leq k \leq p \\ 0, \varepsilon\alpha\nu\, j = k, 1 \leq k \leq p \end{cases}$$

Ενώ το διάνυσμα απορρόφησης των πιθανοτήτων είναι:

$$P_{k\Delta} = \frac{-r_k}{R_{kk}}, 1 \leq \kappa \leq p$$

Και για τον υπολογισμό του χρόνου που απέμειναν στις καταστάσεις (1,2,……p) σε δευτερόλεπτα είναι:

$$m_k = \frac{-1}{T_{kk}}, 1 \le \kappa \le \rho$$

Κεφάλαιο 4ο

Προσαρμογή της Coxian κατανομής

Σε αυτή την ενότητα σκοπός είναι να εκτιμήσουμε τις εκτιμήτριες μέγιστης πιθανοφάνειας προσαρμόζοντας την Coxian κατανομή σε δύο σύνολα δεδομένων. Βέβαια πρώτα θα πρέπει να απαντήσουμε στο ερώτημα γιατί επιλέγουμε αυτή την κατανομή και όχι κάποια άλλη. Οφείλουμε να αναφέρουμε ότι η Coxian κατανομή τις τελευταίες δεκαετίες καθίστανται ολοένα και περισσότερο δημοφιλής σε προβλήματα που αντιπροσωπεύουν το χρόνο επιβίωσης. Στον τομέα της υγειονομικής περίθαλψης, θεωρούνται κατάλληλες για τη μοντελοποίηση του χρόνου παραμονής των ασθενών στο νοσοκομείο και ποιο πρόσφατα για τη μοντελοποιήση του χρόνου αναμονής των ασθενών από τμήματα επειγόντων περιστατικών. Όμως το μειονέκτημα της βρίσκεται ακριβώς στην προσαρμογή της σε δεδομένα. Πραγματικά έχουν γίνει πολλές προσπάθειες στο παρελθόν για να εκτιμήσουμε με ακρίβεια τις παραμέτρους της Coxian κατανομής. Αυτό ακριβώς θα προσπαθήσουμε να επιτεύξουμε σε αυτή την ενότητα.

Η μέθοδος που θα χρησιμοποιήσουμε για τον υπολογισμό εκτιμητριών είναι η εκτιμήτρια μέγιστης πιθανοφάνειας, όπου μεγιστοποιηούμε την λογαριθμική συνάρτηση πιθανοφάνειας, χρησιμοποιώντας το στατιστικό πακέτο R.

Ειδικότερα, στην περίπτωση της κατανομής Coxian η συνάρτηση πυκνότητας πιθανότητας είναι $f(t) = p \cdot \exp(Q \cdot t) q$ επομένως η λογαριθμική συνάρτηση πιθανοφάνειας θα είναι της μορφής:

$$\sum_{i=1}^{n} \log\{p \cdot \exp(Q \cdot t_i) q\}$$

όπου p είναι το αρχικό διάνυσμα πιθανοτήτων ενω ο πίνακας Q διαστάσεως $n \times n$ είναι ο πίνακας μετάβασης πιθανότητων . Στον πίνακα Q όπου λ_i είναι οι διαδοχικές πιθανότητες μετάβασης ανάμεσα στις καταστάσεις από τις $1,...,n-1$ ενω με μ_i συμβολίζονται οι πιθανότητες απο τις μεταβατικές καταστάσεις $1,...,n-1$ στην μοναδική κατάσταση απορρόφησης n. Τέλος το διάνυσμα στήλη q διαστάσεως $n \times 1$ περιλαμβάνει τις πιθανότητες απορρόφησης.

$$p = (1, 0, 0, ..., 0, 0)$$

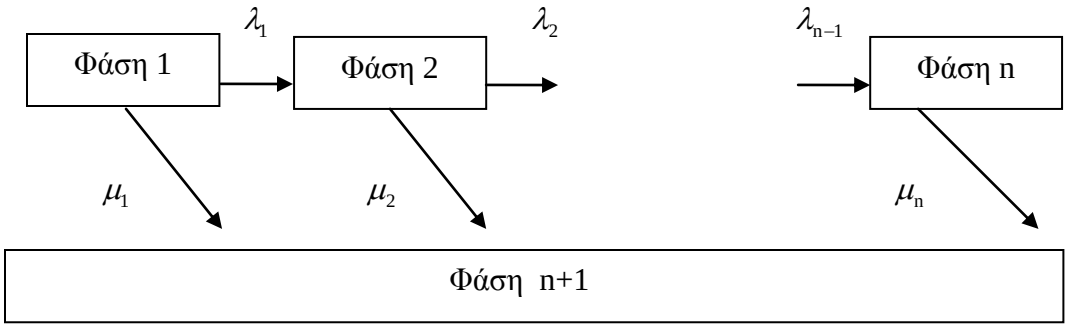
$$Q = \begin{pmatrix} -(\lambda_1 + \mu_1) & \lambda_1 & 0 & \cdots & 0 & 0 \\ 0 & -(\lambda_2 + \mu_2) & \lambda_2 & \cdots & 0 & 0 \\ \vdots & \vdots & \vdots & \vdots & \vdots & \vdots \\ 0 & 0 & 0 & \cdots & -(\lambda_{n-1} + \mu_{n-1}) & \lambda_{n-1} \\ 0 & 0 & 0 & \cdots & 0 & -\mu_n \end{pmatrix}$$

$$q = -Qe = (\mu_1, \quad \mu_2, \quad \cdots, \quad \mu_n)^T$$

ικανοποιώντας τους εξής περιορισμούς: $0 \le \mu_j \le 1, \quad j = 1,...,n$

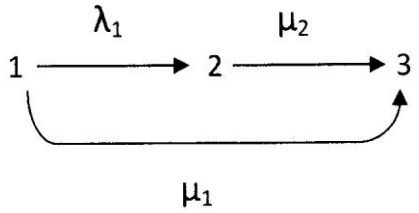
$$0 \le \lambda_j \le 1, \quad j = 1,...,n-1$$

Το παρακάτω διάγραμμα ροής απεικονίζει στην ουσία τις μεταβάσεις που πραγματοποιούνται σε κάθε στάδιο.



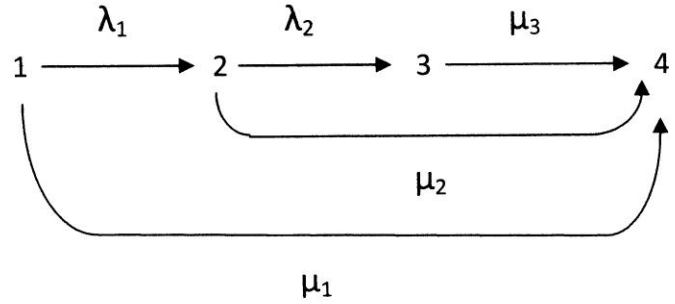
Εικόνα 4 .1 Απεικόνηση της Coxian κατανομής

Έτσι λοιπόν με βάση τα παραπάνω εαν προσαρμόσουμε τη Coxian κατανομή με δύο φάσεις τότε το μοντέλο θα είναι το εξής:



$$\text{όπου } p = (1,0), \quad Q = \begin{pmatrix} -\lambda_1 - \mu_1 & \lambda_1 \\ 0 & -\mu_2 \end{pmatrix}, \quad q = -Qe = \begin{pmatrix} \mu_1 \\ \mu_2 \end{pmatrix}$$

Ενώ αν προσαρμόσουμε τη Coxian κατανομή με τρείς φάσεις τότε το μοντέλο θα είναι το εξής:



$$\text{όπου} \quad \mathrm{p} = (1,0,0), \quad \mathrm{Q} = \begin{pmatrix} -\lambda_1 - \mu_1 & \lambda_1 & 0 \\ 0 & -\lambda_2 - \mu_2 & \lambda_2 \\ 0 & 0 & -\mu_3 \end{pmatrix}, \quad \mathrm{q} = -\mathrm{Q1} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix}$$

4.1 Προσαρμογή της Coxian κατανομής ($1^{ο}$ σύνολο δεδομένων)

Στην υποενότητα αυτή, αρχικά θα προσαρμόσουμε σε δύο σύνολα αποκομμένων δεδομένων τη Coxian κατανομή με δύο φάσεις και στη συνέχεια με τρεις φάσεις. Τα δεδομένα της πρώτης εφαρμογής (Klein & Moeschberger, 1997) έχουν παρθεί από το Channing house, ένα κέντρο περίθαλψης ηλικιωμένων, το οποίο βρίσκεται στο Palo Alto στη California. Τα δεδομένα αυτά περιγράφουν το χρόνο παραμονής (εκφρασμένος σε μήνες) των ηλικιωμένων στο κέντρο αυτό εως τον θάνατο. Επιπλέον το σύνολο του δείγματος ανέρχεται στους 462 ηλικιωμένους, εκ των οποίων οι 97 είναι άντρες ενω οι υπόλοιπες 365 είναι γυναίκες. Επίσης η δειγματοληψία αυτή είχε πραγματοποιηθεί κατα την περίοδο Ιανουαρίου του 1964 εως τον Ιούλιο του 1975. Η προσαρμογή θα γίνει ξεχωριστά για τις γυναίκες και τους άνδρες. Τέλος να αναφέρουμε ότι χαρακτηριστικό γνώρισμα αυτών των ανθρώπων ήταν οτι καλύπτονταν όλοι απο το πρόγραμμα περίθαλψης που τους το

παρείχε το κέντρο φροντίδας ηλικιωμένων χωρίς οικονομική επιβάρυνση απο μέρους τους.

Επιπλέον, όσον αφορά τα δεδομένα και για τις δύο ομάδες (άντρες και γυναίκες) μας δίνονται δύο στήλες. Η πρώτη στήλη μας δίνει την ηλικία του ασθενούς σε μήνες όταν θα εισαχθεί στο κέντρο αυτό, ενω η δεύτερη στήλη μας δίνει την ηλικία του θανάτου του σε μήνες. Άρα λοιπόν επειδή εμείς στην ουσία μελετάμε τη διάρκεια παραμονής των ασθενών στο κέντρο, θα αφαιρέσουμε την πρώτη στήλη από την δεύτερη.

Τα αποτελέσματα που μας δίνει το στατιστικό πακέτο R είναι:

Coxian κατανομή με 2 φάσεις - Γυναίκες	*Εκτιμήτριες μέγιστης πιθανοφάνειας*
$\hat{\lambda}_1$	**0.000001**
$\hat{\mu}_1$	**0.33368848**
$\hat{\mu}_2$	**0.01046703**
log l	**-1201.177**

Πίνακας 4.1 Προσαρμογή της Coxian κατανομής για τις γυναίκες

*Η προσαρμογή της Coxian κατανομής με 3 φάσεις δεν ήταν εφικτή.

Άρα λοιπόν εαν προσαρμόσουμε την Coxian κατανομή με δύο φάσεις οι πίνακες Q και q θα είναι:

$$Q = \begin{pmatrix} -0.000001 - 0.33368848 & 0.000001 \\ 0 & 0.01046703 \end{pmatrix} = \begin{pmatrix} -0.33368948 & 0.000001 \\ 0 & 0.01046703 \end{pmatrix}$$

και

$$q = -Qe = \begin{pmatrix} 0.33368848 \\ 0.01046703 \end{pmatrix}$$

Coxian κατανομή με 3 φάσεις - Άνδρες	*Εκτιμήτριες μέγιστης πιθανοφάνειας*
$\hat{\lambda}_1$	**0.000000001**
$\hat{\mu}_1$	**0.012141517**
$\hat{\lambda}_2$	**0.000357286**
$\hat{\mu}_2$	**0.001554072**
$\hat{\mu}_3$	**0.004992286**
log l	**-276.0831**

Πίνακας 4.2 Προσαρμογή της Coxian κατανομής με 3 φάσεις στους άντρες

*Η προσαρμογή της Coxian κατανομής με 2 φάσεις δεν ήταν εφικτή.

Άρα λοιπόν εαν προσαρμόσουμε την Coxian κατανομή με τρείς φάσεις οι πίνακες Q και q θα είναι:

$$Q=\begin{pmatrix} -0.000000001 & -0.012141517 & 0.000000001 & 0 \\ 0 & & -0.000357286 & -0.001554072 & 0.000357286 \\ 0 & & 0 & & -0.004992286 \end{pmatrix}$$

$$=\begin{pmatrix} -0.012141518 & 0.000000001 & 0 \\ 0 & -0.001911358 & 0.000357286 \\ 0 & 0 & -0.004992286 \end{pmatrix}$$

$$q = -Qe = \begin{pmatrix} 0.012141517 \\ 0.001554072 \\ 0.004992286 \end{pmatrix}$$

Παρατηρούμε λοιπόν ότι ικανοποιούνται οι περιορισμοί για τις παραμέτρους του προσαρμοσμένου μοντέλου: $0 \le \mu_j \le 1, \quad j = 1,2,3$ και $0 \le \lambda_j \le 1, \quad j = 1,2$.

4.2 Προσαρμογή της Coxian κατανομής (2^{o} σύνολο δεδομένων)

Τα δεδομένα της δεύτερης εφαρμογής έχουν παρθεί απο το Πανεπιστήμιο της Massachusetts, στα πλαίσια μιας έρευνας που είχε πραγματοποιηθεί για τις επιπτώσεις του AIDS. Η έρευνα αυτή διέρκησε 5 χρόνια απο το 1989-1994 και το μέγεθος του δείγματος ανερχόταν στις 628 παρατηρήσεις. Χαρακτηριστικό γνώρισμα των ανθρώπων που συμμετείχαν σε αυτή την έρευνα ήταν ότι έκαναν χρήση ναρκωτικών ουσιών. Τα δεδομένα αυτά περιγράφουν το χρόνο ζωής τους ύστερα από την χορήγηση κάποιας θεραπείας. Η θεραπεία αυτή είχε διπλό σκοπό, απο μία να μειώσει την κατάχρηση των ναρκωτικών ουσιών και απο την άλλη να εμποδίσει την συμπεριφορά του ιού HIV, ο οποίος συνέχεια εναλλασσόταν και έτσι οι επιστήμονες δεν μπορούσαν να βρούν το κατάλληλο αντίδοτο για να το εξοντώσουν. Τα αποτελέσματα είναι τα εξής:

Coxian κατανομή με 2 φάσεις	*Εκτιμήτριες μέγιστης πιθανοφάνειας*
$\hat{\lambda}_1$	**0.003221707**
$\hat{\mu}_1$	**0.338847789**
$\hat{\mu}_2$	**0.000001000**
log l	**-6472.408**

Πίνακας 4.3 Προσαρμογή της Coxian κατανομής με 2 φάσεις

Άρα λοιπόν εαν προσαρμόσουμε την Coxian κατανομή με δύο φάσεις οι πίνακες Q και q θα είναι:

$$Q = \begin{pmatrix} -0.003221707-0.338847789 & 0.003221707 \\ 0 & 0.000001000 \end{pmatrix} = \begin{pmatrix} -0.342069496 & 0.003221707 \\ 0 & 0.000001000 \end{pmatrix}$$

$$q = -Qe = \begin{pmatrix} 0.338847789 \\ 0.000001000 \end{pmatrix}$$

Coxian κατανομή με 3 φάσεις	*Εκτιμήτριες μέγιστης πιθανοφάνειας*
$\hat{\lambda}_1$	**0.044423061**
$\hat{\mu}_1$	**0.000001000**
$\hat{\lambda}_2$	**0.008053097**
$\hat{\mu}_2$	**0.000001000**
$\hat{\mu}_3$	**0.030547122**
log l	**-3149.080**

Πίνακας 4.4 Προσαρμογή της Coxian κατανομής με 3 φάσεις

Άρα λοιπόν εαν προσαρμόσουμε την Coxian κατανομή με τρείς φάσεις οι πίνακες Q και q θα είναι:

$$Q = \begin{pmatrix} -0.044423061 & -0.000001000 & 0.044423061 & 0 \\ 0 & & -0.008053097 & -0.000001000 & 0.008053097 \\ 0 & & 0 & -0.030547122 \end{pmatrix}$$

$$= \begin{pmatrix} -0.044424061 & 0.044423061 & 0 \\ 0 & -0.008054097 & 0.008053097 \\ 0 & 0 & -0.030547122 \end{pmatrix}$$

$$q = -Q1 = \begin{pmatrix} 0.000001000 \\ 0.000001000 \\ 0.030547122 \end{pmatrix}$$

Παρατηρούμε λοιπόν ότι καταρχήν, ικανοποιούνται οι περιορισμοί για τις παραμέτρους του προσαρμοσμένου μοντέλου: $0 \le \mu_j \le 1, \quad j = 1,2,3$ και $0 \le \lambda_j \le 1, \quad j = 1,2$. Επιπλέον από τους πίνακες 4.3 και 4.4 οι τιμές της λογαριθμικής συνάρτησης πιθανοφάνειας είναι:

$$n = 2 \qquad \log l = -6472.408$$

$$n = 3 \qquad \log l = -3149.080$$

Συγκρίνοντας τις παραπάνω τιμές της λογαριθμικής συνάρτησης πιθανοφάνειας παρατηρούμε ότι στα δεδομένα μας προσαρμόζεται καλύτερα η Coxian κατανομή

με 3 φάσεις. Προφανώς αυτό σημαίνει ότι για τους ασθενείς που έχουν μολυνθεί απο τον ιό HIV θα πρέπει να υποβληθούν σε μια σειρά απο 3 διαδοχικές εξετάσεις εως ότου αποχωρήσουν απο το σύστημα με ρυθμό μ_3.

Συμπεράσματα

Σε αυτή τη διπλωματική εργασία έγινε μια εισαγωγή στη μορφή των phase type distributions, μιας οικογένειας κατανομών οι οποίες προέρχονται από τις λεγόμενες πινακοεκθετικές κατανομές.

Αρχικά παρουσιάστηκε η μορφή αυτών των κατανομών, έπειτα παραδείγματα κατανομών τα οποία ανήκουν στη κατηγορία των phase type όπως, είναι η εκθετική κατανομή, η Γάμμα κατανομή, Coxian κατανομή, η υπερεκθετική κατανομή, η Erlang κατανομή και άλλα. Στην πρώτη ενότητα δώθηκε ιδιαίτερη έμφαση στον υπολογισμό της συνάρτησης πυκνότητας πιθανότητας, της συνάρτησης κατανομής, της ροπογεννήτριας συνάρτησης, της μέσης τιμής, της διασποράς κτλ αυτών των κατανομών. Επιπλέον στο πρώτο αρχικό κεφάλαιο θεωρήσαμε σημαντικό, ύστερα απο μια ανασκόπηση στην βιβλιογραφία να παρουσιάσουμε μια σειρά απο σχετικές εφαρμογές των phase type κατανομών. Οι περισσότερες εφαρμογές που έχουν πραγματοποιηθεί προέρχονται κυρίως από τον κλάδο της βιοστατιστικής και τον κλάδο της ουράς αναμονής.

Στη συνέχεια ύστερα απο αυτή την αναλυτική καταγραφή των ιδιοτήτων των phase type κατανομών, προχωρήσαμε στην παρουσιάση των μεθόδων για την εκτίμηση των παραμέτρων όταν προσαρμόζουμε τις κατανομές αυτές σε πραγματικά δεδομένα. Η εκτίμηση των παραμέτρων των phase type distributions δεν είναι μια τετριμμένη διαδικασία. Είναι γεγονός ότι υπάρχουν αρκετά προβλήματα στην προσαρμογή των των phase type distributions σε πραγματικά

δεδομένα. Οι δυσκολίες οφείλονται στη μη γραμμικότητα του προβλήματος, την ταυτόχρονη βελτίωση του αριθμού των παραμέτρων και της μη μοναδικότητας της μορφής των PH-κατανομών.

Υπάρχουν τρεις βασικές τεχνικές οι οποίες εξελίσσονται συνεχώς με την πάροδο των χρόνων. Ωστόσο οι περισσότερες επαναληπτικές μέθοδοι βασίζονται κυρίως στην μέθοδο μέγιστης πιθανοφάνειας (maximum likelihood). Επιπλέον σε αυτή την ενότητα κάναμε μια ιδιαίτερη αναφορά στον λεγόμενο αλγόριθμο EM (ecxpectetion-maximaztion), μια επαναληπτική μέθοδος για την εύρεση εκτιμητών μέγιστης πιθανοφάνειας των παραμέτρων μιας δοθείσας κατανομής. Η χρήση του στη εφαρμογή αυτής της διπλωματικής εργασίας παρουσίασε αρκετές δυσκολίες προσαρμογής.

Έτσι λοιπόν στην τελευταία ενότητα όπου πραγματοποιήθηκε η εφαρμογή, αποφασίσαμε να χρησιμοποιήσουμε για την εκτίμηση των παραμέτρων της Coxian κατανομής το στατιστικό πακέτο R εφαρμόζοντας φυσικά την μέθοδο μέγιστης πιθανοφάνειας. Ωστόσο παρουσιάστηκαν δυσκολίες στην εφαρμογή τους. Επιπλέον τα μοντέλα αυτά είναι αρκετά περίπλοκα ως προς το πλήθος των παραμέτρων και για αυτό περιοριστήκαμε στην περίπτωση της Coxian κατανομής με 2 και 3 καταστάσεις.

Συνοψίζοντας, καταλήγουμε στο ότι τα μοντέλα αυτά παρουσιάζουν δυσκολίες στην προσαρμογή τους και φαίνεται να εξαρτώνται πολύ από τις τιμές εκκίνησης των παραμέτρων στη διαδικασία εκτίμησης.

Βιβλιογραφία

Aalen OO (1995) Phase type distributions in survival analysis. *Scandinavian Journal of Statistics* **22**: 447-463.

Asmussen S, Nerman O & Olsson M (1996) Fitting phase-type distributions via the EM algorithm. *Scandinavian Journal of Statistics* **23**: 419-441.

Augustin R & Buscher K-J (1982) Characteristics of the Cox distribution. *ACM Sigmetrics Performance Evaluation Review* **12**: 22-32.

Bladt M (2005) A review on phase type distributions and their use in risk theory. *ASTIN Bulletin* **35**: 145-161.

Bobbio A & Cumani A (1992) ML estimation of the parameters of a PH distribution in triangular canonical form. In: Balbo G & Serazzi G (eds) *Computer Performance Evaluation*. Amsterdam: Elsevier, pp. 33-46.

Botta RF, Harris CM & Marchal WG (1987) Characterizations of generalized hyperexponential distribution functions. *Communications in Statistics: Stochastic Models* **3**: 115-148.

Dempster AP, Laird NL & Rubin DB (1977) Maximum likelihood from incomplete data via the EM algorithm (with Discussion). *Journal of the Royal Statistical Society Series B* **39**: 1-38.

Fackrell M (2009) Modelling healthcare systems with phase-type distributions. *Health Care Management Science* **12**: 11-26.

Faddy MJ (1990) Compartmental models with phase-type residence-time distributions. *Applied Stochastic Models and Data Analysis* **6**:121-127.

Faddy MJ (1993) A structured compartmental model for drug kinetics. *Biometrics* 49:243-248.

Faddy MJ (1994) Examples of fitting structured phase-type distributions. *Applied Stochastic Models in Business and Industry* **10**: 247-255.

Faddy MJ (1998) On inferring the number of phases in a Coxian phase-type distribution. *Communications in Statistics: Stochastic Models* **14**: 407-417.

Faddy MJ & McClean SI (2005) Markov chain modelling for geriatric patient care. *Methods of Information in Medicine* **44**: 369-373.

Faddy MJ & McClean SI (1999) Analysing data on lengths of stay of hospital patients using phase-type distributions. *Applied Stochastic Models in Business and Industry* **15**: 311-317.

Johnson MA (1993) Selecting parameters of phase distributions: combining nonlinear programming, heuristics, and Erlang distributions. *ORSA Journal on Computing* **5**: 69-83.

Johnson MA & Taaffe MR (1989) Matching moments to phase distributions. *Communications in Statistics: Stochastic Models* **5**: 711-743.

Johnson MA & Taaffe MR (1990a) Matching moments to phase distributions: nonlinear programming approaches. *Communications in Statistics: Stochastic Models* **6**: 259-281.

Johnson MA & Taaffe MR (1990b) Matching moments to phase distributions: density function shapes. *Communications in Statistics: Stochastic Models* **6**: 283-306.

Johnson MA & Taaffe MR (1991) An investigation of phase-distribution moment-matching algorithms for use in queueing models. *Queueing Systems* **8**: 129-147.

Klein JP & Moeschberger ML (1997) *Survival Analysis: Techniques for Censored and Truncated Data*. New York: Springer.

Lang A & Arthur JL (1996) Parameter approximation for phase-type distributions. In: Chakravarthy S & Alfa AS (eds) *Matrix-Analytic Methods in Stochastic Models.* Lecture Notes in Pure and Applied Mathematics, Vol. 189. New York: Marcel Dekker, pp. 151-206.

Marshall AH & Zenga M (2009a) Simulating Coxian phase type distributions for patient survival. *International Transactions in Operational Research* **16**: 213-226.

Marshall AH & Zenga M (2009b) Recent developments in fitting Coxian phase-type distributions in healthcare. In Sakalauskas L, Skiadas C & Zavadskas EK (eds) *ASMDA-2009: Selected papers*. Vilnius Gediminas Technical University, pp. 482-485.

Marshall AH & Zenga M (2010) Experimenting with the Coxian phase-type distribution to uncover suitable fits. *Methodology and Computing in Applied Probability*, published online 4th April 2010, DOI: 10.1007/s11009-010-9174-y.

McLachlan GJ & Krishnan T (1997) *The EM Algorithm and Extensions*. New York: Wiley.

Neuts MF (1975) Probability distributions of phase type. In: *Liver amicorum Prof. Emeritus H.Florin.* Louvain: University of Louvain, pp. 173-206.

Neuts MF (1981) *Matrix-geometric solutions in stochastic models: an algorithmic approach*. Baltimore: The Johns Hopkins University Press.

O'Cinneide CA (1989) On non-uniqueness of representations of phase type distributions. *Communications in Statistics: Stochastic Models* **5**: 247-259.

O'Cinneide CA (1990) Characterization of phase type distributions. *Communications in Statistics: Stochastic Models* **6**: 1-57.

O'Cinneide CA (1999) Phase-type distributions: open problems and a few properties. *Communications in Statistics: Stochastic Models* **15**: 731-757.

Okumara H & Dohi T (2006) Building phase-type software reliability. In: *Software Reliability Engineering, 2006: ISSRE '06*, pp. 289-298.

Olsson M (1996) Estimation of phase-type distributions from censored data. *Scandinavian Journal of Statistics* **23**: 443-460.

Parr WC & Schucany WR (1980) Minimum distance and robust estimation. *Journal of the American Statistical Association* **75**: 616-624.

Riska A, Diev V & Smirni E (2002) Efficient fitting of long tailed data sets into PH distributions. *Internet Performance Symposium, IEEE GlobeCom 2002.* Taipei, Taiwan.

Schmickler L (1992) MEDA: Mixed Erlang distributions as phase-type representations of empirical distribution functions. *Communications in Statistics: Stochastic Models* **8**: 131-156.

Καρώνη Χ (2009) *Μοντέλα Αξιοπιστίας και Επιβίωσης*. Εκδόσεις Συμεών, Αθήνα.

Παράρτημα

Δεδομένα – Αποτελέσματα των εφαρμογών με το πακέτο R

Coxian κατανομή με 2 φάσεις – 1ο σετ δεδομένων (Γυναίκες)

```
> library(expm)
> xx<-read.table("c:/EFARMOGH/all-women.txt")
> x<-xx[,1]
> x
 [1] 130 119 118 117 124  89 122  56  40  22  97  79  70 128 106 110  60  41  90 101 115
132 102 119 109 104 115 129  98 112 127 111 125
 [34] 120   9 128  90 113  14 129 108 110  83 111  71 113  15 123  96  56 107  96 108
112  59 127 118 101  50 111  32  61  74  54  18  58
 [67]  31  23  47  22   7  13  67  12 136 100  83  90  73   9  77  34  24  63  68  12  71  67
39  68  81  37  46  19  36  82  57  73  22
[100]  43  38  71  37  19  67  10  86  69  53  11  52  45  44  19   2  56   6   0  39  74   8  39
61  44  93  22  35  74  54  16  15  19
[133] 137  74  98 137  11  65  58  98  67 137 117 137 137  35 137 137  35   8 137   9 137
34 137 137 137  12  68   4 137 123 137 137 137
[166]  27 137  89 132  13  48 137  72  85  65 137 137 137 108 137 137 137   2 137 137
103 137 137 137  41 137  35 137 137 137  32  16   3
[199]   3  87   7 137 137 137  27  52 137  80  27   7 137  89 137 137 137 104 123  59  28
133 137  91 137 137 137  37 137 137  23 137 122
[232] 137 137  12   0 137 137 137  42 137 137  35 137 137 137 137 123   2 137 137  86
64 137 137  31 137  36  18 137 137  28 137 116  46
[265]  52 137 137  87  60  80  95 137 137  35  31  57  15 -66  57  53 137 137 137 137 137
94 137  26 103  33 137  88 137 137 115  58 137
[298] 137 137 137 135 137  38 137  35 137  49   8   3 137  70 137 137 115   4 137  62
137 137  30 137 137 137 120 137 137  36  58 137  18
```

```
[331] 137   7  61   8   0   0  12  92  59   4   6  24   4  18  11 137 105 137 137  99  50 137
137 137 137  41  33   2  17   3  45  26  19   8
[364]  65   5
> c<-xx[,2]
> c
  [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [38] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
 [75] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[112] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[149] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[186] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[223] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[260] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[297] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[334] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

> pp<-function(theta,x,c) {
+  lam1<-theta[1]
+  mu1<-theta[2]
+  mu2<-theta[3]
+  qq<-matrix(c(-lam1-mu1,0,mu1,-mu2), nc=2)
+   p<-matrix(c(1,0),nrow=1)
+  q<-matrix(c(mu1,mu2),nc=1)
+  zz1<-rep(1, times = length(x))
+  zz2<-rep(1, times = length(x))
+  unit<- matrix(c(1,1),nc=1)
+  for (i in 1:length(x))
+  {zz1[i]<-p%*%(expm(qq*x[i]))%*%q}
```

```
+  {zz2[i]<-p%*%(expm(qq*x[i]))%*%unit}
+  ogl<-sum(c*log(zz1)+(1-c)*log(zz2))
+  return(-logl)}
> optim(c(0.0005,0.0005,0.0005),pp,x=x,c=c,method="L-BFGS-B", lower=c(1e-6,1e-6,
1e-6))
$par
[1] 0.00000100 0.33368848 0.01046703
$value
[1] 1201.177
```

Coxian κατανομή με 3 φάσεις – 1ο σετ δεδομένων (Γυναίκες)

Η προσαρμογή δεν ήταν εφικτή

Coxian κατανομή με 2 φάσεις – 1ο σετ δεδομένων (Άνδρες)

Η προσαρμογή δεν ήταν εφικτή

Coxian κατανομή με 3 φάσεις – 1ο σετ δεδομένων (Άνδρες)

```
library(expm)
xx<-read.table("d:/chrys/diplomatikes/metaptyxiakes/rapai/all-men.txt")
x<-xx[,1]
c<-xx[,2]
> x
  [1] 127 108 113  42 120 106 100  82 100 108  43  35 114  17  66 111 118 135  65
 [20]  69  34   1  46  41 -72  31  71  83   8   2  33  73  72  40  74  26  60  72
 [39]  85  22  36  38  41  63  21  34 137  32 137 137  78  74  24  76  26 137 135
 [58] 137 136 110  26 137 103  29 137 137  36 137 137 137   3  27 137 126 137 137
 [77]  22 107  87 100  95 137  31  57   0  40 137  44  34 137 137  36   8  20   4
 [96]  13   5
```

```
> c
 [1] 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
[39] 0 0 0 0 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
[77] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1

> pp<-function(theta,x,c) {
+ lam1<-theta[1]
+ mu1<-theta[2]
+ mu2<-theta[3]
+ lam2<- theta[4]
+ mu3<-theta[5]
+ qq<-matrix(c(-lam1-mu1, 0, 0, lam1, -lam2-mu2, 0, 0, lam2, -mu3), nc=3)
+ p<-matrix(c(1,0,0), nrow=1)
+ q<-matrix(c(mu1,mu2,mu3),nc=1)
+ zz1<-rep(1, times = length(x))
+ zz2<-rep(1, times = length(x))
+ unit<- matrix(c(1,1,1),nc=1)
+ for (i in 1:length(x))
+ {zz1[i]<-p%*%(expm(qq*x[i]))%*%q}
+ {zz2[i]<-p%*%(expm(qq*x[i]))%*%unit}
+ logl<-sum(c*log(zz1)+(1-c)*log(zz2))
+ return(-logl)}
> optim(c(0.0005,0.005,0.0005,0.0001,0.005),pp,x=x,c=c,method="L-BFGS-B",
lower=c(1e-9,1e-6, 1e-6))
$par
[1] 0.000000001 0.012141517 0.001554072 0.000357286 0.004992286

$value
[1] 276.0831
```

Coxian κατανομή με 2 φάσεις – 2ο σετ δεδομένων

```
> library(expm)
> xx<-read.table("c:/EFARMOGH/ef3.txt")
> x<-xx[,1]
> x
  [1] 188  26 207 144 551  32 459  22 210 184   5 212  87 598 260
 [16] 210  84 196  19 441 449 659  21  53 225 161  87  89  44  37
 [31] 523 226 259 289 103 624  68  57  65  79 559  79  87  91 297
 [46]  45 246  37  37 538 541 184 122   5 156 121 231 111  38  15
 [61]  54 127 105  11 153  11  79  46 655 166   6  95  83 151 220
 [76] 227 343 119  43 545  47  15 805 321 167 107 491  35 123 597
 [91] 762  13  31 228 553 190 307  73 208 267   2 169 125 655  70
[106] 398  59 122  29   8  96 1172 734  26  84 171 159   5   7 763
[121] 104 162  90 373 115  67  30   8 168  70 130 285 569  87 310
[136]  87 544 156 658 273 168  83   4 708 137 259 560 586 190 720
[151] 544   3 494 541  94 567  55  93 276  46   4 250 106 552  90
[166] 203  67 559 106 374 630  61 560 547 568 490 222  56 282  35
[181] 603 194 148 354 164  94  65 567 634 633 127 477 436 226 362
[196] 552 144 242 564 299 167 380 120 248 218 115 224 132 148 593
[211]  26 113  32 292  89  21 364 142 188   4  92  56 110 555 220
[226]  23 285  90  59 156 194 142  57 279 118 567 562 239 578 551
[241] 313 560  54 198 164 325  62  45  53 253  51 540 317 437 136
[256] 115 175 442 122 181 180  51 541 121 328   9 166 556 104 102
[271] 533 144 545 537 625   6 307 290  20  74 100 555 152 115  92
[286] 554  92  69  25 501  86  99  87 136 106 220  36 162 116 175
[301] 209 545 245 176  14 113 159 354 174  23  26  98  23 555 290
```

[316] 543 274 536 3 5 119 164 548 175 539 155 14 187 65 159

[331] 96 243 85 4 121 659 260 621 199 565 183 122 170 15 268

[346] 79 23 100 98 81 546 58 569 575 17 91 57 499 123 143

[361] 130 471 74 85 95 36 19 38 539 567 186 546 24 540 157

[376] 86 231 271 14 75 147 105 324 538 300 73 65 568 84 22

[391] 44 7 540 21 537 186 40 287 538 30 516 268 568 131 399

[406] 78 80 102 3 124 80 23 274 10 459 10 176 332 119 217

[421] 285 576 106 81 47 76 348 20 306 192 216 189 403 193 28

[436] 150 99 510 306 101 102 510 69 503 52 547 168 461 538 349

[451] 44 548 12 6 575 589 408 232 143 582 134 7 548 81 170

[466] 29 78 81 369 69 115 361 245 233 227 97 547 224 211 220

[481] 54 192 138 107 597 226 434 106 180 557 556 619 546 85 233

[496] 102 548 99 36 32 78 502 71 59 115 533 10 274 255 503

[511] 256 9 550 386 547 45 58 124 540 243 549 12 51 562 94

[526] 204 238 140 120 154 177 119 83 130 11 159 211 33 72 161

[541] 191 181 546 540 76 7 44 103 79 339 90 542 384 255 431

[556] 587 198 551 110 541 242 537 56 34 567 549 133 226 401 14

[571] 548 224 540 237 354 123 170 203 360 139 215 129 396 547 547

[586] 71 168 228 551 654 51 548 231 280 184 86 560 46 200 244

[601] 182 296 24 142 120 47 519 248 31 567 353 458 554 116 74

[616] 10 355 232 68 48 60 50 51 126 18 35 379 377

> c<-xx[,2]

> c

[1] 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 0 1

[38] 1 1 1 0 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1

[75] 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 0 0 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1

[112] 0 0 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0 1 1 1 1 0 1 1 0 0

```
[149] 1 0 0 1 1 0 1 0 1 1 1 1 1 1 0 1 1 1 1 1 0 1 0 0 0 1 1 1 1 1 0 1 1 1 1
[186] 1 1 0 0 0 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1
[223] 1 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 0 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1
[260] 1 1 1 0 1 1 1 1 0 1 1 0 1 0 0 0 1 1 1 1 1 1 0 1 1 1 0 1 1 1 0 1 1 1 1 1 1
[297] 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 1 1 1 1 0 1 0 1 1 1 1 1 1 1
[334] 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0
[371] 1 0 1 0 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 0 1 0 1 1 1 0 1 1 1 0 1 1 1 1
[408] 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 1 0
[445] 1 0 1 1 0 1 1 0 1 1 0 0 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 1 1
[482] 1 1 1 0 1 1 1 1 0 0 0 0 1 1 1 0 1 1 1 1 1 1 1 1 0 1 1 1 0 1 1 0 1 0 1 1 1
[519] 0 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 1 1 1
[556] 0 1 0 1 0 1 0 1 1 0 0 1 1 1 1 0 1 0 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 0 0 1 0
[593] 1 1 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
```

```
pp<-function(theta,x,c) {
+ lam1<-theta[1]
+ mu1<-theta[2]
+ mu2<-theta[3]
+ qq<-matrix(c(-lam1-mu1,0,mu1,-mu2), nc=2)
+  p<-matrix(c(1,0),nrow=1)
+ q<-matrix(c(mu1,mu2),nc=1)
+ zz1<-rep(1, times = length(x))
+ zz2<-rep(1, times = length(x))
+ unit<- matrix(c(1,1),nc=1)
+ for (i in 1:length(x))
+ {zz1[i]<-p%*%(expm(qq*x[i]))%*%q}
+ {zz2[i]<-p%*%(expm(qq*x[i]))%*%unit}
```

```
+ logl<-sum(c*log(zz1)+(1-c)*log(zz2))
+  return(-logl)}
> optim(c(0.01,0.01,0.01),pp,x=x,c=c,method="L-BFGS-B", lower=c(1e-6,1e-6, 1e-6))
$par
[1] 0.003221707 0.338847789 0.000001000
$value
[1] 6472.408
```

Coxian κατανομή με 3 φάσεις – 2ο σετ δεδομένων

```
> pp<-function(theta,x,c) {
+ lam1<-theta[1]
+ mu1<-theta[2]
+ mu2<-theta[3]
+ lam2<- theta[4]
+ mu3<-theta[5]
+ qq<-matrix(c(-lam1-mu1, 0, 0, lam1, -lam2-mu2, 0, 0, lam2, -mu3), nc=3)
+ p<-matrix(c(1,0,0), nrow=1)
+ q<-matrix(c(mu1,mu2,mu3),nc=1)
+ zz1<-rep(1, times = length(x))
+ zz2<-rep(1, times = length(x))
+ unit<- matrix(c(1,1,1),nc=1)
+ for (i in 1:length(x))
+ {zz1[i]<-p%*%(expm(qq*x[i]))%*%q}
+ {zz2[i]<-p%*%(expm(qq*x[i]))%*%unit}
+ logl<-sum(c*log(zz1)+(1-c)*log(zz2))
+ return(-logl)}
```

> optim(c(0.01,0.01,0.01,0.01,0.01),pp,x=x,c=c,method="L-BFGS-B", lower=c(1e-6,1e-6, 1e-6, 1e-6, 1e-6))

$par

[1] 0.009651864 0.004678539 0.004051199 0.009860004 0.013455351

$value

[1] 3050.798