Πανεπιστήμιο Αιγαίου

Τμήμα Στατιστικής και Αναλογιστικής Επιστήμης

ΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

ΣΤΑΤΙΣΤΙΚΗ ΜΟΝΤΕΛΟΠΟΙΗΣΗ ΤΟΥ ΑΡΙΘΜΟΥ ΤΩΝ ΑΙΤΗΣΕΩΝ ΑΠΟΖΗΜΙΩΣΕΩΝ (OUTSTANDING CLAIM COUNTS) ΣΤΗΝ ΑΝΑΛΟΓΙΣΤΙΚΗ ΕΠΙΣΤΗΜΗ

Γιαγός Βασίλειος

Επιβλέπων καθηγητής: Κατσής Αθανάσιος

15 Απριλίου 2005

Καρλόβασι Σάμος

Στοιχεία της Πτυχιακής Εργασίας

Συγγραφέας:	Γιαγός Βασίλειος
Author:	Giagos Vasilios
Τμήμα:	Τμήμα Στατιστικής και Αναλογιστικής Επιστήμης, Πανεπιστήμιο Αιγαίου
Department:	Department of Statistical and Actuarial Science, University of Aegean
Τίτλος:	Στατιστική μοντελοποίηση του αριθμού των αιτήσεων αποζημιώσεων (oustanding claim counts) στην αναλογιστική επιστήμη.
Title:	Statistical modelling of outstanding claim counts.
Επιβλέπων Καθηγητής:	Κατσής Αθανάσιος, Επίκουρος Καθηγητής
Supervisor	Katsis Athanasios, Assistant Professor
Ακαδημαϊκή Επιτροπή:	Νικολέρης Θεόδωρος, Επίκουρος Καθηγητής Νάκας Χρήστος, Λέκτορας
Academic Comitee:	Nikoleris Theodoros, Assistant Professor Nakas Christos, Lecturer
Keywords:	Generalized Poisson, Hypothesis Tests, Lagrangian Poisson, Markov Chain Monte Carlo, MCMC, Reversible Jump Markov Chain Monte Carlo, RJMCMC, Negative Binomial, Poisson.

Abstract

One of the principal reasons of the increasing popularity of the Bayesian methods is the combination of constant development in the field of statistical simulation and the widely available computational power. Namely, Markov Chain Monte Carlo (MCMC) methods seem to prevail in this particular field. Reversible Jump Markov Chain Monte Carlo (RJMCMC) algorithm is a recent extension which incorporates the model selection problem.

The purpose of this thesis is the comparison of discrete candidate models used for claim count modeling in the actuarial field, from a Bayesian perspective. A general framework of model comparison is proposed using advanced computational techniques to estimate the posterior model odds amongst different distributions for claim counts. Finally, RJMCMC algorithms are constructed for three candidate models and applied to datasets of automobile accidents.

Περίληψη

Ένας από τους λόγους της εξάπλωσης των Μπεϋζιανών μεθόδων είναι ο συνδυασμός της ανάπτυξης στατιστικών μεθόδων προσομοίωσης με την ευρύτατη διάθεση της αυξανόμενης υπολογιστικής δύναμης. Η μέθοδος Markov Chain Monte Carlo (MCMC) είναι από τις πλέον γνωστές στον τομέα της στατιστικής προσομοίωσης. Μια πρόσφατη επέκταση της μεθόδου MCMC είναι ο αλγόριθμος Reversible Jump Markov Chain Monte Carlo (RJMCMC) που συμπεριλαμβάνει και το πρόβλημα της επιλογής ενός μοντέλου ανάμεσα από ένα πλήθος ανταγωνιστικών.

Σκοπός της συγκεκριμένης εργασίας είναι η σύγκριση και ο έλεγχος Μπεϋζιανών μοντέλων διακριτών κατανομών που χρησιμοποιούνται στην Αναλογιστική Επιστήμη για την μοντελοποίηση του αριθμού αιτήσεων αποζημίωσης. Προτείνεται ένα γενικό πλαίσιο σύγκρισης χρησιμοποιώντας υπολογιστικές τεχνικές για να εκτιμήσουμε την posterior πιθανότητα του κάθε μοντέλου. Εφαρμόζουμε τον αλγόριθμο RJMCMC για τρία ανταγωνιστικά μοντέλα και τον υλοποιούμε πάνω σε δεδομένα αυτοκινητιστικών ατυχημάτων.

Περιεχόμενα

Κατάλογος Πινάκων

Κατάλογος Σχημάτων

Κεφάλαιο 1

Εισαγωγή

1.1 Στατιστική και Αναλογιστική Επιστήμη

Στις 21 Δεκεμβρίου 1763 ο Richard Price παρουσίασε μια εργασία του αποθανόντος Thomas Bayes στην Royal Society της Βρετανίας. Η εργασία έχει τίτλο «Toward Solving a Problem in the Doctrine of Change» και θεωρείται ο θεμελιώδης λίθος της Μπεϋζιανής Στατιστικής που για αυτό το λόγο φέρει τιμητικά το όνομα του συγγραφέα της. Περιέχει επίσης κάποιες συνεισφορές του ίδιου του Price ενώ δεν έχει ξεκαθαριστεί αν πρόσθεσε μόνο τα παραρτήματα ή έχει συνεισφέρει και σε άλλα τμήματα της εργασίας. Ο Richard Price ήταν ένας σημαντικός λόγιος με προσφορά σε πολλούς τομείς όπως η φιλοσοφία, η στατιστική, τα δημόσια οικονομικά, η δημογραφία και η αναλογιστική επιστήμη. Συγκεκριμένα στον τελευταίο τομέα, κατασκεύασε τον πίνακα επιβίωσης του Northampton και έγραψε ένα απο τα σημαντικότερα έργα στην αναλογιστική βιβλιογραφία: «Observations on Reversionary Payments: On Schemes for Providing Annuities for Widows and for Persons in Old Age; On the Method of Calculating the Values of Assurance on Lives; and On the National Dept» (Price, 1771). Ο Price γνώριζε τον Bayes και μάλιστα εκτός απο την επιστημονική συνεργασία σε μια εργασία ο Bayes, που πέθανε το 1761, του κληροδότησε 100 λίρες και όλες τις επιστημονικές του εργασίες (Makov, 2001, Anders 1998). Με την σειρά του ο Price επέλεξε την εργασία που αναφέραμε και την παρουσίασε. Με αυτό τον, σχεδόν μυθιστορηματικό, τρόπο η γέννηση της Μπεϋζιανής στατιστικής συνδέεται με την αναλογιστική επιστήμη μέσω του Price που ήταν παρών στη δημιουργία της και η κινητήρια δύναμη που οδήγησε στην εξάπλωσή της.

Η Μπεϋζιανή ανάλυση εφαρμόστηκε στην αναλογιστική επιστήμη πολύ αργότερα από τη γέ-

νεση της. Οι πρώτες εργασίες εμφανίστηκαν στα τέλη της δεκαετίας του 1960 (Bühlmann 1967). Αρχικά υπήρξε έντονη κριτική για την υποκειμενική φύση της και για το αν μπορεί να εφαρμοστεί εύκολα μια και τα προβλήματα που προκύπτουν δεν έχουν «έτοιμες» αναλυτικές λύσεις. Το πρώτο ζήτημα έχει φιλοσοφικές προεκτάσεις είναι προτιμότερο να αφεθεί στην προσωπική κρίση του ερευνητή. Αντιθέτως για το δεύτερο (έλλειψη αναλυτικών λύσεων) έχουν προταθεί εξειδικευμένες τεχνικές αριθμητικής επίλυσης - στατιστικής προσομοίωσης (π.χ. Markov Chain Monte Carlo αλγόριθμοι) με πολύ ικανοποιητικά αποτελέσματα που σε συνδυασμό με την αλματώδη αύξηση της υπολογιστικής ισχύς την καθιστούν ιδιαιτέρως αποτελεσματική. Η Μπεϋζιανή μεθοδολογία χρησιμοποιείται σε διάφορους τομείς της αναλογιστικής επιστήμης αλλά θα αναφέρουμε αυτούς που έχουν συχνότερη εφαρμογή.

Experience rating: είναι η περιοχή που ασχολείται με τον υπολογισμό των ασφαλίστρων με βάση τον αριθμό των αιτήσεων προηγούμενων ετών. Θεωρούμε $\theta_{ij} (i = 1, \ldots, I \; j = 1, \ldots, J)$ την παράμετρο που εκφράζει το ολικό ρίσκο (risk parameter) του συμβολαίου i στην χρονική στιγμή j. Δεδομένου του θ_{ij}, οι πραγματικές αιτήσεις ενός συμβολαίου-i $X_{i1}, X_{i2}, \ldots$ είναι στοχαστικά ανεξάρτητες και ακολουθούν μια κατανομή $f(x_{ij}|\theta_{ij})$. Τα θ είναι ανεξάρτητα ομοιόμορφα κατανεμημένα και ακολουθούν μια prior κατανομή $U(\cdot)$ που ονομάζεται δομική κατανομή (structure distribution). Το ασφάλιστρο που θέλουμε να υπολογίσουμε είναι το:

$$E[X_{i,J+1}|y] = E[\mu(\theta_{ij})|y],$$

όπου y είναι τα δεδομένα και $\mu = E[X_{ij}|\theta_{ij}]$ το δίκαιο ασφάλιστρο (fair premium). Η αρχική μέθοδος υπολογισμού του fair premium ήταν με empirical Bayes τεχνικές, στους Makov et al. (1996), Herzog (1994). Πρόσφατα έχουν χρησιμοποιηθεί και τεχνικές MCMC από τους Makov et al. (1996), και Scollnik (1996).

Experience reserving και Compound claim modeling: είναι η περιοχή που ασχολείται με τον υπολογισμό των αποθεματικών (loss reserves) που πρέπει να κρατά μια ασφαλιστική εταιρεία, με την αποτίμηση των συνολικών αποζημιώσεων καθώς και με την επίπτωση τους στα οικονομικά της εταιρείας αντίστοιχα. Τα αποθεματικά χρειάζονται σε πολλές περιπτώσεις, είτε όταν οι απώλειες παραμένουν απλήρωτες στο τέλος του χρόνου, είτε ως ζημιές που δεν έχουν αναφερθεί (Incurred But Not Reported - IBNR), είτε ως ζημιές που έχουν αναφερθεί και δεν έχουν ακόμη διευθετηθεί Reported But Not Settled (RBNS). Οι αιτήσεις αποζημίωσης αναπαριστώνται ως runoff triangles δηλαδή τριγωνικοί πίνακες που

Χρονιά σε ισχύ	Χρονιά σε εξέλιξη (μετά την ισχύ)						
	1	2	...	t	...	$k-1$	k
1	X_{11}	X_{12}	...	X_{1t}	...	$X_{1,k-1}$	$X_{1,k}$
2	X_{21}	X_{22}	...	X_{2t}		$X_{2,k-1}$	–
3	X_{31}	X_{32}	...	X_{3t}		–	–
$\vdots$						–	–
$k-1$	$X_{k-1,1}$	$X_{k-1,2}$				–	–
k	X_{k1}	–				–	–

Πίνακας 1.1: Παράδειγμα ενός Runoff Triangle

αποτελούνται από τυχαίες μεταβλητές X_{ij} $(i = 1, \ldots, I; j = 1, \ldots, I - i + 1)$ των συχνοτήτων ή των λόγων απωλειών του συμβολαίου της $i-$στης χρονιάς στον j-στο χρόνο μετά την ισχύ του (Πίνακας 1.1). Ο Verral (1990) ασχολήθηκε με το πρόβλημα και πρότεινε μια empirical Bayes μέθοδο που μπορεί να γραφτεί και ως two-way ANOVA μοντέλο:

$$\log(X_{ij}) = \mu + \alpha_I + \beta_J + \epsilon_{ij}$$

Οι Ntzoufras and Dellaportas (2002) υλοποίησαν μια πλήρη Μπεϋζιανή μέθοδο όπου συγκρίνουν (τέσσερα) διαφορετικά μοντέλα που περιλαμβάνουν και την αβεβαιότητα ως προς τον αριθμό των αιτήσεων.

Graduation: είναι η περιοχή που ασχολείται με την κατασκευή πινάκων θνησιμότητας. Σε μια ομιλία του ο Whittaker (1920) πάνω στην ερμηνεία της πιθανότητας, τα ερωτήματα που έθεσε προκάλεσαν γόνιμη συζήτηση με καρποφόρα συμπεράσματα. Στο μοντέλο Graduation που είχε προτείνει, χρειαζόταν να ελαχιστοποιηθεί η συνάρτηση απώλειας:

$$L = F + hS,$$

όπου F είναι ένα μέτρο έλλειψης προσαρμογής, S ένα μέτρο έλλειψης ομαλότητας (smoothness), h μια θετική σταθερά. Σύμφωνα με τους Hickman και Jones (σχολιαστές στον Makov, 2001) η ερμηνεία που δόθηκε από τον διορατικό Whittaker ενθάρρυνε την μελέτη του προβλήματος ακολουθώντας την Μπεϋζιανή προσέγγιση. Η πρώτη εργασία που παρουσίασε μια πλήρη Μπεϋζιανή μελέτη είναι του Carlin (1992) που θεωρείτε ότι περιέχει μία από τις πρώτες εφαρμογές της μεθόδου MCMC στην αναλογιστική επιστήμη.

Θα επικεντρώσουμε το ενδιαφέρον μας στον τομέα του Experience reserving που ενσωματώνει και τον αριθμό των αιτήσεων αποζημιώσεων (outstanding claim counts) για αυτοκινητιστικά ατυχήματα.

1.2 Το ερευνητικό πρόβλημα της μοντελοποίησης του αριθμού των αιτήσεων

Η επιτυχής μοντελοποίηση τυχαίων γεγονότων όπως η βροχόπτωση, η ζήτηση ηλεκτρικής ενέργειας ή η ζήτηση ενός αγαθού είναι στο επίκεντρο του ερευνητικού ενδιαφέροντος. Η ορθή επιλογή στατιστικών κατανομών είναι ύψιστης σημασίας γιατί οδηγεί σε ικανοποιητική αναπαράσταση των παραμέτρων που χαρακτηρίζουν τα γεγονότα. *Ο αριθμός των αιτήσεων αποζημιώσεων* για αυτοκινητιστικά ατυχήματα είναι ένα τυχαίο γεγονός που περιλαμβάνεται στο αντικείμενο μελέτης της αναλογιστικής επιστήμης. Χαρακτηρίζεται και από την οπτική της συχνότητας, δηλαδή πόσο συχνά συμβαίνουν τα ατυχήματα, αλλά και από την οπτική της σφοδρότητας δηλαδή πόσο μεγάλη οικονομική ζημιά τυχαίνει. Συχνά ο χρόνος που μεσολαβεί ανάμεσα στην έλευση της ζημιάς και την αναφορά της στην ασφαλιστική εταιρία είναι μεγάλος, στην περίπτωση αυτή τα ατυχήματα λέμε ότι έγιναν αλλά δεν έχουν αναφερθεί και ονομάζονται «Incurred But Not Reported (IBNR)». Με τον ίδιο όρο χαρακτηρίζονται (λόγω σύμβασης) και οι αιτήσεις που είναι γνωστές στην εταιρία αλλά δεν έχουν αποζημιωθεί πλήρως. Ένας εναλλακτικός τρόπος αντιμετώπισης του ίδιου φαινομένου αφορά την μελέτη *του ποσού των αποζημιώσεων*, κάτι που δεν θα μας απασχολήσει στην παρούσα εργασία καθώς θα ασχοληθούμε μόνο για ατυχήματα που έχουν αναφερθεί, αποκλείοντας επίσης τις περιπτώσεις IBNR.

Η επιλογή της κατάλληλης κατανομής για να οδηγηθούμε σε ένα ικανοποιητικό μοντέλο που θα εκφράζει τον αριθμό αιτήσεων αποζημιώσεων παραμένει φλέγον ζήτημα τόσο στην θεωρητική όσο και στην εφαρμοσμένη έρευνα (Makov, 2001). Στη σχετική βιβλιογραφία έχουν διερευνηθεί διάφορα μοντέλα: ο Ter Berg (1980) προτείνει δύο μοντέλα, ένα για τον αριθμό των αιτήσεων αποζημιώσεων (Loglinear Poisson) και ένα για τα ποσά των αποζημιώσεων (Gamma). Ο Ruohonen (1988) χρησιμοποιείται ένα μοντέλο βασισμένο στην κατανομή Delaporte ενώ στους Ter Berg (1996) και Scollnik (1998) ερευνάται η Γενικευμένη κατανομή Poisson (Generalized ή Lagrangian Poisson distribution). Ειδικά στην τελευταία χρησιμοποιείται Μπεϋζιανή ανάλυση και

η μέθοδος Markov Chain Monte Carlo (MCMC).

Εστιάζουμε το ενδιαφέρον μας στις εργασίες που μελετούν το πρόβλημα των αιτήσεων αποζημίωσης με την βοήθεια της στατιστικής κατά Bayes. Ο de Alba (2002) παρουσιάζει μια Μπεϋζιανή μέθοδος για την πρόβλεψη των αιτήσεων αποζημίωσης είτε ως μοντέλο (Poisson) του αριθμού αποζημιώσεων, είτε ως μοντέλο (Log-Normal) του ποσού (αποθεματικά) των αποζημιώσεων, οι prior κατανομές στα δύο μοντέλα εκφράζουν την άγνοια (μη-πληροφοριακές) και εφαρμόζονται μέθοδοι Monte Carlo για τον υπολογισμό των posterior κατανομών σε πολύ γνωστά δεδομένα της σχετικής βιβλιογραφίας. Ο de Alba παραπάνω καθώς και οι Verrall (1990) , Ntzoufras, Dellaportas (2002) υποστηρίζουν την Μπεϋζιανή προσέγγιση στο πρόβλημα του αριθμού των αιτήσεων αποζημιώσεων, όμως, χρησιμοποιούν αποκλειστικά την κατανομή Poisson για να εκτιμήσουν την posterior κατανομή των άγνωστων παραμέτρων. Θα ήταν χρήσιμο να συμπεριληφθεί στην μελέτη και η *αβεβαιότητα* ως προς την κατανομή που αναπαριστά τον αριθμό των αιτήσεων αποζημιώσεων είτε ελέγχοντας υποθέσεις για διάφορες κατανομές είτε εκτιμώντας την αβεβαιότητα του μοντέλου.

Η συμπερασματολογία κατά Bayes έπεται από την κατασκευή ενός μοντέλου m, τον υπολογισμό της πιθανοφάνειας του $f(y|\boldsymbol{\theta_m}, m)$ και την επιλογή της prior κατανομής $f(\boldsymbol{\theta_m}|m)$, όπου με $\boldsymbol{\theta_m}$ συμβολίζουμε το διάνυσμα των παραμέτρων του μοντέλου m, με y τα δεδομένα που στην περίπτωσή μας αναπαριστώνται με ένα διάνυσμα. Τα τελικά συμπεράσματα για το μοντέλο εξάγονται από την posterior κατανομή $f(\boldsymbol{\theta_m}|y, m)$ ενώ η αβεβαιότητα του μοντέλου ποσοτικοποιείται μέσω της *posterior πιθανότητας του μοντέλου* $f(m|y)$.

Για παράδειγμα, έστω ότι έχουμε δύο «ανταγωνιστικά» μοντέλα m_0 και m_1. Θέλουμε να εκμεταλλευτούμε την όποια γνωστή πληροφορία (π.χ. δεδομένα) ώστε να συγκρίνουμε το ένα με το άλλο. Αν $f(m)$ είναι η prior πιθανότητα του μοντέλου m τότε χρησιμοποιώντας το θεώρημα του Bayes η posterior πιθανότητα PO_{01} του μοντέλου m_0 ως προς το μοντέλο m_1 δίνεται από τον τύπο:

$$PO_{01} = \frac{f(m_0|y)}{f(m_1|y)} = \frac{f(y|m_0)}{f(y|m_1)} \frac{f(m_0)}{f(m_1)} = B_{01} \frac{f(m_0)}{f(m_1)} \tag{1.1}$$

Όπου B_{01} ονομάζεται Bayes Factor και το κλάσμα $\frac{f(m_0)}{f(m_1)}$ είναι η prior πιθανότητα του μοντέλου m_0 ως προς το m_1.Διαισθητικά, μπορούμε να πούμε ότι ο Bayes Factor εκφράζει την πιθανότητα του posterior μοντέλου προς το prior. Μια πιο ακριβής ερμηνεία (Lavine και Schervish, 1999 καθώς και στο βιβλίο των Carlin και Louis, 2001) είναι ότι εκφράζει την *αλλαγή* στην πιθανότητα

ως προς το μοντέλο 1 καθώς μετακινούμαστε από την prior στη posterior πληροφόριση. Οι Kass και Raftery (1994) βασιζόμενοι στην πρωτότυπη εργασία του Jeffreys για τους Bayes Factors προτείνουν ένα πίνακα για διευκόλυνση στη διεξαγωγή συμπερασμάτων (Πίνακας 1.2).

$\log_{10}(B_{10})$	B_{10}	Ένδειξη κατά της H_0
$0 - \frac{1}{2}$	$1 - 3.2$	Αξίζει να γίνει μια απλή αναφορά
$\frac{1}{2} - 1$	$3.2 - 10$	Είναι υπαρκτή
$1 - 2$	$10 - 100$	Δυνατή
> 2	> 100	Αποφασιστική

Πίνακας 1.2: Ερμηνεία των Bayes Factors έναντι της μηδενικής υπόθεσης.

Οπότε, μεγάλες τιμές του B_{01}, συνήθως μεγαλύτερες του 12, υποστηρίζουν το μοντέλο m_0 έναντι του m_1 χρησιμοποιώντας την posterior πληροφορία. Η $f(y|m)$ ονομάζεται *περιθώρια πιθανοφάνεια* του $f(y|m) = \int f(y|\boldsymbol{\theta_m}, m) f(\boldsymbol{\theta_m}|m) d\boldsymbol{\theta_m}$. Ο Bayes Factor B_{01} του μοντέλου m_1 ως προς το μοντέλο m_0 εκτιμά τόσο αν υπάρχει ένδειξη *ενάντια* στην μηδενική υπόθεση, όπως γίνεται και στους κλασσικούς ελέγχους σημαντικότητας όσο και αν υπάρχει ένδειξη *υπέρ* της μηδενικής υπόθεσης, που αδυνατούν να κάνουν οι κλασσικοί έλεγχοι (Kass και Raftery, 1994). Στην περίπτωση που θεωρούμε ένα σύνολο από ανταγωνιστικά μοντέλα $\mathcal{M} = \{m_1, m_2, \ldots, m_{|\mathcal{M}|}\}$ εστιάζουμε την προσοχή μας στην posterior πιθανότητα του μοντέλου $m \in \mathcal{M}$ που ορίζεται ως:

$$f(m|y) = \frac{f(y|m)\ f(m)}{\sum\limits_{m_l \in \mathcal{M}} f(y|m_l)\ f(m_l)} = \left(\sum_{m_l \in \mathcal{M}} PO_{m_l,m} \right)^{-1}, \tag{1.2}$$

όπου $\mathcal{M}$ είναι το σύνολο των μοντέλων και $|\mathcal{M}|$ ο αριθμός των μοντέλων που εξετάζονται.

Με τον Bayes Factor έχουμε ένα τρόπο να αξιολογήσουμε τα ανταγωνιστικά μοντέλα. Για την *επιλογή ενός* από αυτά εφαρμόζονται αρχές από την Θεωρία Ωφελιμότητας ή Χρησιμότητας. Ορίζεται μία συνάρτηση χρησιμότητας και το μοντέλο που την μεγιστοποιεί επιλέγεται έναντι των άλλων (Bernado and Smith 1994, Chipman et al. 2002). Εναλλακτικά χρησιμοποιείται ο όρος συνάρτηση απώλειας ως μια συνάρτηση *απώλειας πληροφορίας* που επιλέγουμε να ελαχιστοποιήσουμε. Η απόφαση που παίρνουμε και στις δύο περιπτώσεις είναι η ίδια γιατί πρόκειται για ένα δυικό πρόβλημα. Συνήθως επιλέγονται συναρτήσεις ωφελιμότητας $0 - 1$, για την επιλογή των εσφαλμένων και ορθών μοντέλων αντίστοιχα.

Στο ζήτημα της επιλογής ενός μοντέλου κάνουμε την παραδοχή ότι το πλέον κατάλληλο μοντέλο είναι ανάμεσα στα ανταγωνιστικά που έχουμε συμπεριλάβει, χωρίς να έχουμε εκφράσει

κάποια «αβεβαιότητα» στην διαδικασία επιλογής του μοντέλου. Στην περίπτωση αυτή είμαστε απολύτως σίγουροι για την καταλληλότητα του μοντέλου. Μία μέθοδος που έχει προταθεί για να βασίζουμε τα συμπεράσματα σε πολλά μοντέλα είναι η Bayesian Model Averaging (BMA). Συγκεκριμένα, η συμπερασματολογία βασίζεται σε ένα σταθμισμένο μέσο πάνω στο χώρο των μοντέλων (Hoeting, 2002 , Hoeting et al., 1999 , Carlin and Louis, 2000). Αν Δ είναι η ποσότητα που μας ενδιαφέρει όπως η επίδραση μιας ενέργειας, η επιτυχής πρόβλεψη ή η χρησιμότητα ενός τρόπου ενεργειών τότε η posterior κατανομή δοθέντος των δεδομένων y είναι:

$$f(\Delta|y) = \sum_{l \in \mathcal{M}} f(\Delta|y, m_l) f(m_l|y), \tag{1.3}$$

όπου $f(m_l|y)$ είναι η posterior πιθανότητα του μοντέλου m_l, $l \in \mathcal{M}$ όπως ορίστηκε στη (1.2) και $f(\Delta|y, m_l)$ η posterior κατανομή για τη Δ στο l-μοντέλο. Αν για παράδειγμα θέλουμε την BMA εκτίμηση για μια παράμετρο θ:

$$\hat{\theta}_{\text{BMA}} = \sum_{l \in \mathcal{M}} \hat{\theta}_l f(m_l|y)$$

όπου $\hat{\theta}_l$ ο posterior μέσος του μοντέλου m_l.

Σε οποιοδήποτε μοντέλο/α καταλήξουμε δημιουργούνται συνήθως υπολογιστικά ζητήματα καθώς είτε η διάσταση του παραμετρικού χώρου που πρέπει να ολοκληρώσουμε είναι πολύ μεγάλη είτε ο αριθμός των διαφορετικών μοντέλων είναι πολύ μεγάλος. Για την επίλυση των προβλημάτων έχουν προταθεί ασυμπτωτικές προσεγγίσεις και εξειδικευμένες υπολογιστικές τεχνικές. Μία πολύ γνωστή τεχνική υπολογισμού με τη βοήθεια στατιστικής προσομοίωσης είναι η Markov Chain Monte Carlo (Gilks 1996, Scollnik 2001) και η πρόσφατη εξέλιξη της από τον Green (1995) για μοντέλα διαφορετικών διαστάσεων με τον αλγόριθμο Reversible Jump Markov Chain Monte Carlo (RJMCMC) για περισσότερες πληροφορίες: στους Carlin και Louis (2000), στους Chen et al., (2000), στους Han και Carlin (2001). Για τον υπολογισμό του BMA έχουν προταθεί αντίστοιχες τεχνικές υπολογισμού: Occam's window των Madigan, Raftery (1994) που βασίζεται στην μείωση του χώρου των μοντέλων και Markov Chain Monte Carlo Model Composition $(\text{MC})^3$ των Madigan, York (1995) που βασίζεται στη μέθοδο MCMC.

Σε αυτή την εργασία εστιάζουμε στην σύγκριση διακριτών κατανομών που χρησιμοποιούνται για τον αριθμό των αιτήσεων αποζημιώσεων και την εκτίμηση των posterior πιθανοτήτων για κάθε μοντέλο ακολουθώντας τη Μπεϋζιανή μεθοδολογία. Ειδικότερα, ακολουθούμε τη Μπεϋζιανή μεθοδολογία των Ntzoufras et al. (2005), που αναφέρεται στην εφαρμογή του RJMCMC αλγόριθμου σε πραγματικά δεδομένα από τον χώρο της Αναλογιστικής επιστήμης. Η Μπεϋζιανή

προσέγγιση μας προσφέρει αρκετά πλεονεκτήματα έναντι της κλασσικής. Αρχικά, μπορούμε να χρησιμοποιήσουμε ένα μείγμα μοντέλων, ανταγωνιστικών μεταξύ τους, εφαρμόζοντας μία «στάθμιση» ανάλογα με τις posterior πιθανότητες του καθενός (Bayesian Model Averaging), για να πάρουμε πιο ακριβή αποτελέσματα, ή με την βοήθεια μια συνάρτησης χρησιμότητας να επιλέξουμε το καλύτερο. Επίσης, δεν υπάρχουν οι περιορισμοί που ισχύουν στην κλασική στατιστική όπου μόνο nested μοντέλα μπορούν να συγκριθούν. Θα επικεντρωθούμε στη μελέτη τριών διαφορετικών μοντέλων αλλά ο τρόπος που έχει προταθεί είναι γενικός και μπορούν να συμπεριληφθούν στην ανάλυση ακόμα περισσότερα.

Κεφάλαιο 2

Κατανομές

Για την μοντελοποίηση των αιτήσεων αποζημιώσεων είδαμε πως έχουν προταθεί πλήθος διαφορετικών κατανομών. Θα μας απασχολήσουν τρεις κατανομές που έχουν μεγάλο εύρος εφαρμογών. Η γνωστή κατανομή Poisson (Ter Berg, 1980), η Αρνητική Διωνυμική ή Negative Binomial (Verral, 2000) και η Γενικευμένη (Generalized ή Lagrangian) Poisson (Ter Berg, 1996, Scollnik, 1998). Μια πλήρης περιγραφή για κάθε κατανομή δίνεται στο βιβλίο των Johnson et al. (1993).

2.1 Poisson

Η κατανομή Poisson μπορεί να αντιμετωπιστεί ως ειδική περίπτωση της Αρνητικής Διωνυμικής ή της Γενικευμένης Poisson. Έστω τα δεδομένα $y_i,\ i = 1, \ldots, n$. Το μοντέλο Poisson δίνεται από:

$$y_i|\lambda \sim Poisson(\lambda)$$

με συνάρτηση πυκνότητας πιθανότητας:

$$f(y_i|\lambda) = \frac{\lambda^{y_i} \exp(-\lambda)}{y_i!},\ \lambda > 0 \tag{2.1}$$

Η κατανομή Poisson έχει την ιδιότητα η μέση τιμή να είναι ίση με την διακύμανση που οδηγεί τον δείκτη της σκέδασης (Dispersion Index ή DI) να παίρνει την τιμή $\mu_2/\mu = 1$. Η ιδιότητα αυτή χρησιμοποιείται σε διάφορους τομείς (Johnson et. al., 1993, σελ. 157) π.χ. για να ανιχνευθεί η ύπαρξη σκέδασης είτε με τη μορφή υπερβολικής σκέδασης (overdispersion όταν $\mu_2 > \mu$), είτε ως έλλειψη σκέδασης (underdisperion όταν $\mu_2 < \mu$). Ωστόσο, η έλλειψη σκέδασης συναντάται

σπάνια σε πραγματικά δεδομένα στα οποία η δειγματική διασπορά συνήθως ξεπερνά την δειγματική μέση τιμή (overdispersion). Στις περιπτώσεις αυτές παίρνουμε τιμές πολύ διαφορετικές από $DI = 1$ με αποτέλεσμα να μην ευσταθεί η υπόθεση του απλού μοντέλου Poisson. Για τον λόγο αυτό έχουν προταθεί εναλλακτικά μοντέλα που εμφανίζουν overdispersion ενώ σχετίζονται με το απλό μοντέλο Poisson.

2.2 Αρνητική Διωνυμική - Negative Binomial

Η Αρνητική Διωνυμική κατανομή μπορεί να κατασκευαστεί προσθέτοντας ένα ακόμη ιεραρχικό επίπεδο στο απλό μοντέλο Poisson, συγκεκριμένα:

$$y_i|\epsilon_i, \lambda \sim Poisson(\epsilon_i\lambda),\ \epsilon_i|\vartheta \sim Gamma(\vartheta, \vartheta)$$

όπου $\vartheta > 0$ και $Gamma(a, b)$ είναι μια κατανομή Γάμμα με μέση τιμή a/b και διασπορά a/b^2. Η συνάρτηση πυκνότητας πιθανότητας δίνεται από:

$$\begin{aligned}
f(y_i|\lambda,\vartheta) &= \int_0^\infty f(y_i|\lambda,\epsilon_i)f(\epsilon_i)d\epsilon_i \\
&= \int_0^\infty \frac{e^{-\lambda\epsilon_i}\lambda\epsilon_i{}^{y_i}}{y_i!}\ \frac{\vartheta^\vartheta\epsilon_i^{\vartheta-1}e^{-\vartheta\epsilon_i}}{\Gamma(\vartheta)}d\epsilon_i \\
&= \frac{\lambda^{y_i}\vartheta^\vartheta}{\Gamma(y_i+1)\Gamma(\vartheta)}\int_0^\infty \epsilon_i^{y_i+\vartheta-1}e^{-\epsilon_i(y_i+\vartheta)}d\epsilon_i \\
&= \frac{1}{\Gamma(y_i+1)\Gamma(\vartheta)}\ \frac{\lambda^{y_i}\vartheta^\vartheta}{(\lambda+\vartheta)^{y_i+\vartheta}}\int_0^\infty z^{y_i+\vartheta-1}e^{-z}dz, \quad \epsilon_i > 0
\end{aligned}$$

άρα έχει την μορφή:

$$f(y_i|\lambda,\vartheta) = \frac{\Gamma(y_i+\vartheta)}{\Gamma(y_i+1)\Gamma(\vartheta)}\left(\frac{\lambda}{\lambda+\vartheta}\right)^{y_i}\left(\frac{\vartheta}{\lambda+\vartheta}\right)^{\vartheta}, \quad \vartheta > 0 \tag{2.2}$$

όπου $\Gamma(x)$ η συνάρτηση Γάμμα, ενώ η κατανομή έχει $E(y_i) = \lambda$ και $Var(y_i) = \lambda + \lambda^2/\vartheta$. Διαισθητικά θα λέγαμε ότι η Αρνητική Διωνυμική εκφράζει μια «ροπή σε ατυχήματα» όπου η κατά μέσο όρο πιθανότητα να συμβεί ένα ατύχημα είναι $\lambda\epsilon_i$ και αυτή με την σειρά της εξαρτάται από μια κατανομή Γάμμα. Το μοντέλο Poisson είναι η οριακή κατανομή της (2.2) για την περίπτωση

$\vartheta \to \infty$. Στη συνέχεια υπολογίζουμε τον δείκτη DI, ενώ χρησιμοποιούμε την παραμετροποίηση $\phi = \lambda/\vartheta$:

$$DI = \frac{Var(y_i)}{E(y_i)} = 1 + \lambda/\vartheta = 1 + \phi$$

Για την περίπτωση $\phi \to 0$ η παραπάνω κατανομή γίνεται η απλή Poisson.

2.3 Γενικευμένη (Generalized) ή Lagrangian Poisson

Οι Consul και Jane (1973) πρότειναν την Γενικευμένη (Generalized) ή Lagrangian Poisson η οποία στη συνέχεια χρησιμοποιήθηκε στην αναλογιστική επιστήμη. Προκύπτει από κατάλληλη μετατροπή της κατανομής *Tanner-Borel*:

$$Pr[Y = y] = \frac{n}{(y-n)!} y^{y-n-1} (\ell\beta)^{y-n} e^{-t\beta\gamma}, \quad y = n, n+1, \ldots$$

όπου εκφράζει τον συνολικό αριθμό των πελατών Y που εξυπηρετούνται πριν η ουρά αναμονής αδειάσει δεδομένου ότι έχουμε μια ουρά αναμονής με τυχαίους χρόνους άφιξης πελατών με σταθερό ρυθμό ℓ και με συγκεκριμένο χρόνο β εξυπηρέτησης. Η μετατροπή αφορά την μετατόπιση της Tanner-Borel δηλαδή μετασχηματίζοντας την τ.μ. $Y = X - n$ ώστε να υποστηρίζει πλέον ενδεχόμενα $\{0, 1, 2, \ldots\}$. Η Γενικευμένη Poisson ανήκει στην Lagrangian οικογένεια κατανομών, δηλαδή οι πιθανότητες μπορούν να προκύψουν επεκτείνοντας (Lagrangian expansion) την γεννήτρια συνάρτηση πιθανότητας. Η γεννήτρια συνάρτηση πιθανότητας είναι της μορφής:

$$G(z) = e^{\vartheta\{g(z)-1\}}$$

δηλαδή η κατανομή είναι μία Poisson stopped-sum, εκφράζει ένα Poisson αριθμό ανεξαρτήτων κατανεμημένων τυχαίων μεταβλητών με γεννήτρια συνάρτηση πιθανότητας $g(z)$.

Το μοντέλο της *Γενικευμένης (Generalized) ή Lagrangian Poisson* με παραμέτρους ζ, ω ορίζεται ως εξής:

$$f(y_i|\zeta, \omega) = \frac{\zeta(\zeta + \omega y_i)^{y_i - 1}}{y_i!} e^{-(\zeta + \omega y_i)}, \quad \zeta > 0, \quad 0 \leq \omega < 1 \tag{2.3}$$

Η μέση τιμή της κατανομής είναι $E(y_i) = \zeta(1-\omega)^{-1}$ ενώ η διακύμανση δίνεται από $Var(y_i) = \zeta(1-\omega)^{-3}$. Σύμφωνα με τον Ter Berg (1996) οι τιμές που μπορεί να πάρει το ω είναι μέσα στο διάστημα $[0, 1)$. Θεωρητικά η κατανομή ορίζεται και για οποιαδήποτε τιμή του $\omega \in \Re$ μόνο που σε κάθε περίπτωση πλην του διαστήματος $[0, 1)$ οι τιμές της αθροιστικής συνάρτησης πιθανότητας

(α. σ. π.) δεν αθροίζουν στην τιμή 1. Στις περιπτώσεις αυτές είναι σύνηθες να χρησιμοποιείται μια «κόλουρη» εκδοχή της ώστε να ικανοποιούνται οι περιορισμοί της α. σ. π. (Scollnik 1998). Πέρα από αυτό τον περιορισμό οι αρνητικές τιμές του ω οδηγούν στην εμφάνιση underdispersion, μια ιδιότητα που δεν είναι τόσο συνήθεις για τον αριθμό των αιτήσεων αποζημιώσεων. Για αυτό τον λόγο θα μείνουμε στην μελέτη της κατανομής για $\omega \in [0, 1)$. Για $\omega = 0$ παίρνουμε την απλή Poisson με μέσο ζ.

Τέλος για να διευκολύνουμε τις συγκρίσεις ανάμεσα στα τρία μοντέλα που προτείναμε, παραμετροποιούμε την παραπάνω κατανομή χρησιμοποιώντας $\lambda = \zeta/(1-\omega)$. Η «καινούργια» κατανομή έχει $E(y_i) = \lambda$, $Var(y_i) = \lambda(1-\omega)^{-2}$, $DI = (1-\omega)^{-2}$. Αντίστοιχα η συνάρτηση πυκνότητας πιθανότητας είναι:

$$f(y_i|\lambda, \omega) = (1-\omega)\lambda \frac{\{(1-\omega)\lambda + \omega y_i\}^{y_i - 1}}{y_i!} e^{-\{(1-\omega)\lambda + \omega y_i\}}. \tag{2.4}$$

Η παραμετροποίηση μας ωφελεί γιατί μπορούμε πλέον να ερμηνεύσουμε την παράμετρο λ με τον ίδιο τρόπο ανάμεσα στα τρία μοντέλα και μας διευκολύνει στην υλοποίηση του αλγόριθμου MCMC που θα αναλύσουμε παρακάτω (Παράρτημα Α′).

2.4 Καθορισμός των a-priori κατανομών

Οι posterior πιθανότητες των μοντέλων είναι πολύ ευαίσθητες στον βαθμό των a priori διασπορών έχοντας μια τάση να ευνοούν τα μοντέλα με πιο απλή δομή καθώς οι prior διασπορές αυξάνουν (Bartlett, 1957, Lindley, 1957, Sinharay και Stern, 2002). Επομένως η επιλογή των prior είναι καίριας σημασίας για την a posteriori υποστήριξη των μοντέλων. Στην δημοσίευσή του ο Lindley (1957) δίνει έμφαση στην επίδραση του μεγέθους του δείγματος στις posterior πιθανότητες των μοντέλων και στα αντιφατικά αποτελέσματα μεταξύ των Μπεϋζιανών και κλασσικών ελέγχων σημαντικότητας. Ο Bartlett (1957) αντίστοιχα τονίζει την επίδραση της prior κατανομής στις posterior πιθανότητες των μοντέλων. Το πρόβλημα γίνεται πιο εμφανές όταν συγκρίνονται (nested) μοντέλα που το ένα περιλαμβάνεται στο άλλο. Συγκεκριμένα το Poisson μοντέλο περιλαμβάνεται και στο μοντέλο της Αρνητικής Διωνυμικής και στο μοντέλο της Γενικευμένης Poisson.

Με βάση τα παραπάνω ορίζουμε τις prior κατανομές και καταλήγουμε στην εξής ιεραρχική

δομή των παραμέτρων των μοντέλων:

$$\begin{aligned} f(\lambda, m_1) &= f(\lambda|m_1)f(m_1) \\ f(\lambda, \vartheta, m_2) &= f(\vartheta|\lambda, m_2)f(\lambda|m_2)f(m_2) \\ f(\lambda, \omega, m_3) &= f(\omega|\lambda, m_3)f(\lambda|m_3)f(m_3). \end{aligned}$$

Επιλέγουμε για prior κατανομή του δείκτη του μοντέλου m την *ομοιόμορφη κατανομή* πάνω στο χώρο των μοντέλων $\mathcal{M}$, καταλήγοντας στην $f(m_i) = \frac{1}{3}, \quad i = 1, 2, 3$. Με την συγκεκριμένη prior κατανομή δίνουμε την ίδια προτίμηση και στα τρία μοντέλα, θέτοντας την πιθανότητα του κάθε μοντέλου ίση. Φυσικά η επιλογή της prior κατανομής είναι υποκειμενική και σε καμιά περίπτωση δεν μπορεί να θεωρηθεί τέλεια.

Επίσης όσες παράμετροι είναι κοινές ανάμεσα στα μοντέλα που προτείναμε θα πρέπει να τις διέπουν οι ίδιες κατανομές ώστε να είναι συνεπείς με τις αρχικές πεποιθήσεις μας. Στην περίπτωση που εξετάζουμε, η παράμετρος λ είναι κοινή σε όλα τα μοντέλα επομένως χρησιμοποιούμε την ίδια prior κατανομή $f(\lambda|m)$, που είναι μια κατανομή Γάμμα, $Gamma(a, b)$. Εφόσον εκφράζουμε την έλλειψη κάποιας πληροφορίας για την λ θέτουμε τις υπερ-παραμέτρους (hyperparameters) a, b ίσους με κάποιον πολύ μικρό θετικό αριθμό: συγκεκριμένα χρησιμοποιήσαμε $a = b = 0.0001$ που οδηγεί σε prior μέση τιμή $E(\lambda) = 1$ και διασπορά $Var(\lambda) = 10000$. Η επίδραση αυτής της επιλογής στις posterior πιθανότητες των μοντέλων είναι ελάχιστη διότι η παράμετρος λ είναι παρούσα σε όλα τα μοντέλα που εξετάζουμε (Kass and Raftery, σελ. 24, 1994).

Τέλος μας μένει να ορίσουμε τις prior κατανομές $f(\vartheta|\lambda, m_2)$ και $f(\omega|\lambda, m_3)$ ώστε οι κατανομές που επηρεάζουν τον DI να είναι ουσιαστικά η ίδια και στα δυο μοντέλα. Για αυτό τον λόγο καθορίζουμε την μια κατανομή και υπολογίζουμε την άλλη εξισώνοντας τους δείκτες σκέδασης DI των δυο κατανομών. Με αυτό τον τρόπο καταφέρνουμε να έχουν την ίδια κατανομή των DI και στα δύο μοντέλα. Εξισώνοντας στο μοντέλο της Αρνητικής Διωνυμικής και στο μοντέλο της Γενικευμένης Poisson έχουμε:

$$\phi = \lambda/\vartheta = \frac{\omega(2-\omega)}{(1-\omega)^2} \quad \text{ή} \quad \omega = 1 - \frac{1}{\sqrt{1+\phi}}. \tag{2.5}$$

Όπως είδαμε στην §2.3 η παράμετρος ω ορίζεται στο διάστημα $[0, 1)$ και μια συνήθης κατανομή που μπορεί να εκφράσει την αβεβαιότητα μας είναι η ομοιόμορφη δίνοντας ίση πιθανότητα σε οποιοδήποτε διάστημα ίσου μήκους. Οπότε η αντίστοιχη prior κατανομή για το ϕ είναι μια Βήτα τύπου II (Beta type II) με παραμέτρους 1, 1/2 και συνάρτηση πυκνότητας πιθανότητας:

$$f_\phi(\phi|\lambda, m_2) = \frac{1}{2}(1+\phi)^{-\frac{3}{2}},$$

Επειδή $\vartheta = \lambda/\phi$ η prior για το ϑ δίνεται από:

$$f_\vartheta(\vartheta|\lambda, m_1) = f_\phi(\frac{\lambda}{\vartheta}|\lambda, m_2)\lambda\vartheta^{-2} = \frac{1}{2}\lambda\vartheta^{-2}(1+\frac{\lambda}{\vartheta})^{-\frac{3}{2}}, \tag{2.6}$$

Η οποία είναι μια Βήτα τύπου II (Beta type II) υπό κλίμακα.

Κεφάλαιο 3

Ο αλγόριθμος RJMCMC για τις αιτήσεις αποζημιώσεων

Σε αυτό το κεφάλαιο θα επικεντρωθούμε στον υπολογισμό των posterior πιθανοτήτων των μοντέλων χρησιμοποιώντας την μεθοδολογία που προτάθηκε από τον Green (1995) Reversible Jumb Markov Chain Monte Carlo (RJMCMC). Όπως έχουμε ήδη διαπιστώσει, εκτός από την περίπτωση των συζυγών κατανομών, ο υπολογισμός των posterior κατανομών είναι ιδιαίτερα δύσκολος λόγων των ολοκληρωμάτων που εμπλέκονται μέσω του θεωρήματος του Bayes. Για αυτό τον λόγο έχουμε καταφύγει σε μια σειρά από: αναλυτικές προσεγγίσεις, αριθμητικές μεθόδους επίλυσης, Monte Carlo εκτιμήσεις, εκτιμήσεις που βασίζονται στα αποτελέσματα προσομοιώσεων MCMC για κάθε μοντέλο και τέλος εκτιμήσεις MCMC για μοντέλα διαφορετικών διαστάσεων. Επιλέξαμε τον αλγόριθμο RJMCMC για τη μεγάλη προσαρμοστικότητα που έχει, διότι μπορεί να χειριστεί πολλά και διαφορετικά ανταγωνιστικά μοντέλα μέσα σε μια MCMC αλυσίδα. Με αυτό τον τρόπο καταφέρνουμε να υπολογίσουμε τις άγνωστες παραμέτρους των μοντέλων μας ενώ ταυτόχρονα υπολογίζουμε τις posterior πιθανότητες που θα μας βοηθήσουν στην επιλογή ενός από των μοντέλων ή στους υπολογισμούς του BMA.

3.1 Ο γενικός αλγόριθμος RJMCMC

Η Reversible Jump μεθοδολογία ή μεθοδολογία «αναστρέψιμου άλματος» προτάθηκε, όπως είδαμε, ως μια στατιστική μέθοδος τον Green (1995). Επεκτείνει τις τεχνικές MCMC καθώς η δειγματοληψία βασίζεται και στο χώρο των παραμέτρων και στο χώρο των μοντέλων που τα

τελευταία μπορούν να διαφέρουν ως προς τις διαστάσεις τους. Παράγει μια Μαρκοβιανή αλυσίδα (Markov Chain) που μπορεί να «μεταπηδά» μεταξύ μοντέλων διαφορετικών διαστάσεων ενώ διατηρεί τις συνθήκες της απεριοδικότητας, της μη-αναγωγής (irreducibility) και της λεπτομερής ισορροπίας (detailed balance) ώστε να εξασφαλίζεται η σωστή οριακή κατανομή (περαιτέρω πληροφορίες μπορείτε να βρείτε στους Carlin και Louis, 2001, Chen et al., 2000, Han και Carlin (2001).

Ας υποθέσουμε ότι έχουμε ένα σύνολο ανταγωνιστικών μοντέλων $\mathcal{M}$. Μια βοηθητική μεταβλητή $m \in \mathcal{M}$ επισημαίνει το κάθε μοντέλο και $\boldsymbol{\theta_m}$ είναι το αντίστοιχο διάνυσμα παραμέτρων. Ο αλγόριθμος λειτουργεί πάνω στο χώρο της ένωσης, $\mathcal{M} \times \bigcup_{m\in\mathcal{M}} \boldsymbol{\theta_m}$, που θα συζητήσουμε παρακάτω. Αν η τρέχουσα κατάσταση της αλυσίδας Markov είναι $(m, \boldsymbol{\theta_m})$, όπου $\boldsymbol{\theta_m}$ έχει διάσταση d_m, γενική διατύπωση του αλγόριθμου είναι ως εξής:

- Προτείνουμε ένα καινούργιο μοντέλο m' με πιθανότητα $j(m, m')$.
- Παράγουμε το $\boldsymbol{u}$ από μια πυκνότητα προσφοράς (proposal density) $q(\boldsymbol{u}|\boldsymbol{\theta_m}, m, m')$.
- Προτείνουμε ένα καινούργιο διάνυσμα παραμέτρων $\boldsymbol{\theta'_{m'}}$ θέτοντας $(\boldsymbol{\theta'_{m'}}, \boldsymbol{u'}) = g_{m,m'}(\boldsymbol{\theta_m}, \boldsymbol{u})$ όπου $g_{m,m'}$ είναι μια καθορισμένη αντιστρέψιμη συνάρτηση.
- Για να επιτύχουμε την σωστή οριακή κατανομή, δεχόμαστε την προτεινόμενη μετακίνηση στο μοντέλο m' με πιθανότητα:

$$\alpha = \min\left(1, \frac{f(\boldsymbol{y}|m', \boldsymbol{\theta}'_{m'}) f(\boldsymbol{\theta}'_{m'}|m') f(m') j(m', m) q(\boldsymbol{u'}|\boldsymbol{\theta}_m, m', m)}{f(\boldsymbol{y}|m, \boldsymbol{\theta}_m) f(\boldsymbol{\theta}_m|m) f(m) j(m, m') q(\boldsymbol{u}|\boldsymbol{\theta}_{m'}, m, m')} \left|\frac{\partial g(\boldsymbol{\theta}_m, \boldsymbol{u})}{\partial(\boldsymbol{\theta}_m, \boldsymbol{u})}\right|\right). \tag{3.1}$$

Σημαντικά χαρακτηριστικά για την αποδοτικότητα και για την υλοποίηση του αλγόριθμου είναι οι πυκνότητες προσφοράς (proposal densities) $q(\boldsymbol{u}|\boldsymbol{\theta_m}, m, m')$ και η συνάρτηση αντιστοίχησης $g_{m,m'}$. Η $g_{m,m'}$ αντιστοιχεί τον χώρο των παραμέτρων ενός μοντέλου στο χώρο παραμέτρων ενός άλλου χρησιμοποιώντας τα διανύσματα $\boldsymbol{u}$ και $\boldsymbol{u'}$ ώστε $d_{m'} + d_{\boldsymbol{u'}} = d_m + d_{\boldsymbol{u}}$. Η συνηθισμένη πρακτική που ακολουθείται είναι να θέτουμε είτε το $d_{\boldsymbol{u'}}$ είτε το $d_{\boldsymbol{u}}$ ίσο με μηδέν ανάλογα με ποιο μοντέλο έχει τις λιγότερες παραμέτρους. Όταν έχουμε $d_m < d_{m'}$, θέτουμε $d_{\boldsymbol{u'}} = 0$, παράγουμε το $\boldsymbol{u}$ από το $q(\boldsymbol{u}|\boldsymbol{\theta}_m, m, m')$ και υπολογίζουμε το $\boldsymbol{\theta'_{m'}}$ χρησιμοποιώντας την συνάρτηση αντιστοίχησης $g_{m,m'}$. Διαφορετικά όταν $d_{m'} < d_m$, θέτουμε $\boldsymbol{u'} = 0$ και κατευθείαν υπολογίζουμε το $\boldsymbol{\theta_{m'}}$ χρησιμοποιώντας την συνάρτηση αντιστοίχησης $g_{m,m'}$ χωρίς να χρειαστεί να παράγουμε συμπληρωματικές παραμέτρους.

Οι αντίστοιχες κατανομές προσφοράς (proposal distributions) κατασκευάζονται από ξεχωριστές προσομοιώσεις MCMC για κάθε μοντέλο (Dellaportas et al., 2002). Τελείως διαφορετικά κατασκευάζουμε την συνάρτηση αντιστοίχησης $g_{m,m'}$ λαμβάνοντας υπόψην την δομή του κάθε μοντέλου και τις πιθανές συσχετίσεις τους. Επίσεις ισχύει λόγω συμμετρίας $g_{m,m'} = g^{-1}_{m,m'}$.

Είναι χρήσιμο να ξεκαθαρίσουμε λίγο την έννοια της «αντιστοίχισης» ακολουθώντας το παράδειγμα των Carlin και Louis (2001). Ας υποθέσουμε ότι συγκρίνουμε δύο μοντέλα, το $m = 1$ έχει ένα διάνυσμα παραμέτρων την $\theta_1 \in \Re$ και το άλλο $m = 2$ έχει αντίστοιχα ένα διάνυσμα παραμέτρων $\theta_2 \in \Re^2$. Αν θ_1 είναι ένα διάνυσμα που περιλαμβάνεται στο θ_2, τότε όταν μετακινούμαστε από $m = 1$ σε $m' = 2$ παίρνουμε το $\boldsymbol{u} \sim q(\boldsymbol{u}|\theta_1, m, m')$ και θέτουμε:

$$\theta'_{2'} = (\theta_1, \boldsymbol{u})$$

οπότε η Ιακωβιανή ορίζουσα στο 3.1 είναι ίση με ένα καθώς θέλουμε $(\boldsymbol{\theta'_{2'}}) = g_{m,m'}(\boldsymbol{\theta_1}, \boldsymbol{u})$ που ισχύει όταν $g_{m,m'}$ είναι η ταυτοτική συνάρτηση.

Σε πολλές περιπτώσεις δεν μπορούμε να θεωρήσουμε το $\boldsymbol{\theta_1}$ ότι περιλαμβάνεται στο $\boldsymbol{\theta_2}$. Ο Green(1995) αναλύει την περίπτωση ενός change-point model δηλαδή ένα μοντέλο που μπορεί να αλλάζει τις παραμέτρους. Ως change-point model μπορεί να θεωρηθεί ένα μοντέλο παλλινδρόμησης στο διάστημα $[0, L]$. Στην περίπτωση αυτή, θα χρησιμοποιούμε μια συνάρτηση βήματος με $k \in \mathcal{K} = \{0, 1, 2, \ldots\}$ βήματα. Επομένως παράγουμε $k + 1$ διαστήματα όπου στο καθένα το μοντέλο μας έχει σταθερές παραμέτρους αλλά διαφορετικές από κάθε άλλο διάστημα. Ακολουθεί ένα άλλο παράδειγμα που διαφοροποιείται ως προς την συνάρτηση ταύτισης.

Έστω ότι έχουμε να συγκρίνουμε δύο μοντέλα χρονολογικών σειρών το μοντέλο 1 έχει σταθερό μέσο επίπεδο $\boldsymbol{\theta_1}$ και το μοντέλο 2 έχει δύο επίπεδα $\boldsymbol{\theta_{2,1}}$, πριν από ένα σημείο αλλαγής - παρέμβασης (change point), $\boldsymbol{\theta_{2,2}}$ μετά. Αν θέλουμε να μεταπηδήσουμε από το μοντέλο 2 στο μοντέλο 1 δεν θα μπορούσαμε να θεωρήσουμε το $\boldsymbol{\theta_{2,1}}$ ή το $\boldsymbol{\theta_{2,2}}$ υποδιάνυσμα του $\boldsymbol{\theta'_1}$. Πιο φυσικό είναι να θεωρήσουμε την μετάβαση:

$$\boldsymbol{\theta'_1} = \frac{\boldsymbol{\theta_{2,1}} + \boldsymbol{\theta_{2,2}}}{2}$$

μια και η μέση τιμή των δύο επιπέδων δίνουν μια ανταγωνιστική τιμή για να μεταπηδήσουμε στο μοντέλο 1. Τέλος για να εξασφαλίσουμε την αντιστροφή αυτής της κίνησης, για να μεταπηδήσουμε δηλαδή από το μοντέλο 1 στο μοντέλο 2 μπορούμε να πάρουμε ένα $\boldsymbol{u} \sim q(\boldsymbol{u}|\boldsymbol{\theta_1})$ τέτοιο ώστε να θέσουμε:

$$\boldsymbol{\theta'_{1,2}} = \boldsymbol{\theta_1} - \boldsymbol{u} \quad \text{και} \quad \boldsymbol{\theta'_{2,2}} = \boldsymbol{\theta_1} + \boldsymbol{u}$$

και λαμβάνουμε μια $1 - 1$ αντιστρέψιμη συνάρτηση ως $g_{m,m'}$.

3.2 Εφαρμογή στις αιτήσεις αποζημίωσης

Στόχος μας είναι να κατασκευάσουμε ένα αλγόριθμο RJMCMC για τις αιτήσεις αποζημίωσης, χρησιμοποιώντας τα τρία ανταγωνιστικά μοντέλα που παρουσιάσαμε στο κεφάλαιο 2. Η μεταβλητή που υποδεικνύει το κάθε μοντέλο m παίρνει τιμές $m \in \{m_1, m_2, m_3\}$ όπου m_1 το Poisson μοντέλο, m_2 το μοντέλο της Αρνητικής Διωνυμικής και m_3 το μοντέλο της Γενικευμένης Poisson. Επιπλέον, οι παράμετροι για το μοντέλο Poisson συμβολίζονται με $\boldsymbol{\theta_{m1}} = \lambda$, για το μοντέλο Αρνητικής Διωνυμικής $\boldsymbol{\theta_{m2}} = (\lambda, \vartheta)^T$ και για το μοντέλο της Γενικευμένης Poisson $\boldsymbol{\theta_{m3}} = (\lambda, \omega)^T$.

Θεωρούμε ότι η Μαρκοβιανή αλυσίδα είναι στην κατάσταση $(m, \boldsymbol{\theta_m})$, ο αλγόριθμος RJMCMC για την σύγκριση των μοντέλων που μας ενδιαφέρουν μπορεί να διατυπωθεί ως εξής:

1. Παράγουμε τις παραμέτρους $\boldsymbol{\theta_m}$ του μοντέλου από την posterior κατανομή $f(\boldsymbol{\theta_m}|\boldsymbol{y}, m)$ με την βοήθεια μίας απλής MCMC εξομοίωσης (λεπτομέρειες στο παράρτημα Α′).

2. Προτείνουμε ένα άλμα από το m στο m', $m \neq m'$ με πιθανότητα $j(m, m') = (|\mathcal{M}| - 1)^{-1}$.

3. (α′) Αντιστοιχούμε τις παραμέτρους του παλιού με του νέου μοντέλου μελετώντας όλες τις περιπτώσεις:

   i. Εάν $m = m_1$ (Poisson) και $m' = m_2$ (Αρνητική Διωνυμική) τότε παράγουμε μια προτεινόμενη τιμή για το ϑ από την κατανομή προσφοράς $q_\vartheta(\vartheta|m)$.

   ii. Εάν $m = m_1$ (Poisson) και $m' = m_3$ (Γενικευμένη Poisson) τότε παράγουμε μια προτεινόμενη τιμή για το ω από την κατανομή προσφοράς $q_\omega(\omega|m)$.

   iii. Εάν $m = m_2$ (Αρνητική Διωνυμική) και $m' = m_3$ (Γενικευμένη Poisson) τότε παράγουμε μια προτεινόμενη τιμή για το ω από την συνάρτηση αντιστοίχησης:

   $$\omega = h_{m_2,m_3}(\vartheta) = 1 - \left(1 + \lambda\vartheta^{-1}\right)^{-1/2} \tag{3.2}$$

   έπεται από την (2.5), εξισώνοντας τους δείκτες σκέδασης DI των δύο κατανομών.

   iv. Εάν $m = m_2$ (Αρνητική Διωνυμική) ή $m = m_3$ (Γενικευμένη Poisson) και $m' = m_1$ τότε δεν χρειάζεται να παράγουμε επιπλέον παραμέτρους.

   v. Αν $m = m_3$ (Γενικευμένη Poisson) και $m' = m_2$ (Αρνητική Διωνυμική) τότε θέτουμε:

   $$\vartheta = h^{-1}_{m_2,m_3}(\omega) = h_{m_3,m_2}(\omega) = \frac{\lambda(1-\omega)^2}{\omega(2-\omega)}. \tag{3.3}$$

(β′) Δεχόμαστε το προτεινόμενο άλμα με πιθανότητα $\alpha(m,m') = min\{1, \delta(m,m')\}$, όπου:

$$\begin{aligned}
\delta(m_1,m_2) &= \frac{f(\boldsymbol{y}|\lambda,\vartheta,m_2)f(\lambda,\vartheta|m_2)f(m_2)}{f(\boldsymbol{y}|\lambda,m_1)f(\lambda|m_1)f(m_1)q_\vartheta(\vartheta|m_1)} \\
\delta(m_1,m_3) &= \frac{f(\boldsymbol{y}|\lambda,\omega,m_3)f(\lambda,\omega|m_3)f(m_3)}{f(\boldsymbol{y}|\lambda,m_1)f(\lambda|m_1)f(m_1)q_\omega(\omega|m_1)} \\
\delta(m_2,m_3) &= \frac{f(\boldsymbol{y}|\lambda,\omega,m_3)f(\lambda,\omega|m_3)f(m_3)}{f(\boldsymbol{y}|\lambda,\vartheta,m_2)f(\lambda,\vartheta|m_2)f(m_2)} \times \frac{1}{2}\left(1+\lambda\vartheta^{-1}\right)^{-3/2}\lambda\vartheta^{-2}.
\end{aligned}$$

Για τις αντίστροφες μεταβάσεις χρησιμοποιούμε την ιδιότητα: $\delta(m,m') = 1/\delta(m',m)$.

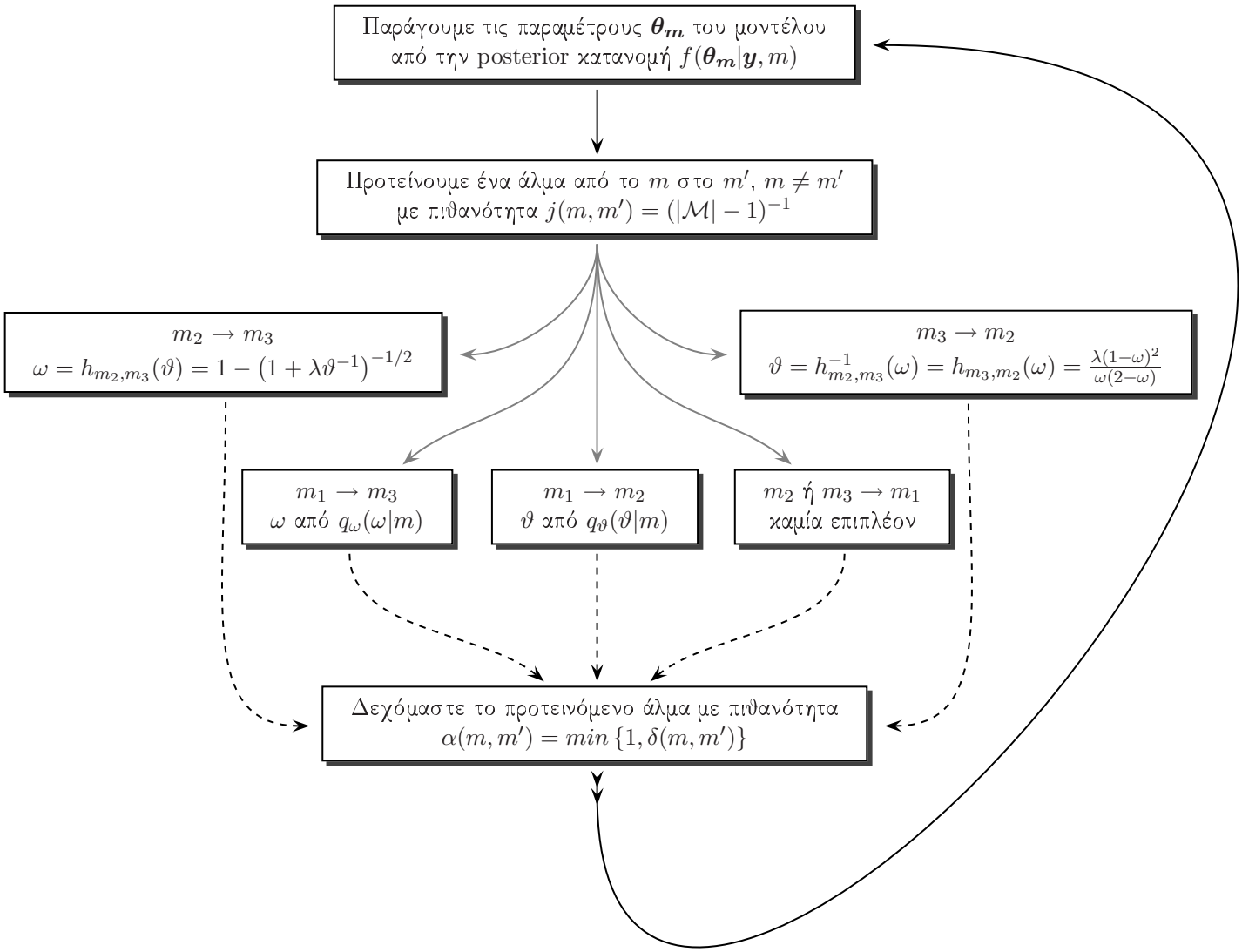
Για μια πιο διαισθητική εικόνα ανατρέξτε στο σχήμα (3.1) όπου αναπαριστάται ο παραπάνω αλγόριθμος. Η ροή των βημάτων είναι από πάνω προς τα κάτω, τα βήματα που έχουν πολλές δυνατές επιλογές στην πραγματικότητα περιορίζονται ανάλογα με τα μοντέλα που έχουν επιλεγεί στη δεδομένη χρονική στιγμή (επανάληψη).

Για το μοντέλο της απλής κατανομής Poisson έχουμε χρησιμοποιήσει *συζυγή prior* κατανομή και γνωρίζουμε ότι η posterior $f(\lambda|\boldsymbol{y},m_1)$ είναι μια κατανομή Γάμμα με παραμέτρους $Gamma(\sum_{i=1}^n y_i+\alpha, n+b)$. Συνεπώς στο βήμα 1, όταν έχουμε $m=m_1$ παράγουμε την παράμετρο λ κατευθείαν από την posterior κατανομή.

Στον παραπάνω αλγόριθμο, η σύγκριση μεταξύ των μοντέλων m_2 και m_3 μπορεί να γίνει χρησιμοποιώντας τον αλγόριθμο Metropolized Carlin Chib (ή έναν independence sampler). Ο πρώτος περιγράφεται από τους Dellaportas et al. (2002) και είναι μια παραλλαγή του αλγόριθμου των Carlin και Chib (1995). Σε αυτή την περίπτωση οι παράμετροι ϑ και ω λαμβάνονται από τις κατανομές προσφοράς $q_\vartheta(\vartheta|m)$ στο βήμα 3-(α′)-i και $q_\omega(\omega|m)$ στο 3-(α′)-ii ενώ συγκρίνονται υπολογίζοντας το $\delta(m_2,m_3) = \delta(m_1,m_3)/\delta(m_1,m_2)$.

Σε περίπτωση μας, έχουμε υποθέσει μια σχέση μεταξύ των DI (οι δείκτες σκέδασης) και αυτό μας επιτρέπει να επιλέξουμε μια πιο αυτοματοποιημένη παραλλαγή του RJMCMC για την σύγκριση των μοντέλων m_2, m_3 ακολουθώντας την προσέγγιση των Ntzoufras et al. (2003). Με αυτό τον τρόπο, αποφεύγουμε να παράγουμε τιμές και για τις δύο παραμέτρους ϑ, ω διότι μπορούμε να παράγουμε παραμέτρους για το ένα μοντέλο και οι όποιες επιπλέον παράμετροι παράγονται εξισώνοντας τον DI. Οδηγούμαστε έτσι σε $1-1$ μετασχηματισμό όπως είδαμε στο (2.5) με την ιδιότητα να κρατάμε το DI σταθερό όταν προτείνεται μια μετάβαση από το μοντέλο της Αρνητικής Διωνυμικής στη Γενικευμένη Poisson και αντίστροφα.

Τέλος χρησιμοποιώντας τις prior κατανομές που παρουσιάσαμε στο Κεφάλαιο 2 ο παραπάνω λόγος απλοποιείται σε μια σύγκριση των πιθανοφανειών (σταθμισμένες από τις prior πιθανότητες



Σχήμα 3.1: Ο αλγόριθμος RJMCMC

του κάθε μοντέλου) σε κάθε επανάληψη του αλγόριθμου:

$$\delta(m_2, m_3) = \frac{f(\boldsymbol{y}|\lambda, \omega, m_3)f(m_3)}{f(\boldsymbol{y}|\lambda, \vartheta, m_2)f(m_2)}.$$

3.3 Καθορισμός των κατανομών προσφοράς

Η σωστή και προσεκτική επιλογή των κατανομών προσφοράς $q_\vartheta(\vartheta|m)$ και $q_\omega(\omega|m)$ καθορίζει την αποτελεσματικότητα του αλγόριθμου RJMCMC. Οι κατανομές προσφοράς παράγουν τιμές για τις παραμέτρους που «λείπουν» από ένα μοντέλο όταν επιχειρείται μια μετάβαση σε ένα άλλο. Για αυτό το λόγο, ένας αποτελεσματικός RJMCMC αλγόριθμος πρέπει να δίνει τιμές κοντά στην posterior κατανομή του μοντέλου m'. Αν αυτό δεν συμβαίνει τότε οι προτεινόμενες τιμές συνεχώς θα απορρίπτονται, αντίστοιχα ο αλγόριθμος είτε θα επικεντρωθεί σε ένα μοντέλο είτε θα συγκλίνει προς την σωστή posterior κατανομή πολύ αργά.

Στη δική μας περίπτωση χρησιμοποιούμε σε κάθε μοντέλο δοκιμαστικές MCMC προσομοιώσεις των 1000 επαναλήψεων. Οι τιμές που προκύπτουν μας εξυπηρετούν ως προσεγγιστικές εκτιμήτριες για τις posterior κατανομές του κάθε μοντέλου. Επίσης, μας βοηθούν ώστε οι προτεινόμενες τιμές και η αντίστοιχη posterior κατανομή να μη διαφέρουν πολύ. Για την παράμετρο ϑ της Αρνητικής Διωνυμικής (που παίρνει θετικές τιμές) χρησιμοποιούμε μια Log-Normal κατανομή $q_\vartheta(\vartheta|m) = LN(\overline{\log\vartheta}, \bar{\sigma}^2_{\log\vartheta})$, όπου $\overline{\log\vartheta}$ είναι η μέση τιμή και $\bar{\sigma}^2_{\log\vartheta}$ η διασπορά των $\log\vartheta$ που λάβαμε από τη δοκιμαστική προσομοίωση, για αναλυτική περιγραφή και περισσότερες λεπτομέρειες ανατρέξτε στο παράρτημα Α′.

Φυσικά, μπορούμε να χρησιμοποιήσουμε οποιαδήποτε κατανομή που ορίζεται στο διάστημα $(0, \infty)$ θέτοντας τις παραμέτρους ώστε να ταυτίζονται με την posterior μέση τιμή και διακύμανση από της δοκιμαστικής προσομοίωσης. Η *αποτελεσματικότητα κάθε κατανομής* προσφοράς εξαρτάται από το πόσο κοντά είναι από την posterior κατανομή των δοκιμαστικών προσομοιώσεων. Είναι αναγκαίο να τονίσουμε ότι οι posterior πιθανότητες του κάθε μοντέλου πρέπει να επιβεβαιώνονται από διαφορετικές κατανομές προσφοράς ώστε να επιτυγχάνουμε ένα κάλο «μείγμα» των μοντέλων. Από την άλλη μεριά, *η αποδοτικότητα του αλγόριθμου* εξαρτάται από τις διαφορετικές επιλογές των κατανομών προσφοράς, συνεπώς, και ο αριθμός των επαναλήψεων που χρειάζεται για να συγκλίνει εξαρτάται από την επιλογή της κατανομής. Οι Dellaportas et al. (2002) και οι Brooks et al. (2003) μελετούν την επιλογή των κατανομών προσφοράς και προτείνουν μία μέθοδο αυτόματου καθορισμού των κατανομών.

Παρόμοια για την παράμετρο ω της Γενικευμένης Poisson, που παίρνει τιμές στο διάστημα $[0, 1]$, χρησιμοποιούμε $q_\omega(\omega|m) = Beta(\bar{a}, \bar{b})$. Οι παράμετροι $\bar{a}$ και $\bar{b}$ υπολογίζονται εξισώνοντας την μέση τιμή και την διακύμανση της κατανομής Βήτα με την δειγματική μέση τιμή ($\bar{\omega}$) και

δειγματική διακύμανση $\bar{\sigma}^2_\omega$, αντίστοιχα, από το δείγμα της δοκιμαστικής MCMC προσομοίωσης για το μοντέλο της Γενικευμένης Poisson, δηλαδή:

$$\bar{\omega} = \frac{\bar{a}}{\bar{a}+\bar{b}}, \quad \bar{\sigma}^2_\omega = \frac{\bar{a}\bar{b}}{(\bar{a}+\bar{b})^2(\bar{a}+\bar{b}+1)}$$

και καταλήγουμε:

$$\bar{a} = \bar{\omega}\left(\frac{\bar{\omega}(1-\bar{\omega})}{\bar{\sigma}^2_\omega} - 1\right), \quad \bar{b} = \bar{a}\,\frac{1-\bar{\omega}}{\bar{\omega}}. \tag{3.4}$$

Η αλυσίδα MCMC είναι δυνατόν να συγκλίνει γρηγορότερα αν αυξήσουμε ή μειώσουμε, ανάλογα με την περίπτωση, την διακύμανση της κατανομής προσφοράς ώστε να έχουμε υψηλούς ρυθμούς αποδοχής.

3.4 Ανάλυση του δείγματος RJMCMC

Με το πέρας των L επαναλήψεων του αλγόριθμου RJMCMC λαμβάνουμε ένα δείγμα με τιμές των παραμέτρων $m^{(k)}, \lambda^{(k)}, \vartheta^{(k)}, \omega^{(k)}$ για κάθε επανάληψη $k = 1, \ldots, L$. Τις πρώτες B επαναλήψεις δεν τις λαμβάνουμε υπόψη μας ως μια περίοδο burn-in που μας βοηθά να απαλείψουμε την όποια πιθανή επίδραση από τις αρχικές τιμές. Η μεταβλητή $m^{(k)} \in \{1, 2, 3\}$ είναι ένας δείκτης του μοντέλου που ο αλγόριθμος επισκέπτεται στην k-στη επανάληψη. Ο αλγόριθμος δρα ως εξής για κάθε μοντέλο που επισκέπτεται:

- Όταν επισκέπτεται το m_1 (Poisson) τότε έχουμε $m = 1$ και $\vartheta = \omega = 0$.
- Όταν επισκέπτεται το m_2 (Αρνητικής Διωνυμικής) τότε έχουμε $m = 2$ και $\omega = 0$.
- Όταν επισκέπτεται το m_3 (Γενικευμένης Poisson) τότε έχουμε $m = 3$ και $\vartheta = 0$.

Από το RJMCMC δείγμα εκτιμούμε την posterior πιθανότητα του μοντέλου $f(m_i|\boldsymbol{y})$ για $i = 1, 2, 3$ από τον τύπο:

$$\hat{f}(m_i|\boldsymbol{y}) = \frac{1}{L-B}\sum_{k=B+1}^{L} I_{m_i}(m^{(k)}) \tag{3.5}$$

όπου η δείκτρια συνάρτηση $I_{m_i}(m^{(k)}) = 1$ αν $i = m^{(k)}$ και μηδέν διαφορετικά. Η παραπάνω εκτιμήτρια μας βοηθά να υπολογίσουμε τους posterior λόγους των μοντέλων (posterior model

odds) και τους Bayes Factors που δίνονται από τους τύπους:

$$\begin{aligned} \widehat{PO}_{ij} &= \frac{\hat{f}(m_i|\boldsymbol{y})}{\hat{f}(m_j|\boldsymbol{y})} = \widehat{BF}_{ij}\frac{\hat{f}(m_i)}{\hat{f}(m_j)} \Leftrightarrow \\ \widehat{BF}_{ij} &= \widehat{PO}_{ij}\frac{\hat{f}(m_j)}{\hat{f}(m_i)}. \end{aligned} \tag{3.6}$$

Κεφάλαιο 4

Εφαρμογή σε δεδομένα

Σε αυτό το κομμάτι της εργασίας θα εφαρμόσουμε τον RJMCMC αλγόριθμο σε δύο σετ δεδομένων που χρησιμοποιούνται και από τον Denuit (1997). Πρόκειται για τον αριθμό των αιτήσεων αποζημίωσης που αφορούν αυτοκινητιστικά ατυχήματα στην Ελβετία το 1961 και το Βέλγιο το 1993. Συνοψίζουμε τα δεδομένα στον πίνακα συχνοτήτων 4.1 και τα αναπαριστούμε στα ιστογράμματα του σχήματος 4.1 όπου δίνονται σε σχέση με τη σχετική συχνότητα. Κάθε κλάση του ιστογράμματος είναι δεξιά ανοιχτή δηλαδή δεν περιέχει τον αριθμό στα δεξιά της και πάνω από κάθε μια δίνεται ο αριθμός των αιτήσεων για την κλάση των ατυχημάτων.

		Αριθμός ατυχημάτων							
	Χώρα	0	1	2	3	4	5	6	7
1.	Ελβετία 1961	103704	14075	1766	255	45	6	2	0
2.	Βέλγιο 1993	57178	5618	446	50	8	0	0	0

Πίνακας 4.1: Πίνακας συχνοτήτων για τις αιτήσεις αποζημίωσης αυτοκινητιστικών ατυχημάτων.

Τρέξαμε τον αλγόριθμο για 21000 επαναλήψεις για κάθε χώρα από τις οποίες αφαιρέσαμε τις πρώτες 1000 ως burn-in περίοδο. Οι αρχικές τιμές υπολογίστηκαν χρησιμοποιώντας εκτιμήσεις με την μέθοδο των ροπών. Οι παράμετροι των κατανομών προσφοράς βασίζονται σε δοκιμαστικές MCMC εξομοιώσεις των 3.000 επαναλήψεων αφού αφαιρέθηκαν επίσης οι πρώτες 1000 ως burn-in περίοδο, με τις διακυμάνσεις τέτοιες ώστε να επιτυγχάνουμε μεγάλο ρυθμό αποδοχής, μεγαλύτερο του 80%, για την περίπτωση της Αρνητικής Διωνυμικής, κοντά στο 20% για την περίπτωση της Γενικευμένης Poisson. Στα σχήματα 4.2 και 4.3 δίνονται οι τιμές που παιρνουν οι παράμετροι των

δύο μοντέλων κατά την διάρκεια του RJMCMC αλγόριθμου για τα δεδομένα της Ελβετίας και του Βελγίου αντίστοιχα. Στην MCMC αλυσίδα κάθε μοντέλου έχουν αφαιρεθεί οι κενές τιμές, που προκείπτουν όταν το μοντέλο δεν έχει επιλεχθεί.

Αφού έχουμε τα αποτελέσματα του αλγόριθμου είμαστε σε θέση να γνωρίζουμε τις posterior μέσες τιμές των παραμέτρων και των posterior πιθανοτήτων των μοντέλων. Αρχικά μπορούμε να υπολογίσουμε αναλυτικά τις posterior τιμές του μοντέλου Poisson λόγω συζυγών κατανομών (Πίνακας 4.2).

		Posterior Τιμές			
		a'	b'	Μέση τιμή	Τυπική Απόκλιση
	Δεδομένα	$\{\sum_{i=1}^{n} y_i + a\}$	$\{n+b\}$	$\{E(\lambda\|\boldsymbol{y}) = a'/b'\}$	$\{S_{\lambda\|\boldsymbol{y}} = \sqrt{a'}/b'\}$
1.	Ελβετία 1961	18594	119853	0.155	0.0011
2.	Βέλγιο 1993	6691	63299	0.106	0.0013

Πίνακας 4.2: Posterior Εκτιμήσεις για το μοντέλο Poisson

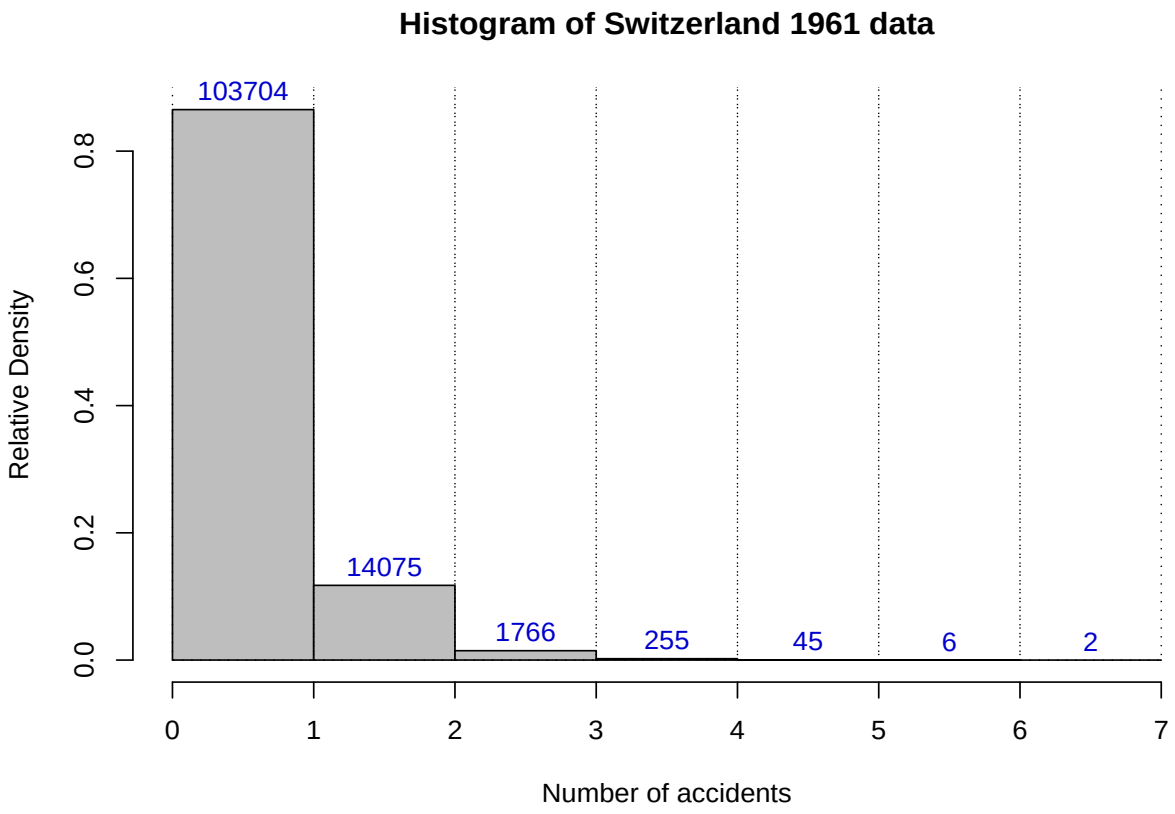
Από τον αλγόριθμο RJMCMC παίρνουμε τις εκτιμήσεις των παραμέτρων για τα μοντέλα της Αρνητικής Διωνυμικής κατανομής (Πίνακας 4.3) και της Γενικευμένης Poisson (Πίνακας 4.4). Παρατηρούμε, και στις δύο χώρες, ότι οι εκτιμήσεις της παραμέτρου λ είναι παραπλήσιες καθώς έχουμε παράγει το δείγμα από την συζυγή posterior κατανομή.

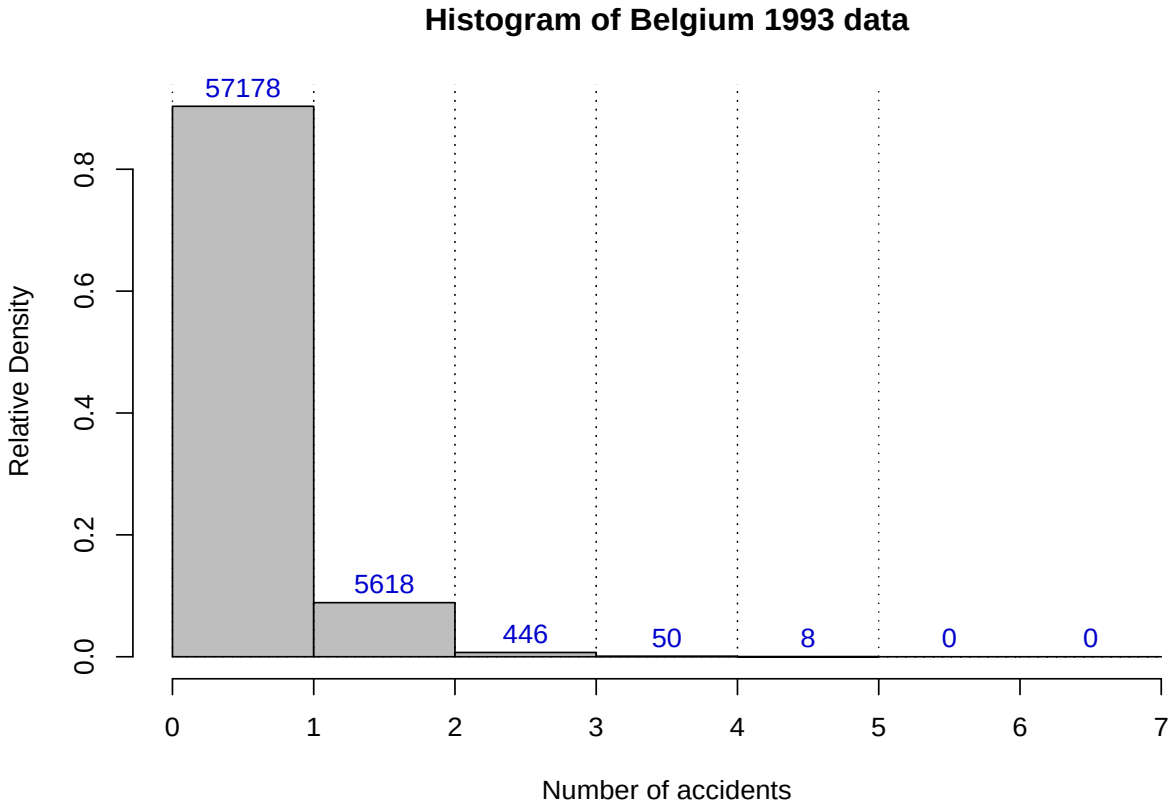
		Posterior Μέση τιμή ± Τυπική Απόκλιση	
		λ	ϑ
1.	Ελβετία 1961	0.15508±0.000796	0.98372±0.00025
2.	Βέλγιο 1993	0.10570±0.000934	1.20907±0.00019

Πίνακας 4.3: Posterior Εκτιμήσεις για το μοντέλο της Αρνητικής Διωνυμικής

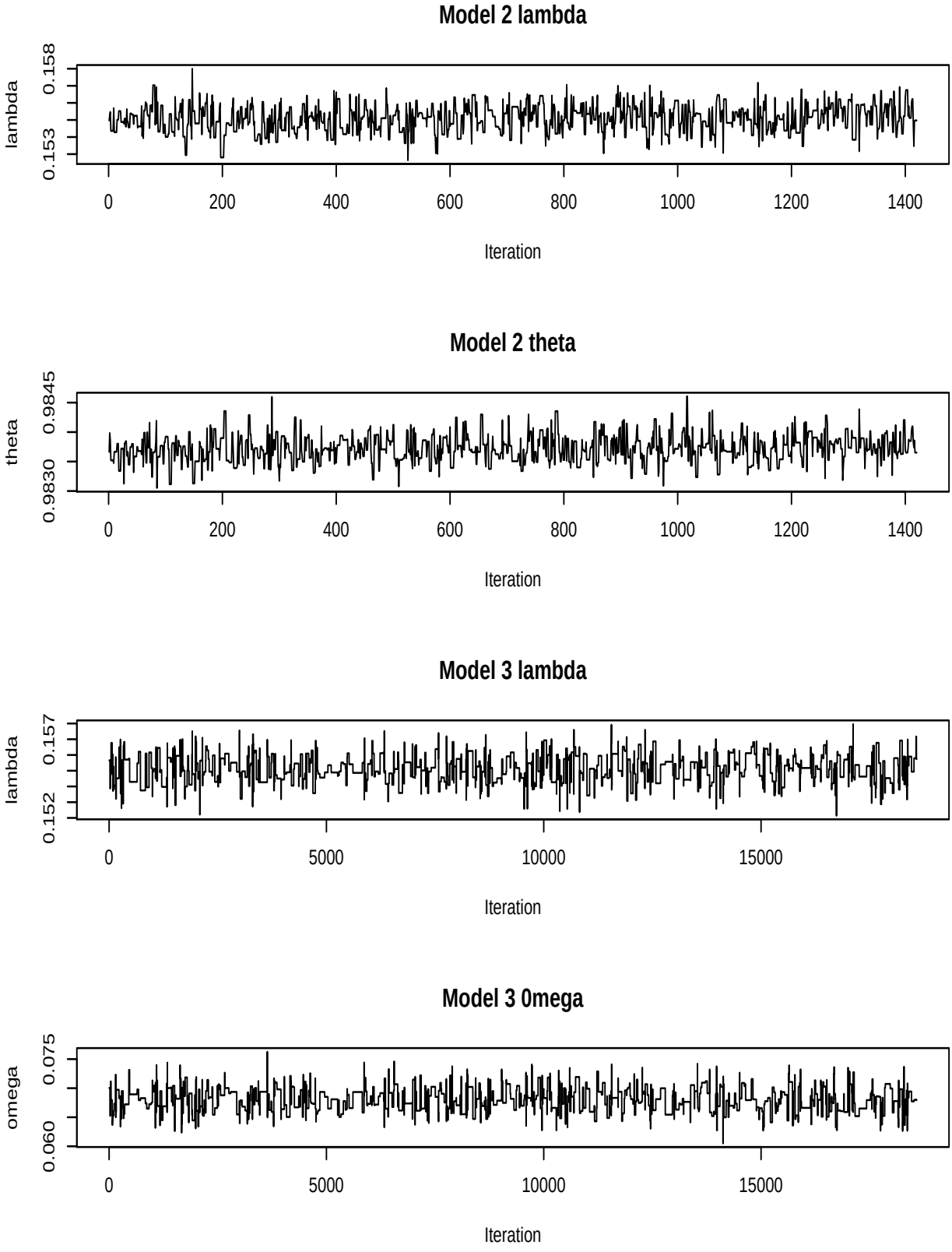
Δίνουμε τις posterior πιθανότητες κάθε μοντέλου (Πίνακας 4.5). Είναι σαφές από τις πιθανότητες των μοντέλων ότι το απλό μοντέλο Poisson δεν υποστηρίζεται καθόλου από τα δεδομένα και ο αλγόριθμος δεν το επισκέπτεται καθόλου.

Για να συγκρίνουμε τα υπόλοιπα μοντέλα υπολογίζουμε τους Bayes Factors (Πίνακας 4.6). Το Poisson μοντέλο έχει μηδενικές posterior πιθανότητες στα δεδομένα των δύο χωρών. Έχουμε επιλέξει για prior πιθανότητα κάθε μοντέλου $f(m_1) = f(m_2) = f(m_3) = 1/3$, ενώ ισχύει από (3.6):

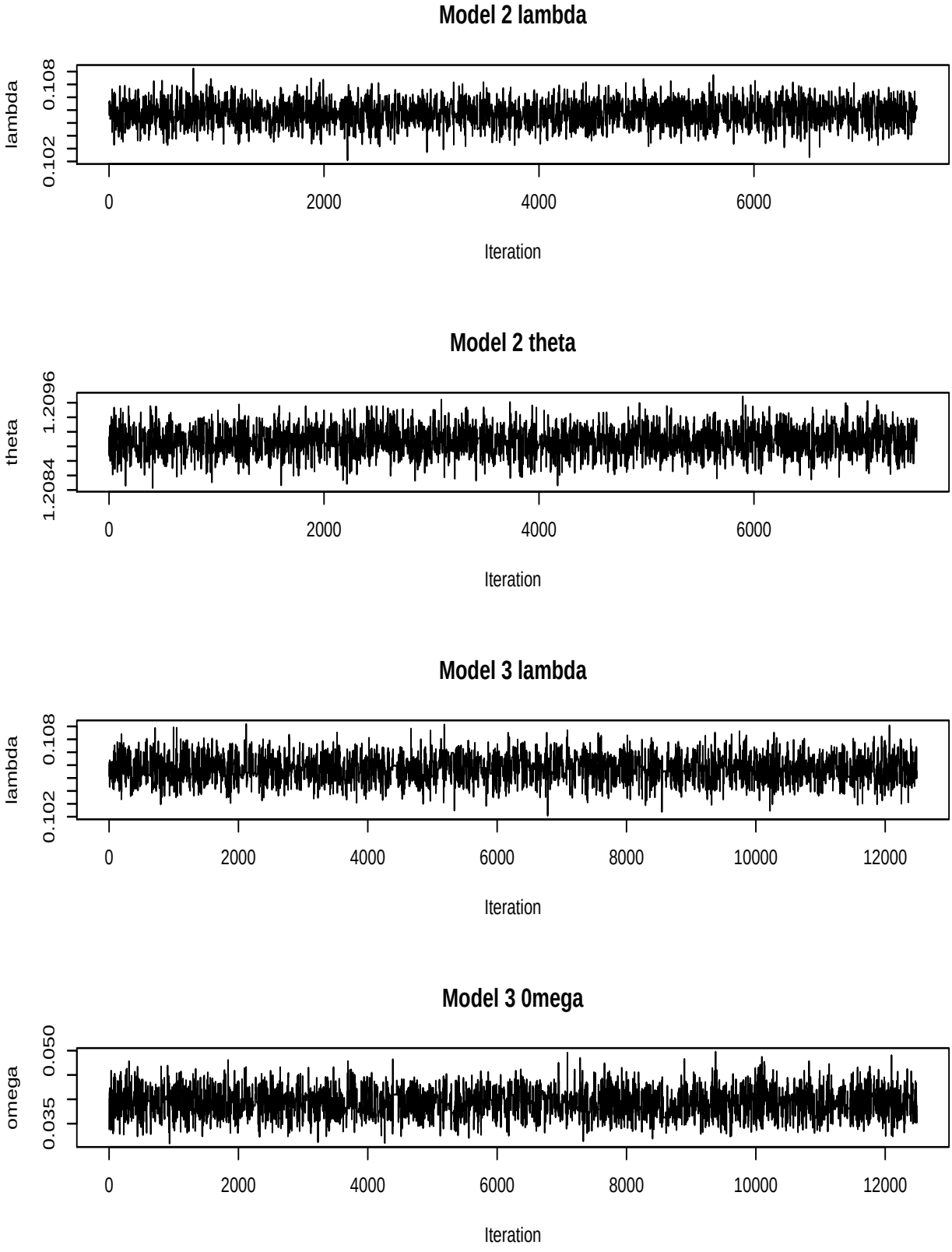




Σχήμα 4.1: Ιστογράμματα των δεδομένων Ελβετίας και Βελγίου



Σχήμα 4.2: RJMCMC Output των παραμέτρων για τα δεδομένα της Ελβετίας



Σχήμα 4.3: RJMCMC Output των παραμέτρων για τα δεδομένα του Βελγίου

		Posterior Μέση τιμή ± Τυπική Απόκλιση	
		λ	ω
1.	Ελβετία 1961	0.15514±0.000854	0.06805±0.001995
2.	Βέλγιο 1993	0.10570±0.000912	0.03925±0.002611

Πίνακας 4.4: Posterior Εκτιμήσεις για το μοντέλο της Γενικευμένης Poisson

	Posterior πιθανότητα του Μοντέλου		
	m_1	m_2	m_3
Ελβετία 1961	0	0.07099645	0.9290035
Βέλγιο 1993	0	0.3756812	0.6243188

Πίνακας 4.5: Posterior πιθανότητες των μοντέλων

	Bayes Factors των μοντέλων		
	m_2 v.s. m_1	m_3 v.s. m_1	m_3 v.s. m_2
Ελβετία 1961	∞	∞	13.08521
Βέλγιο 1993	∞	∞	1.661831

Πίνακας 4.6: Bayes Factors

$$\widehat{BF}_{ij} = \widehat{PO}_{ij}\frac{\hat{f}(m_j)}{\hat{f}(m_i)} = \widehat{PO}_{ij} = \frac{\hat{f}(m_i|\boldsymbol{y})}{\hat{f}(m_j|\boldsymbol{y})} \tag{4.1}$$

Με τη βοήθεια του πίνακα 1.2 θα εξάγουμε τα συμπεράσματά μας. Όπως είδαμε το μοντέλο της απλής κατανομής Poisson δεν υποστηρίζεται καθόλου ανεξαρτήτως των δεδομένων και των μοντέλων. Για τα δεδομένα της Ελβετίας βλέπουμε ότι υπάρχει δυνατή ένδειξη για το μοντέλο της Γενικευμένης Poisson έναντι του μοντέλου της Αρνητικής Διωνυμικής. Στα δεδομένα του Βελγίου υποστηρίζεται επίσης το μοντέλο της Γενικευμένης Poisson αλλά σε πολύ μικρότερο βαθμό για να τη θεωρήσουμε ως ισχυρή ένδειξη για την Γενικευμένη Poisson. Επιπλέον, εντοπίζουμε το μέγεθος των διαφορών δίνοντας στους Πίνακες 4.7 και 4.8 τις συχνότητες των δεδομένων που προκύπτουν από τις κατανομές πρόβλεψης με τις παραμέτρους που υπολογίσαμε.

Οι Πίνακες 4.7,4.8 εκτιμούν τον αριθμό των αιτήσεων που κατατέθηκαν από πελάτες που είχαν Y ατυχήματα. Για παράδειγμα, αν θέλουμε να δούμε στον Πίνακα 4.7 τις εκτιμήσεις των μοντέλων για το πόσοι πελάτες έκαναν αιτήσεις για ένα ατύχημα ($Y = 1$), παρατηρούμε ότι το μοντέλο Poisson εκτιμά 15909.81 πελάτες, το μοντέλο της Αρνητικής Διωνυμικής 13902.50 πελάτες, της Γενικευμένης Poisson 14009.45 πελάτες ενώ ο πραγματικός αριθμός των πελατών είναι 14075.

Πρόβλεψη με τη μέση τιμή των συχνοτήτων

Αριθμός ατυχημάτων: Y

Μοντέλο	$Y=0$	$Y=1$	$Y=2$	$Y=3$	$Y=4$	$Y=5$	$Y=6$
Poisson	102643.9	15909.81	1233.010	63.70553	2.468589	0.076526	0.001976
Neg.Binomial	103778.4	13902.50	1877.835	254.3361	34.49455	4.682177	0.6358892
Gen.Poisson	103718.7	14009.45	1836.825	247.8467	34.46168	4.914465	0.7154753
Παρατηρούμενα	103704	14075	1766	255	45	6	2

Πίνακας 4.7: Περίληψη των κατανομών πρόβλεψης για τα δεδομένα της Ελβετίας

Το μοντέλο Poisson παρουσιάζει μεγάλη απόκλιση σε σχέση με τις τιμές των υπολοίπων μοντέλων και των πραγματικών. Στον Πίνακα 4.8 το μοντέλο παρουσιάζει πολύ μικρές τιμές ($< 10^{-4}$) οι οποίες αναφέρονται ως 00. Και στις δύο συλλογές δεδομένων δίνεται μια προτίμηση του μοντέλου της Γενικευμένης Poisson έναντι της Αρνητικής Διωνυμικής από τους Bayes Factors. Στα δεδομένα της Ελβετίας έχουμε για $Y = 0, 1, 2$ καλύτερη προσαρμογή του m_3 μοντέλου με μεγάλες διαφορές στις συχνότητες ενώ για τους υπόλοιπους αριθμούς ατυχημάτων

οι προβλεπόμενες συχνότητες είναι πολύ κοντά, με το m_3 μοντέλο να έχει ελαφρώς καλύτερη προσαρμογή (εξαίρεση για $Y = 3$). Στα δεδομένα του Βελγίου παρατηρούμε επίσης το μοντέλο της Γενικευμένης Poisson να έχει ελαφρώς καλύτερη προσαρμογή πλην της ουράς ($Y \geq 3$) όπου το μοντέλο m_2 προσεγγίζει κάπως καλύτερα τις παρατηρούμενες συχνότητες.

Πρόβλεψη με τη μέση τιμή των συχνοτήτων

Αριθμός ατυχημάτων: Y

Μοντέλο	$Y = 0$	$Y = 1$	$Y = 2$	$Y = 3$	$Y = 4$
Poisson	63230.41	69.55345	0.0382544	00	00
Neg.Binomial	57199.68	5560.179	493.7555	42.46329	3.59239
Gen.Poisson	57187.4	5583.981	483.3304	41.3744	3.57393
Παρατηρούμενα	57178	5618	446	50	8

Πίνακας 4.8: Περίληψη των κατανομών πρόβλεψης για τα δεδομένα του Βελγίου

Κεφάλαιο 5

Συμπεράσματα - Μελλοντική Έρευνα

Σε αυτή την εργασία χρησιμοποιήσαμε τη Μπεϋζιανό υπόδειγμα και τεχνικές MCMC για να εκτιμήσουμε και να συγκρίνουμε τρεις κατανομές που χρησιμοποιούνται ευρέως στην αναλογιστική επιστήμη για την μοντελοποίηση του αριθμού των αιτήσεων αποζημίωσης. Το Μπεϋζιανό υπόδειγμα μας επιτρέπει να χρησιμοποιήσουμε συναρτήσεις χρησιμότητας για την επιλογή ενός μοντέλου ή τεχνικές στάθμισης για την επιλογή ενός μείγματος μοντέλων καθώς και να συγκρίνουμε μοντέλα τα οποία δεν έχουν απαραίτητα κάποια δομική ομοιότητα (non-nested models). Τα αποτελέσματα από τις συλλογές δεδομένων που χρησιμοποιήσαμε δείχνουν, ξεκάθαρα, ότι η κατανομή Poisson δεν είναι κατάλληλη για να εκφράσει τον αριθμό αιτήσεων αποζημίωσης. Μεταξύ της Αρνητικής Διωνυμικής και της Γενικευμένης Poisson δίνεται ελαφρώς μεγαλύτερη posterior υποστήριξη στην δεύτερη αν και οι διαφορές που υπάρχουν στις προβλεπόμενες τιμές δεν είναι ιδιαίτερα μεγάλες.

Η ομοιότητα που παρατηρούμε μεταξύ των δύο τελευταίων κατανομών μπορεί να εξηγηθεί από τον Douglas (1994) που συμπεραίνει ότι για δεδομένα με μικρό αριθμό αιτήσεων πολλές διακριτές κατανομές μπορούν να εφαρμοστούν ικανοποιητικά. Στην πράξη επιβεβαιώσαμε την επιλογή μας από τους Bayes Factors καθώς διαπιστώσαμε διαφορές οι οποίες θα ήταν μεγαλύτερες για περισσότερα δεδομένα.

Η μεθοδολογία που χρησιμοποιήσαμε μπορεί να επεκταθεί και στον πιο γενικό τομέα των αποζημιώσεων όπου χρησιμοποιείται η Poisson κατανομή για να αποδώσει τον αριθμό των αιτήσεων, για να ελεχθεί πόσο ικανοποιητικά αναπαριστά τα δεδομένα. Ένας άλλος τομέας έρευνας που μπορεί επεκταθεί είναι στην πρόβλεψη των αποθεματικών για αποζημιώσεις (claim reserving). Μπορούν επίσης να προστεθούν και άλλα μοντέλα στον αλγόριθμο RJMCMC. Με αυτό τον

τρόπο συγκρίνονται περισσότερα μοντέλα αλλά και μπορούν να χρησιμοποιηθούν σε ένα σταθμισμένο μείγμα μοντέλων (Bayesian Model Averaging) το οποίο θα οδηγήσει σε ακριβέστερες προβλέψεις.

Τέλος ως προς την υλοποίηση του αλγόριθμου χρησιμοποιήθηκε το στατιστικό πακέτο R, συμβατό με την γλώσσα προγραμματισμού S3, ενώ ακολουθήθηκε μια αντικειμενοστραφή (object-oriented) προσέγγιση. Φυσικά στον κώδικα μπορούν να εισαχθούν πλήθος βελτιστοποιήσεων, ειδικά αν χρησιμοποιήσουμε την δυνατότητα εισαγωγής precompiled βιβλιοθηκών που θα χειρίζονταν την μνήμη του υπολογιστή ώστε να έχουμε ταχύτερη πρόσβαση στα αντικείμενα των ίδιων μοντέλων. Μια άλλη αντιμετώπιση του προβλήματος θα επέτρεπε την χρήση vectorizing, τεχνικής που σύμφωνα με τους δημιουργούς του R βελτιώνει την ταχύτητα εκτέλεσης. Προτιμήθηκε η αντικειμενοστραφή προσέγγιση καθώς επιτρέπει την εύκολη συντήρηση του κώδικα και την εισαγωγή νέων κλάσεων ώστε να μπορεί να προστεθεί εύκολα μια διαφορετική κατανομή αιτήσεων αποζημίωσης.

Βιβλιογραφία

[1] Anders, H. *A History of mathematical statistics from 1750 to 1930.* John Wiley & Sons, 1st edition, 1998.

[2] Bartlett, M.S. Comment on D.V. Lindley's Statistical Paradox. *Biometrika*, 44:533–534, 1957.

[3] Brooks, S.P., Giudici, P., Roberts, G.O. Efficient Construction of Reversible Jump Markov Chain Monte Carlo Proposal Distributions. *Journal of the Royal Statistical Society B*, 65:3–56, 2003.

[4] Bühlmann, H. Experience Rating and Probability. *ASTIN Bulletin*, 4:199–207, 1967.

[5] Carlin, B. A Simple Monte Carlo Approach to Bayesian Graduation. *Transactions of the Society of Actuaries*, 44:55–76, 1992.

[6] Carlin, B. and Thomas, L. *Bayes and Empirical Bayes Methods for Data Analysis.* Chapman & Hall, 2nd edition, 2000.

[7] Carlin, B.P. and Chib, S. Bayesian Model Choise via Markov Chain Monte Carlo Methods. *Journal of the Royal Statistical Society, B*, 157:473–484, 1995.

[8] Chen, M.H., Shao, Q.M. and Ibrahim, J.G. *Monte Carlo Methods in Bayesian Computation.* New York: Springer - Verlag, first edition, 2000.

[9] Consul, P.C., Jain, G.C. A Generalization of the Poisson Distribution. *Technometrics*, 15:791–799, 1973.

[10] de Alba, E. Bayesian Estimation of Outstanding Claim Reserves. *North American Actuarial Journal*, Volume 6 Number 4:1–20, 2002.

[11] Dellaportas, P., Forster, J.J. and Ntzoufras, I. On Bayesian Model and Variable Selection Using MCMC. *Statistics and Computing*, 12:27–36, 2002.

[12] Denuit, M. A new distribution of poisson-type for the number of claims. *Astin Bulletin*, 27:229–242, 1997.

[13] Douglas, J.B. Empirical fitting of discrete distributions. *Biometrics*, 50:576–579, 1994.

[14] Gilks, W.R., Richardson S., Spiegelhalter, D.J. *Markov Chain Monte Carlo in Practice*. Chapman and Hall, UK, 1996.

[15] Green, P. Reversible Jump Markov Chain Monte Carlo computation and Bayesian model determination. *Biometrika*, 82:711–732, 1995.

[16] Han, C. and Carlin, B. MCMC Methods for Computing Bayes Factors: a Comparative Review. *Journal of the American Statistical Association*, 96:1122–1132, 2001.

[17] Herzog, T. N. *Introduction to Credibility Theory*. Winsted: AC-TEX Publications, 1994.

[18] Hoeting, J. Methology for Bayesian Model Averaging: An Update. *to appear*, 2002.

[19] Hoeting, J.A., Madigan, D., Raftery, A.E. and Volinski, C.T. Bayesian Model Averaging: A Tutorial. *Statistical Science*, 14:382–417, 1999.

[20] Johnson N., Kotz S., Kemp A. *Univariate Discrete Distributions*. John Wiley & Sons, 2nd edition, 1993.

[21] Kass, R.E. and Raftery, A.E. Bayes factors. Technical Report 254 (Revision 3), Department of Statistics, University of Washington, July 6 1994.

[22] Lavine, J. and Schervish, M.J. Bayes Factors: What They are and What They are Not. *The American Statistician*, 53:119–123, 1999.

[23] Lindley, D.V. A Statistical Paradox. *Biometrika*, 44:187–192, 1957.

[24] Madigan, D. and York, Y. Bayesian Graphical Models for Discrete Data. *International Statistican Review*, 63:215–232, 1995.

[25] Madigan, D., Raftery, A. E. Model selection and accounting for model uncertainty in graphical models using Occam's window. *American Statistician*, 89:1535–1546, 1994.

[26] Makov, U. Principal Applications of Bayesian Methods in Actuarial Science: A Perspective. *North American Actuarial Journal*, Volume 5 Number 4:53–73, 2001.

[27] Makov, U. E., Smith, A. F. M., Liu, Y-H. Bayesian Methods in Actuarial Science. *The Statistician*, 45:503–515, 1996.

[28] Ntzoufras I. and Dellaportas P. Bayesian Modelling of Outstanding Liabilities Incorporating Claim Count Uncertainty. (with discussion). *North American Actuarial Journal*, Volume 6 Number 1:113–128, 2002.

[29] Ntzoufras, I., Katsis, A. and Karlis D. Bayesian assessment of the distribution of insurance claim counts using reversible jump MCMC. *To appear in*, 2005.

[30] Price, R. *Observations on Reversionary Payments: On Schemes for Providing Annuities for Widows and for Persons in Old Age; On the Method of Calculating the Values of Assurance on Lives; and On the National Dept.* London: T. Caddel and W. Davis, 1771.

[31] Ruohonen, M. On a model for the claim number process. *Astin Bulletin*, 18:57–68, 1988.

[32] Scollnik, D. An Introduction to Monte Carlo Markov Chain and Their Actuarial Applications. *Casualty Actuarial Society*, 158:114–165, 1996.

[33] Scollnik, D. On the Analysis of the Truncated Generalized Poisson Distribution using a Bayesian Method. *Astin Bulletin*, 28:135–152, 1998.

[34] Scollnik, D.P.M. Actuarial modeling with MCMC and BUGS. *North American Actuarial Journal*, 56:96–125, 2001.

[35] Sinharay, S. and Stern, H.S. On the Sensitivity of Bayes Factor to the Prior Distributions. *The American Statistician*, 56:196–201, 2002.

[36] Ter Berg, P. On the loglinear poisson and gamma model. *Astin Bulletin*, 11:35–40, 1980.

[37] Ter Berg, P. A Loglinear Lagrangian Poisson Model. *Astin Bulletin*, 26(1):123–129, 1996.

[38] Verrall, R. Bayes and Empirical Bayes estimation of the chain-ladder model. *Astin Bulletin*, 20(2):217–243, 1990.

[39] Verrall, R. An investigation into stochastic claims reserving models and the chain-ladder technique. *Insurance: Mathematics and Economics*, 26:91–99, 2000.

[40] Whittaker, E.T. On Some Disputed Questions of Probability. *Transactions of the Faculty of Actuaries*, 8:163–206, 1920.

Παράρτημα Α′

MCMC για κάθε μοντέλο

Στο πρώτο βήμα του αλγόριθμου RJMCMC (3.2) αναφερθήκαμε σε MCMC προσομοιώσεις που χρησιμοποιούμε για να παράγουμε τιμές από την posterior κατανομή της παραμέτρου για κάθε μοντέλο. Αρχικά θα περιγράψουμε το γενικό πλαίσιο ενός MCMC αλγορίθμου που ακολουθεί την μεθοδολογία των Metropolis - Hastings ώστε να καταλήξουμε να παράγουμε τιμές από την posterior κατανομή των παραμέτρων (λ, ϑ) αν επιλέγουμε το m_2 μοντέλο ή των (λ, ω) αν επιλέγουμε το m_3 μοντέλο. Το m_1 μοντέλο (Poisson) έχει συζυγή posterior κατανομή και μπορούμε να την υπολογίσουμε αναλυτικά.

Αρχικά θεωρούμε αυθαίρετες τιμές $\boldsymbol{\theta}_m^{(0)}$ και επαναλαμβάνουμε τα ακόλουθα βήματα έως η σύγκλιση να έχει εξασφαλιστεί. Θέτουμε για τα μοντέλα μας, $\boldsymbol{\theta}_{m1} = \lambda$, $\boldsymbol{\theta}_{m2} = (\lambda, \vartheta)^T$, $\boldsymbol{\theta}_{m3} = (\lambda, \omega)^T$. Αφαιρούμε τις πρώτες 1000 από τις συνολικά 3000 τιμές θεωρώντας ότι αποτελούν μια «burn-in period» δηλαδή εξαλείφουμε την επιρροή των αυθαίρετων αρχικών τιμών. Στα δεδομένα που χρησιμοποιούμε επιλέγουμε τις εξής αρχικές τιμές:

$$\lambda^{(0)} = \bar{y}, \quad \vartheta^{(0)} = \max\left\{ 0.01,\ \bar{y}^2/(s_y^2 - \bar{y}) \right\}, \quad \omega^{(0)} = \max\left\{ 0.01,\ 1 - \sqrt{\bar{y}/s_y^2} \right\} \qquad (\text{A}'.1)$$

όπου $\bar{y}$ είναι η δειγματική μέση τιμή των δεδομένων και s_y^2 η δειγματική διακύμανση.

Για να υλοποιήσουμε ένα MCMC αλγόριθμο για ένα μοντέλο m επαναλαμβάνουμε τα εξής βήματα για $t = 1, \ldots, T$:

ΒΗΜΑ 1: Θέτουμε $\boldsymbol{\theta} = \boldsymbol{\theta}_m(t)$.

ΒΗΜΑ 2: Για όλες τις παραμέτρους του μοντέλου $j = 1, \ldots, d_m$

1. Προτείνουμε θ'_j από $q(\theta'_j|\boldsymbol{\theta}_{\setminus j})$, όπου θ_j είναι το j-στο στοιχείο του διανύσματος[1] $\boldsymbol{\theta}$ και $\boldsymbol{\theta}_{\setminus j} = (\theta_1, \ldots, \theta_{j-1}, \theta_{j+1}, \ldots, \theta_{d_m})^T$.

2. Υπολογίζουμε την πιθανότητα αποδοχής:

$$a = \min\left\{1, \frac{f(\boldsymbol{y}|\boldsymbol{\theta}'_m, m) f(\boldsymbol{\theta}'_m|m) q(\theta'_j|\boldsymbol{\theta}_{\setminus j})}{f(\boldsymbol{y}|\boldsymbol{\theta}_m, m) f(\boldsymbol{\theta}_m|m) q(\theta_j|\boldsymbol{\theta}_{\setminus j})}\right\}$$

   όπου $\boldsymbol{\theta'} = (\theta_1, \ldots, \theta_{j-1}, \theta'_j, \theta_{j+1}, \ldots, \theta_{d_m})^T$

3. Παράγουμε την τ.μ. $u \sim U(0,1)$, όπου $U(0,1)$ είναι η ομοιόμορφη κατανομή στο διάστημα $(0,1)$.

4. Αν $\alpha > u$ τότε θέτουμε $\theta_j = \theta'_j$ διαφορετικά αφήνουμε το θ_j όπως έχει.

ΒΗΜΑ 3: Θέτουμε $\boldsymbol{\theta}_m^{(t+1)} = \boldsymbol{\theta}_m$.

Όπως είδαμε η posterior κατανομή του μοντέλου Poisson είναι συζυγής και παράγουμε τις τιμές του λ κατευθείαν από αυτή.

Για τα άλλα δύο μοντέλα, η κατανομή προσφοράς που αφορά την παράμετρο λ βασίζεται στην posterior κατανομή της παραμέτρου του απλού μοντέλου Poisson. Οπότε $q(\lambda'|\lambda)$ είναι μια κατανομή Γάμμα με παραμέτρους $Gamma(a + n\bar{y}, b + n)$. Η επιλογή αυτή λειτούργησε πολύ αποτελεσματικά στα δεδομένα που δοκιμάσαμε τον αλγόριθμο.

Σε ό,τι αφορά τις παραμέτρους ϑ και ω έχουμε χρησιμοποιήσει παραλλαγές του random walk Metropolis. Συγκεκριμένα, για την παράμετρο ϑ χρησιμοποιήσαμε την συνάρτηση προσφοράς $q(\vartheta'|\vartheta) = LN(\log\vartheta, C_\vartheta^2)$, όπου $LN(\mu, s^2)$ είναι η κατανομή Log-Normal με παραμέτρους μ, s^2 και συνάρτηση πυκνότητας πιθανότητας:

$$f(x) = \frac{1}{\sqrt{2\pi}sx} \exp\left\{-\frac{1}{2}\left(\frac{\log x - \mu}{s^2}\right)^2\right\}. \tag{Α'.2}$$

Για το ω χρησιμοποιήσαμε την κατανομή προσφοράς $q(\omega'|\omega) = Beta\left(C_\omega \frac{\omega}{1-\omega}, C_\omega\right)$ με μέση τιμή ω και συνάρτηση πυκνότητας πιθανότητας:

$$q(\omega'|\omega) = \frac{\Gamma\left(\frac{C_\omega}{1-\omega}\right)}{\Gamma\left(C_\omega \frac{\omega}{1-\omega}\right)\Gamma\left(C_\omega\right)} \omega'^{C_\omega \frac{\omega}{1-\omega}-1}(1-\omega')^{C_\omega - 1}. \tag{Α'.3}$$

Στις παραπάνω κατανομές προσφοράς υπάρχουν οι ποσότητες C_ϑ, C_ω είναι παράμετροι που τις ρυθμίζουμε κατάλληλα ώστε να επιτυγχάνουμε ρυθμό αποδοχής ανάμεσα στο $30 - 50\%$.

[1] Η j-στή παράμετρος του διανύσματος παραμέτρων.

MCMC για το μοντέλο της Αρνητικής Διωνυμικής

Εφαρμόζουμε τον παραπάνω αλγόριθμο πάνω στο μοντέλο της Αρνητικής Διωνυμικής, m_2, ώστε να παράγουμε παραμέτρους λ και θ. Θεωρώντας ότι (λ, θ) είναι οι τρέχουσες τιμές του αλγορίθμου έχουμε:

1. Παίρνουμε μια τιμή της λ από την $f(\lambda|\vartheta, \boldsymbol{y}, m_2)$ χρησιμοποιώντας την μέθοδο independent Metropolis:

   (α′) Προτείνουμε μια νέα υποψήφια τιμή λ' από την κατανομή $Gamma(a + n\bar{y}, b + n)$.

   (β′) Δεχόμαστε την προτεινόμενη τιμή με πιθανότητα:

$$\alpha = \min\left\{1, \frac{\lambda'}{\lambda}\left(\frac{\lambda+\vartheta}{\lambda'+\vartheta}\right)^{n\bar{y}+n\vartheta+3/2} e^{-n(\lambda-\lambda')}\right\}. \tag{A′.4}$$

2. Παίρνουμε μια τιμή της ϑ από την $f(\vartheta|\lambda, \boldsymbol{y}, m_2)$ χρησιμοποιώντας την μέθοδο Metropolis-Hastings:

   (α′) Προτείνουμε μια υποψηφια τιμή ϑ' από την κατανομή $LN(\log\vartheta, C_\vartheta^2)$ (A′.2).

   (β′) Δεχόμαστε την προτεινόμενη τιμή με πιθανότητα $\alpha = min\{1, A\}$, όπου A δίνεται από:

$$\begin{aligned}\log A \;=\; & \sum_{i=1}^{n} \log\frac{\Gamma(y_i+\vartheta')}{\Gamma(y_i+\vartheta)} + n\log\frac{\Gamma(\vartheta)}{\Gamma(\vartheta')} \\ & +(n\vartheta'+1/2)\log\vartheta' - (n\vartheta+1/2)\log\vartheta \\ & +(n\bar{y}-3/2)\log\frac{\lambda+\vartheta'}{\lambda+\vartheta} + n\vartheta'\log(\lambda+\vartheta') \\ & -n\vartheta\log(\lambda+\vartheta).\end{aligned} \tag{A′.5}$$

MCMC για το μοντέλο της Γενικευμένης Poisson

Αντίστοιχα στο μοντέλο m_3 (Γενικευμένη Poisson) εφαρμόζουμε τον γενικό αλγόριθμο για να παράγουμε τις παραμέτρους λ και ω. Θεωρώντας ότι (λ, ω) είναι οι τρέχουσες τιμές του αλγορίθμου έχουμε:

1. Παίρνουμε μια τιμή της λ από την $f(\lambda|\omega, \boldsymbol{y}, m_3)$ χρησιμοποιώντας την μέθοδο independent Metropolis:

   (α′) Προτείνουμε μια νέα υποψήφια τιμή λ' από την κατανομή $Gamma(a + n\bar{y}, b + n)$.

(β′) Δεχόμαστε την προτεινόμενη τιμή με πιθανότητα:

$$\alpha = \min\left\{1, \left(\frac{\lambda'}{\lambda}\right)^{n-n\bar{y}} e^{n\omega(\lambda'-\lambda)} \left[\prod_{i=1}^{n}\left(\frac{(1-\omega)\lambda' + \omega y_i}{(1-\omega)\lambda + \omega y_i}\right)^{y_i-1}\right]\right\}. \qquad (A'.6)$$

2. Παίρνουμε μια τιμή της ω από την $f(\omega|\lambda, \boldsymbol{y}, m_3)$ χρησιμοποιώντας εξής μέθοδο Metropolis-Hastings:

   (α′) Προτείνουμε μια υποψηφια τιμή ω' από την κατανομή $Beta\left(C_\omega \frac{\omega}{1-\omega}, C_\omega\right)$ (A′.3).

   (β′) Δεχόμαστε την προτεινόμενη τιμή με πιθανότητα $\alpha = min\{1, A\}$, όπου A δίνεται από:

$$\begin{aligned}
\log A &= (n - C_\omega + 1)\log\frac{1-\omega'}{1-\omega} - n(\bar{y}-\lambda)(\omega'-\omega) \\
&\quad + \sum_{i=1}^{n}\left((y_i - 1)\log\frac{(1-\omega')\lambda + \omega' y_i}{(1-\omega)\lambda + \omega y_i}\right) \\
&\quad + \log\frac{\Gamma\left(\frac{C_\omega}{1-\omega'}\right)}{\Gamma\left(\frac{C_\omega}{1-\omega}\right)} + \log\frac{\Gamma\left(C_\omega\frac{\omega}{1-\omega}\right)}{\Gamma\left(C_\omega\frac{\omega'}{1-\omega'}\right)} \\
&\quad + \left(C_\omega\frac{\omega'}{1-\omega'} - 1\right)\log\omega - \left(C_\omega\frac{\omega}{1-\omega} - 1\right)\log\omega'. \qquad (A'.7)
\end{aligned}$$

Με αυτή την υλοποίηση λαμβάνουμε τα δείγματα των (λ, ϑ) και (λ, ω). Και στις δύο συλλογές δεδομένων που χρησιμοποιήσαμε επιλέξαμε για τις σταθερές (C_ϑ, C_ω) τις τιμές $(0.001, 100)$.

Παράρτημα Β΄

Υλοποίηση στο R

Σε αυτό το παράρτημα δίνουμε τις κυριότερες συναρτήσεις που υλοποιήθηκαν στο στατιστικό πακέτο R, συμβατό με την γλώσσα προγραμματισμού S3, ενώ ακολουθήθηκε μια αντικειμενοστραφή (object-oriented) προσέγγιση.

Με αυτό τον τρόπο διαχωρίζονται τα μοντέλα σε κλάσεις και σε κάθε βήμα του αλγόριθμου προτείνεται ένα μοντέλο με αντίστοιχες παραμέτρους, υπό την μορφή αντικειμένου που αποτελεί μία πραγματοποίηση της αντίστοιχης κλάσης. Πλέον τα αντικείμενα μπορούν να συμπεριληφθούν σε δομές δεδομένων όπως λίστες και να επεξεργαστούν με εξειδικευμένες μεθόδους π.χ. print.rjmcmc για την εκτύπωση μια αλυσίδας που παράγει ο RJMCMC. Οι κλάσεις ορίστηκαν κατά αναλογία με τα μοντέλα, δηλαδή m1 για το m_1 μοντέλο κ.ο.κ. Το τελικό αποτέλεσμα, δηλαδή η αλυσίδα RJMCMC δίνεται από μια λίστα αναφορών σε αντικείμενα, δηλαδή κάθε βήμα του αλγόριθμου δίνεται από μια διπλή αναφορά (reference). Αρχικά εισάγουμε τα δεδομένα:

```
InsDat <- data.frame(
array(c(
  103704,14075,1766,255,45,6,2,0,
  57178,5618,446,50,8,0,0,0,
  c(8,2)),
  row.names=c(0,1,2,3,4,5,6,7))
names(InsDat)<-c("Switzerland 1961","Belgium 1993")
```

Στη συνέχεια ορίζουμε την κεντρική συνάρτηση που ορίζει τον αλγόριθμο RJMCMC.

```
rjmcmc <- function(y,iter=5000){
  y<<-y#make y global
  #y is a vector of data ex. InsDat[,1]
  #Count accepted tries to jump
  accm1m2<<-0
  accm1m3<<-0
  accm2m3<<-0
  accm2m1<<-0
  accm3m2<<-0
  accm3m1<<-0
  #Count total tries to jump
  totm1m2<<-0
  totm1m3<<-0
  totm2m3<<-0
  totm2m1<<-0
```

```
  totm3m2<<-0
  totm3m1<<-0
  #Step 1: Do pilot runs - generate parameters
  pilotruns()

  #choose a random model to start
  old <- generate()
  listmodels<-list(old)
  for(i in 1:iter){
      #select new model
      new <- select(old)
      #try to jump
      old<-accept(old,new)
      listmodels <- c(listmodels,list(old))
  }
  class(listmodels)<-"rjmcmc"
  return(listmodels)
}
```

Ορίζουμε μια συνάρτηση που θα εκτυπώνει με κατάλληλο τρόπο τα αντικείμενα που ανήκουν στη κλάση rjmcmc. Η συνάρτηση παίρνει την MCMC αλυσίδα του αλγόριθμου, τη διατρέχει ενώ ταυτόχρονα δημιουργεί νέες, μία σε κάθε μοντέλο. Κάθε νέα αλυσίδα περιέχει όλες τις επισκέψεις του αλγόριθμου στο εν λόγω μοντέλο. Ουσιαστικά είναι ένας γρήγορος τρόπος να εκφράσουμε τις δείκτριες συναρτήσεις που μας χρειάζονται για να εκτιμήσουμε την posterior πιθανότητα του μοντέλου (3.5).

```
print.rjmcmc <-function(blah){
cutx<-function(x,i=1){return(x[i+1:length(x)])} #cuts the first i elements
blah<-blah[1001:length(blah)]
xm1<-0
xm2a<-0
xm2b<-0
xm3a<-0
xm3b<-0
for(i in 1:length(blah)) {
  if(class(blah[[i]])=="m1")
  {
    xm1<-c(xm1,blah[[i]]$lambda)
  }
  if(class(blah[[i]])=="m2")
  {
    xm2a<-c(xm2a,blah[[i]]$lambda)
    xm2b<-c(xm2b,blah[[i]]$theta)
  }
  if(class(blah[[i]])=="m3")
  {
    xm3a<-c(xm3a,blah[[i]]$lambda)
    xm3b<-c(xm3b,blah[[i]]$omega)
  }
}
if(length(xm1)==1){length(xm1)<-0}
else{xm1<-xm1[2:length(xm1)]}
xm2a<-xm2a[2:length(xm2a)]
xm2b<-xm2b[2:length(xm2b)]
xm3a<-xm3a[2:length(xm3a)]
```

```
xm3b<-xm3b[2:length(xm3b)]

#Plots
def.par <- par(no.readonly = TRUE)# save default, for resetting...
    op <- par(mfrow = c(2, 2),bg="grey")
    plot(xm2a,type='l',main="Model 2 lambda",ylab="lambda",xlab="Iteration")
    plot(xm2b,type='l',main="Model 2 theta",ylab="theta",xlab="Iteration")
    plot(xm3a,type='l',main="Model 3 lambda",ylab="lambda",xlab="Iteration")
    plot(xm3b,type='l',main="Model 3 Omega",ylab="omega",xlab="Iteration")
  par(op)
  par(def.par)# reset to default

#Informations about the models
tot<-length(xm1)+length(xm2a)+length(xm3a)
if(length(xm1)==0){
      cat("Model 1:\nmean lambda: NA sd: NA\nPosterior Prob: 0\n")
      }
else{
      cat("Model 1:\nmean lambda: ",mean(xm1)," sd: ",sd(xm1,T),
            "\nPosterior Prob:",length(xm1)/tot,"\n")
    }
cat("Model 2:\nmean lambda: ",mean(xm2a)," sd: ",sd(xm2a,T),
      "\nmean theta: ",mean(xm2b)," sd: ",sd(xm2b,T),
      "\nPosterior Prob:",length(xm2a)/tot,"\n")
cat("Model 3:\nmean lambda: ",mean(xm3a)," sd: ",sd(xm3a,T),
      "\nmean omega: ",mean(xm3b)," sd: ",sd(xm3b,T),
      "\nPosterior Prob:",length(xm3a)/tot,"\n")
cat("Acceptance rates:\n")
cat("m1m2:",accm1m2/totm1m2,"\tm1m3:",accm1m3/totm1m3,"\n")
cat("m2m1:",accm2m1/totm2m1,"\tm2m3:",accm2m3/totm2m3,"\n")
cat("m3m1:",accm3m1/totm3m1,"\tm3m2:",accm3m2/totm3m2,"\n")
}
```

Ορίζουμε την συνάρτηση κατανομής πιθανότητας για την Γενικευμένη Poisson.

```
dgpoisson<- function(x,lambda,omega){
  if(omega<0 || omega>=1){stop("illegal value in omega [0,1)")}
  return(
            (1-omega)*lambda*
            (
                  ((1-omega)*lambda + omega * x)^(x-1)
            ) *
            exp(-((1-omega)*lambda + omega*x)) /
            gamma(x+1)
      )
}
```

Ορίζουμε την συνάρτηση Πιθανοφάνειας η οποία ανάλογα με τον τύπο του μοντέλο που θα έχει ως όρισμα θα επιλέγει την αντίστοιχη εξειδικευμένη συνάρτηση που φέρει στο τέλος τον τύπο του μοντέλου, π.χ. .m1 για το μοντέλο m_1.

```
loglikelihood <- function(x,...){
      UseMethod("loglikelihood")
}
loglikelihood.m1 <- function(x){
  lambda<-x$lambda
```

```
  l<-0
  for(i in 1:length(y)){
      l<-l+y[i]*log(dpois(i,lambda))
  }
  return(l)
}
loglikelihood.m2 <- function(x){
  l<-0
  lambda<-x$lambda
  theta<-x$theta
  for(i in 1:length(y)){
      l<-l+y[i]*(log(dnbinom(i,size=theta,mu=lambda)))
  }
  return(l)
}
loglikelihood.m3 <- function(x){
  l<-0
  lambda<-x$lambda
  omega<-x$omega
  for(i in 1:length(y)){
      l<-l+y[i]*(log(dgpoisson(i,lambda,omega)))
  }
  return(l)
}
loglikelihood.default <- function(x){
      stop("Not an Candidate Model, cannot compute likelhood")
}
```

Δεδομένου ότι γνωρίζουμε το μοντέλο που είμαστε επιλέγουμε στην τύχη ένα από τα υπόλοιπα δύο κάθε φορά (υλοποιείται στη μέθοδο select στη συνέχεια). Υπολογίζουμε την πιθανότητα ενός άλματος στο νέο μοντέλο και αποφασίζουμε αν τελικά γίνει (Βήματα 2 και 3 του RJMCMC αλγόριθμου).

```
accept <- function(old,...){
      UseMethod("accept")
}
accept.m1 <- function(old,new){
      if(class(new)=="m2"){#m1->m2
            a<-min(1,m1m2(old,new))
            totm1m2<<-totm1m2+1
            if(runif(1)<a){
                  accm1m2<<-accm1m2+1
                  return(new)
            }
            else{return(old)}
      }
      if(class(new)=="m3"){#m1->m3
            a<-min(1,m1m3(old,new))
            totm1m3<<-totm1m3+1
            if(runif(1)<a){
                  accm1m3<<-accm1m3+1
                  return(new)
            }
            else{return(old)}
      }

}
```

```
accept.m2 <- function(old,new){
      if(class(new)=="m1"){#m2->m1
            a<-min(1,1/m1m2(new,old))
            totm2m1<<-totm2m1+1
            if(runif(1)<a){
                  accm2m1<<-accm2m1+1
                  return(new)
            }
            else{return(old)}
      }
      if(class(new)=="m3"){#m2->m3
            a<-min(1,m2m3(old,new))
            totm2m3<<-totm2m3+1
            if(runif(1)<a){
                  accm2m3<<-accm2m3+1
                  return(new)
            }
            else{return(old)}
      }
}

accept.m3 <- function(old,new){
      if(class(new)=="m1"){#m3->m1
            a<-min(1,1/m1m3(new,old))
            totm3m1<<-totm3m1+1
            if(runif(1)<a){
                  accm3m1<<-accm3m1+1
                  return(new)
            }
            else{return(old)}
      }
      if(class(new)=="m2"){#m3->m2
            a<-min(1,1/(m2m3(new,old)))
            totm3m2<<-totm3m2+1
            if(runif(1)<a){
                  accm3m2<<-accm3m2+1
                  return(new)
            }
            else{return(old)}
      }
}

m1m2<-function(old,new){
      mistheta<-rlnorm(1,ThetaLogNormA,ThetaLogNormB)
      ## Constructed theta for the m1 model
      return(exp(loglikelihood(new)-loglikelihood(old))*
                  dbetaii(new$theta,new$lambda)*
                  dgamma(new$lambda,0.0001,0.0001)/
                  (dgamma(old$lambda,0.0001,0.0001)*
                   dlnorm(mistheta,ThetaLogNormA,ThetaLogNormB))
            )
}

m1m3<-function(old,new){
      misomega<-rbeta(1,OmegaBetaA,OmegaBetaB)
      ## Constructed omega for the m1 model
```

```
      return(exp(loglikelihood(new)-loglikelihood(old))*
                  dunif(new$omega,0,1)*
                  dgamma(new$lambda,.0001,.0001)/
                  (dgamma(old$lambda,.0001,.0001)*
                  dbeta(misomega,OmegaBetaA,OmegaBetaB))
            )
}
m2m3<-function(old,new){
      return(     exp(loglikelihood(new)-loglikelihood(old))*
                  (
                    .5*(1+old$lambda/old$theta)^(-3/2) *
                    (old$lambda/(old$theta^2))
                  )*
                  dunif(new$omega,0,1)*
                  dgamma(new$lambda,.0001,.0001)/
                  (
                    dbetaii(old$theta,old$lambda)*
                    dgamma(old$lambda,.0001,.0001)
                  )
            )
}
```

Επιλέγουμε ένα μοντέλο στην τύχη και το κατασκευάζουμε. Αυτό μας διευκολύνει για να αρχίσουμε τον αλγόριθμο RJMCMC από ένα τυχαίο μοντέλο.

```
generate <- function(x,...){
  prob<-runif(1)
  if (prob<=1/3){
      #Negative Binomial
      #THETA ~ LogNormal
      new <- list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=rlnorm(1,ThetaLogNormA,ThetaLogNormB),omega=0)
      class(new)<-"m2"
      return(new)
  }
  if(prob<=2/3){
      #Generalized Poisson
      #OMEGA ~ BETA(a,b)
      new <- list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=0,omega=rbeta(1,OmegaBetaA,OmegaBetaB))
      class(new)<-"m3"
      return(new)
  }
  else{
      #lambda of Poisson: LAMBDA|X ~ GAMMA(SUM YI + a, n+b)
      new<-list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=0,omega=0)
      class(new)<-"m1"
      return(new)
  }
}
```

Ορίζουμε μια μέθοδο που ανάλογα με το μοντέλο που είμαστε δημιουργεί στην τύχη ένα υποψήφιο. (Βήμα 1 του RJMCMC αλγόριθμου).

```
select <- function(x,...){
      #Method dispatching
      UseMethod("select")
```

```
}

#We have a m1 model
select.m1 <- function(x){
  if (runif(1)<=.5){
      #Theta of Negative Binomial
      #THETA ~ LogNormal
      new <- list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=rlnorm(1,ThetaLogNormA,ThetaLogNormB),omega=0)
      class(new)<-"m2"
      return(new)
  }
  #Generalized Poisson
  #OMEGA ~ BETA(a,b)
  new <- list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=0,omega=rbeta(1,OmegaBetaA,OmegaBetaB))
  class(new)<-"m3"
  return(new)
}

#We have a m2 model
select.m2 <- function(x){
  if (runif(1)<=.5){
      #lambda of Poisson: LAMBDA|X ~ GAMMA(SUM YI + a, n+b)
      new<-list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=0,omega=0)
      class(new)<-"m1"
      return(new)
  }
  #Generalized Poisson
  #OMEGA ~ BETA(a,b)
  new <- list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=0,omega=rbeta(1,OmegaBetaA,OmegaBetaB))
  class(new)<-"m3"
  return(new)
}

#We have a m3 model
select.m3 <- function(x){
  if (runif(1)<=.5){
      #lambda of Poisson: LAMBDA|X ~ GAMMA(SUM YI + a, n+b)
      new<-list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=0,omega=0)
      class(new)<-"m1"
      return(new)
  }
  #Theta of Negative Binomial
  #THETA ~ LogNormal
  new <- list(lambda=rgamma(1,LambdaGammaA,LambdaGammaB),
            theta=rlnorm(1,ThetaLogNormA,ThetaLogNormB),omega=0)
  class(new)<-"m2"
  return(new)
}
```

Συνάρτηση πυκνότητας πιθανότητας της Βήτα τύπου II (Beta type II) υπό κλίμακα (2.6).

```
dbetaii<-function(theta,lambda){
      return(.5*(lambda/theta^2)*(1+lambda/theta)^(-3/2))
```

```
}
```

Για τα μοντέλα m_2, m_3 ξεκινά τις απλές MCMC εξομοιώσεις , παίρνει τα outputs για να υπολογίσει τις παραμέτρους για τις κατανομές προσφοράς. Για το μοντέλο m_1 και υπολογίζει την posterior κατανομή του λ.

```
pilotruns<-function(){
  #Initial values for THETA|X
  inittheta<-log(samplebnb(y,3000)$theta)
  inittheta<-inittheta[1001:length(inittheta)]
  #THETA|X ~LOGNORMAL(lognorma,lognormb)
  ThetaLogNormA<<-mean(inittheta)
  ThetaLogNormB<<-var(inittheta)

  #Inital Values for OMEGA|X
  initomega<-samplegp(y,3000)$omega
  initomega<-initomega[1001:length(initomega)]
  #OMEGA ~ BETA(omegabetaa,omegabetab)
  minitomega<-mean(initomega)
  vinitomega<-var(initomega)
  OmegaBetaA<<-minitomega*(
              (minitomega*(1-minitomega))/
                  vinitomega -1
              )
  OmegaBetaB<<-OmegaBetaA*(1-minitomega)/minitomega

  #Inital Values for Lambda|x
  #LAMDA ~ GAMMA(LambdaGammaA,LambdaGammaB)
  #The posterior distribution is conjugate
  LambdaGammaA<<-sum(y*c(0:7))+0.0001
  LambdaGammaB<<-sum(y)+0.0001
}
```

Υλοποιεί τον MCMC αλγόριθμο για το μοντέλο της Γενικευμένης Poisson (Παράρτημα Α′).

```
"samplegp" <- function(y,iter=1000 ,lambda = sum(y*c(0:7))/sum(y),
      omega = max(0.01,1-sqrt((sum(y*c(0:7))/sum(y))/
            (( (sum(y) * sum( c(0:7)^2 ) )  - sum(y*c(0:7))^2 )/
            ( sum(y)*(sum(y)-1) )))))
{
 #Initalize
 if(is.vector(y)){
      n <- sum(y)
      my<-sum(y*c(0:7))/sum(y)
 }
 else{
      cat("y not a vector\n")
      return()
}
 j <- 1
 #Numeric Vectors for lambda-omega
 lamdaarr <- numeric(iter)
 omegaarr <- numeric(iter)
 lamdaarr[1]<-lambda
```

```
omegaarr[1] <- omega
acclambda <- 0
accomega <- 0
comega<-100  #needed in omega's Prior (Beta), LOGLIKEHOOD RATIO
ga <- .0001  #Needed in lambda's Prior (Gamma)
gb <- .0001  #Needed in lambda's Prior (Gamma)
while(j < iter){
     ############################################################
     #1. Sample lambda
     ############################################################

     #1.a. Propose new candidate lambdanew from GAMMA(ny+a,n+b)
     lambdanew <- rgamma(1,shape = (n*my+ga),rate =(n+gb))

     #1.b. Accept the proposed value
     lfact <- 0
     for (i in 1:8) {
           lfact <- lfact + (i-2)*y[i]*log(((1-omega)*lambdanew+omega*(i))/
           ((1-omega)*lambda+omega*(i)))
     }
     accratio<-(lambdanew/lambda)^(n-n*my) * exp(n*omega*(lambdanew-lambda)) *
                 exp(lfact)

     if(is.na(accratio)){accratio<-0}
     if(!is.finite(accratio)){accratio<-1}
     alpha <- min(1,accratio)

     if (alpha ==1){
           lambda<-lambdanew
           acclambda <- acclambda + 1
     }
     else{
           if( runif(1) <= alpha ){
                 lambda<-lambdanew
                 acclambda <- acclambda + 1
           }
     }

     ############################################################
     #2. Sample Omega
     ############################################################

     #2.a. Propose omeganew

     omeganew <- rbeta(1, comega*(omega/(1-omega)), comega)

     #2.b. Accept the proposed value
     omegafact <- 0
     for (i in 1:8) {
           omegafact <- omegafact + y[i]*(i-2)*
                 log( ((1-omeganew) * lambda + omeganew * (i-1)) /
                 ((1-omega) * lambda + omega * (i-1)))
     }
     logaccratio <- (n - comega + 1) * log((1-omeganew)/(1-omega)) -
           n*(my-lambda)*(omeganew-omega) +
           omegafact +
           log(gamma(comega/(1-omeganew))/gamma(comega/(1-omega)))+
```

```
            log(  gamma((comega*omega)/(1-omega))/
                  gamma((comega*omeganew)/(1-omeganew))) +
            (comega*(omeganew/(1-omeganew))-1)*log(omega) -
            (comega*(omega/(1-omega))-1)*log(omeganew)

      accratio<-exp(logaccratio)
      if(is.na(accratio)){accratio<-0}
      if(!is.finite(accratio)){accratio<-1}
      alpha <- min(1,accratio)
      if (alpha ==1){
            omega<-omeganew
            accomega <- accomega + 1
      }
      else{
            if( runif(1) <= alpha ){
                  omega<-omeganew
                  accomega <- accomega + 1
            }
      }
      j <- j + 1
      lamdaarr[j]<-lambda
      omegaarr[j] <- omega
 }
 return(list(lambda = lamdaarr,
             omega = omegaarr,
             acclambda = (acclambda/iter),
             accomega=(accomega/iter)))
}
```

Υλοποιεί τον MCMC αλγόριθμο για το μοντέλο της Αρνητικής Διωνυμικής (Παράρτημα Α′).

```
"samplebnb" <- function(y,iter=1000 ,lambda = sum(y*c(0:7))/sum(y),
theta = max(0.01,
     (sum(y*c(0:7))/sum(y))^2/(
     ( (sum(y) * sum( y*c(0:7)^2 ) )  - sum(c(0:7))^2 )/ (sum(y)*(sum(y)-1))
     -(sum(y*c(0:7))/sum(y)))
     ))
{
 #Initalize
 if(is.vector(y)){
      n <- sum(y)
      my<-sum(y*c(0:7))/sum(y)
 }
 else{
      stop("y not a vector\n")
}
 iter<-iter
 lamdaarr <- numeric(iter)
 thetaarr <- numeric(iter)
 lamdaarr[1]<-lambda
 thetaarr[1] <- theta
 j <- 1
 acclambda <- 0
 acctheta <- 0
 ga <- .0001
 gb <- .0001
 while(j < iter){
```

```
#############################################################
#1. Sample lambda
#############################################################

#1.a. Propose new candidate lambdanew from GAMMA(ny+a,n+b)
lambdanew <- rgamma(1,shape = (n*my+ga),rate =(n+gb))

#1.b. Accept the proposed value
logaccratio<-log(lambdanew)-
      log(lambda)+
      (n*my+n*theta+1.5)*(log(lambda+theta)-
      log(lambdanew+theta))-
      n*(lambda-lambdanew)

accratio<-exp(logaccratio)

if(!is.finite(accratio)){accratio=1}
alpha <- min(1,accratio)

if (alpha ==1){
      lambda<-lambdanew
      acclambda <- acclambda + 1
}
else{
      if( runif(1) <= alpha ){
            lambda<-lambdanew
            acclambda <- acclambda + 1
      }
}


#############################################################
#2. Sample Theta
#############################################################
#2.a. Propose thetanew

thetanew <- rlnorm(1, log(theta), sdlog = 0.001)

#2.b. Accept the proposed value
A <- 0
for (i in 1:8){
  A<- A + y[i]*log(gamma(i-1+thetanew)/gamma(i-1+theta))
}

A<- A+ n*log(gamma(theta)/gamma(thetanew)) +
  (n*thetanew+.5)*log(thetanew)-
  (n*theta+.5)*log(theta)+
  (n*my-3/2)*log((lambda+thetanew)/(lambda+theta))+
  (n*thetanew)*log(lambda+thetanew)-
   n*theta*log(lambda+theta)

alpha<-exp(A)
alpha <- min(1,accratio)
#Test for acceptance
if (alpha ==1){
      theta<-thetanew
      acctheta <- acctheta + 1
```

```
      }
      else{
            if( runif(1) <= alpha ){
                  theta<-thetanew
                  acctheta <- acctheta + 1
            }
      }
      #Update
      j <- j + 1
      lamdaarr[j] <- lambda
      thetaarr[j] <- theta
 }
 return(list(lambda = lamdaarr,
             theta = thetaarr,
             acclambda = (acclambda/iter),
             acctheta=(acctheta/iter)))
}
```