
# ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ

ΔΠΜΣ «ΣΥΣΤΗΜΑΤΑ ΑΥΤΟΜΑΤΙΣΜΟΥ»

ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ

***Ανάλυση Μπεϋζιανών Δικτύων και Εφαρμογή για Πρόβλεψη Κινδύνου σε Νανοϋλικά***

**Ευάγγελος Ρουμελιώτης**

**Επιβλέπων Καθηγητής: Χαράλαμπος Σαρίμβεης**

Αθήνα, Οκτώβριος 2020

# Ευχαριστίες

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα της διπλωματικής μου εργασίας , Χαράλαμπο Σαρίμβεη, Καθηγητή Ε.Μ.Π της Σχολής Χημικων Μηχανικών, τόσο για την εμπιστοσύνη που μου έδειξε όσο και για την ευκαιρία που μου έδωσε να ασχοληθώ με ένα τόσο ενδιαφέρον θέμα.

Επίσης, θα ήθελα να ευχαριστήσω τον υποψήφιο διδακτορικό φοιτητή Περικλή Τσίρο για την συνεχή καθοδήγηση του σε όλα τα στάδια εκπόνησης της παρούσας διπλωματικής εργασίας, η οποία ήταν καταλυτική για την πραγματοποίησή της.

Θα ήθελα ακόμη να εκφράσω τις ειλικρινείς μου ευχαριστίες στους καθηγητές του Διατμηματικού Προγράμματος Μεταπτυχιακών Σπουδών “Συστήματα Αυτοματισμού”, τα μαθήματα των οποίων παρακολούθησα, για τη συμβολή που είχαν στην ολοκλήρωση των σπουδών μου.

Τέλος, θα ήθελα να ευχαριστήσω την οικογένεια μου, για την συμπαράσταση και την υπομονή που επέδειξαν καθόλη τη διάρκεια των σπουδών μου.

# Περίληψη

Στο πλαίσιο της παρούσας διπλωματικής εργασίας έγινε μία εις βάθος ανάλυση των μπεϋζιανών δικτύων και των αλγορίθμων που σχετίζονται με την εύρεση δομής, στατιστικού μοντέλου και πρόβλεψης σε αυτά. Σε κάθε κεφάλαιο δημιουργήθηκαν πολλαπλά παραδείγματα για την πλήρη εμπέδωση εννοιών και αλγορίθμων. Την εκτενή θεωρητική κάλυψη του αντικειμένου ακολούθησε μία μελέτη περίπτωσης, όπου προβλήθηκε η δυνατότητα αιτιακής μοντελοποίησης και πρόβλεψης του κινδύνου νανοϋλικών.

Η εργασία ξεκινάει με θεμελίωση βασικών εννοιών Στατιστικής και Πιθανοτήτων, όπως η έννοιας της πιθανότητας, η έννοια της στοχαστικής ανεξαρτησίας και το θεώρημα του Bayes, εργαλεία απαραίτητα για να γίνουν πλήρως αντιληπτές οι έννοιες που παρουσιάζονται σε επόμενα κεφάλαια. Στη συνέχεια, αναλύονται δομικά στοιχεία της θεωρίας των γραφημάτων που είναι απαραίτητα για τον ορισμό και την κατανόηση των μπεϋζιανών δικτύων, όπως τα κατευθυνόμενα ακυκλικά γραφήματα (directed acyclic graphs) και το κριτήριο του διαχωρισμού κατεύθυνσης (direction separation) μεταξύ των κόμβων.

Στη συνέχεια παρουσιάζεται η ανάλυση των μπεϋζιανών δικτύων, ξεκινώντας με τον ορισμό και τις ιδιότητες τους. Ακολουθεί η παρουσίαση των βασικών μεθόδων εύρεσης παραμέτρων, όπως η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας, η μπεϋζιανή προσέγγιση και ο αλγόριθμος EM (Expectation Maximization). Επιπλέον, αναλύονται οι δύο υπερκατηγορίες αλγορίθμων εύρεσης δομής μπεϋζιανών δικτύων, δηλαδή οι αλγόριθμοι με περιορισμούς (constraint-based) και οι αλγόρθμοι που βασίζονται στη βαθμολόγηση (score-based). Η ανάλυση κλείνει με αναφορά σε μεθοδολογίες ακριβούς και προσεγγιστικής συμπερασματολογίας/πρόβλεψης.

Το δεύτερο μέρος της εργασίας ασχολείται με την εφαρμογή των προηγούμενων μεθόδων σε ένα πρόβλημα κατηγοριοποίησης. Συγκεκριμένα, ο στόχος αυτού του μέρους ήταν η πρόβλεψη των κλάσεων κινδύνου νανοϋλικών χρησιμοποιώντας το σετ δεδομένων των Marvin *et al.* (2017). Δοκιμάστηκαν διάφορες στρατηγικές κατασκευής μοντέλου, οι οποίες περιελάμβαναν ένα μεγάλο εύρος τεχνικών και αλγορίθμων, από χρήση έτοιμου δικτύου μέχρι κατασκευή δικτύου και από στατική μέχρι στοχαστική πρόβλεψη. Με χρήση διάφορων στατιστικών μέτρων προκρίθηκε η καλύτερη στρατηγική, η οποία εφαρμόστηκε εξ αρχής στο ίδιο σετ δεδομένων αλλά

ακολουθώντας τον διαχωρισμό μεταξύ σετ εκπαιδευσης (training set) και δοκιμής (test set) που ακολούθησαν και οι Marvin *et al.* (2017), ώστε να μπορεί να διεξαχθεί σύγκριση μεταξύ των δύο μεθοδολογιών.Μετά την παρουσίαση των αποτελεσμάτων ακολουθεί η αναφορά των συμπερασμάτων που εξήχθησαν μέσα από αυτή την εργασία.

# Abstract

In the present thesis, an in-depth analysis of Bayesian networks was performed, stretchiing from algorithms related to finding their structure amd statistical model to prediction algorithms. In each chapter, multiple examples were created for attaining a complete understanding of concepts and algorithms. The extensive theoretical coverage of the subject was followed by a case study, which demonstrated the possibility of causally modeling and predicting the hazard of nanomaterials.

The thesis begins with the establishment of basic concepts of Statistics and Probability, such as the concept of probability, the concept of stochastic independence and Bayes' theorem, tools necessary for fully understanding the concepts presented in the subsequent chapters. On top of that, structural elements of graph theory that are necessary for the definition and understanding of Bayesian networks are analyzed, such as directed acyclic graphs and the d-separation criterion between nodes.

Following that, an analysis of Bayesian networks takes place, starting with their definition and properties. After that the basic parameter estimation methods are presented, such as the maximum likelihood estimation, the Bayesian approach and the EM (Expectation Maximization) algorithm. In addition, the two subcategories of Bayesian network structure learning algorithms are analyzed, namely constraint-based and score-based algorithms. The theoretical analysis comes to an end with a reference to accurate and approximate inference.

The second part of this work deals with the application of the previous methods to a classification problem. Specifically, the purpose of this section was to predict the hazard of nanomaterials using the data set of Marvin et al. (2017). Various model building strategies were tested, which included a wide range of techniques and algorithms, from predefined structure to structure learning and from static to stochastic prediction. Using various statistical measures, the best strategy was selected, which was then applied from the beginning in the same data set, but following the separation between training and test set presented in Marvin et al. (2017), so that a comparison can be made between the two methodologies. After the presentation of the results, the conclusions drawn from this work are reported.

# Περιεχόμενα

Ευχαριστίες..... II

Περίληψη..... III

Abstract ..... V

Κατάλογος Εικόνων ..... IX

Κατάλογος Πινάκων..... XI

1 Εισαγωγή ..... - 1 -

2 Στοιχεία Πιθανοτήτων..... - 6 -

2.1 Βασικές Έννοιες Πιθανοτήτων ..... - 7 -

2.1.1 Στατιστική Πιθανότητα ..... - 7 -

2.1.2 Υποκειμενική Πιθανότητα ..... - 8 -

2.1.3 Αξιωματικός Ορισμός Πιθανότητας ..... - 8 -

2.1.4 Δεσμευμένη Πιθανότητα και Αποτελέσματα..... - 9 -

2.1.5 Θεώρημα του Bayes ..... - 10 -

2.1.6 Ανεξαρτησία Ενδεχομένων..... - 10 -

2.2 Τυχαίες Μεταβλητές και Ιδιότητες τους ..... - 13 -

2.2.1 Τυχαία Μεταβλητή..... - 13 -

2.2.2 Συνάρτηση Κατανομής Πιθανότητας..... - 13 -

2.2.3 Διακριτή Τυχαία Μεταβλητή και Συνάρτηση Μάζας Πιθανότητας .. - 13 -

2.2.4 Συνεχής Τυχαία Μεταβλητή και Συνάρτηση Πυκνότητας Πιθανότητας.. - 14 -

2.3 Πολυδιάστατες Τυχαίες Μεταβλητές, Κατανομές και Ιδιότητές τους..... - 16 -

2.3.1 Συνάρτηση Κατανομής Πιθανότητας Τυχαίου Διανύσματος ..... - 16 -

2.3.2 Διακριτές Τυχαίες Μεταβλητές και η Από Κοινού Συνάρτηση Μάζας Πιθανότητας ..... - 16 -

2.3.3 Συνεχείς Τυχαίες Μεταβλητές και η Από Κοινού Συνάρτηση Πυκνότητας Πιθανότητας - 17 -
2.3.4 Περιθώριες Κατανομές - 18 -
2.3.5 Δεσμευμένες Κατανομές - 19 -
2.3.6 Ανεξαρτησία Τυχαίων Μεταβλητών - 19 -
2.4 Μέση Τιμή Τυχαίων Μεταβλητών - 21 -
2.4.1 Μέση Τιμή τυχαίας μεταβλητής - 21 -
2.4.2 Μέση Τιμή συνάρτησης τυχαίας μεταβλητής - 22 -
2.4.3 Μέση Τιμή Τυχαίου Διανύσματος - 23 -
3 Στοιχεία Γραφημάτων - 24 -
3.1 Βασικές Έννοιες Γραφημάτων - 25 -
3.2 Βασικές Συνδέσεις Μεταβλητών/Κόμβων σε DAG - 29 -
3.3 Κριτήριο D-Separation και η Κουβέρτα του Markov - 36 -
4 Μπεϋζιανά Δίκτυα - 43 -
4.1 Μπεϋζιανά Δίκτυα και Πίνακες Δεσμευμένων Πιθανοτήτων - 44 -
4.2 Παραμετροποίηση Μπεϋζιανού Δικτύου - 50 -
4.3 Εύρεση Παραμέτρων - 53 -
4.3.1 Μέθοδος Εκτίμησης Μέγιστης Πιθανοφάνειας (ΕΜΠ) - 53 -
4.3.1.1 Η ΕΜΠ σε Μπεϋζιανό Δίκτυο και η Ολική Αποσύνθεση Πιθανοφάνειας - 55 -
4.3.1.2 Η ΕΜΠ σε Μπεϋζιανό Δίκτυο και η Τοπική Αποσύνθεση Πιθανοφάνειας - 58 -
4.3.2 Μπεϋζιανή Εκτίμηση Παραμέτρων - 69 -
4.3.2.1 Μπεϋζιανή Εκτίμηση Παραμέτρων σε Μπεϋζιανό Δίκτυο - 72 -
4.3.2.2 Μετα-Δίκτυο και Ολική Παραμετρική Ανεξαρτησία - 75 -
4.3.2.3 Ολική Αποσύνθεση της εκ των Υστέρων Κατανομής - 79 -
4.3.2.4 Τοπική Αποσύνθεση της εκ των Υστέρων Κατανομής - 81 -
4.3.2.5 Εκ των Προτέρων Κατανομές - 83 -

4.3.3 Εύρεση Παραμέτρων Χωρίς Πλήρη Δεδομένα με τον Αλγόριθμο ΕΜ- 87 -
4.4 Εύρεση Δομής Μπεϋζιανού Δικτύου ..... - 95 -
4.4.1 Constraint based μέθοδοι ..... - 96 -
4.4.2 Score based μέθοδοι ..... - 96 -
4.4.3 Δομικός Αλγόριθμος ΕΜ ..... - 101 -
4.5 Διαδικασία Πρόβλεψης ..... - 105 -
4.5.1 Μέθοδος Διαγραφής Μεταβλητών ..... - 110 -
4.5.2 Στάθμιση Πιθανοφάνειας (Likelihood Weighting) ..... - 116 -
5 Μελέτη Περίπτωσης ..... - 121 -
5.1 Περιγραφή του Προβλήματος ..... - 122 -
5.2 Μεθοδολογία ..... - 126 -
5.2.1 Επεξεργασία Δεδομένων ..... - 126 -
5.2.2 Κατασκευή Μπεϋζιανών Δικτύων ..... - 127 -
5.2.3 Αξιολόγηση Δικτύων ..... - 132 -
5.2.3.1 Πίνακας Σύγχυσης (Confusion Matrix) ..... - 133 -
5.2.3.2 Στατιστικοί Δείκτες ..... - 135 -
5.2.3.3 Cross Validation ..... - 136 -
5.3 Αποτελέσματα ..... - 138 -
5.3.1 Επιλογή Βέλτιστης Στρατηγικής Εκπαίδευσης ..... - 138 -
5.3.2 Σύγκριση Αποτελεσμάτων με Marvin et al. (2017) ..... - 145 -
5.3.2.1 Χρήση Διαφορετικού Αρχικού Δικτύου ..... - 150 -
6 Συμπεράσματα ..... - 155 -
Βιβλιογραφία ..... - 158 -
Παράρτημα ..... - 161 -

# Κατάλογος Εικόνων

Εικόνα 1. Είδη συνδέσεων μεταξύ δύο μεταβλητών $vi - vj$ ..... - 25 -
Εικόνα 2. Παράδειγμα σύνδεσης μεταξύ των κόμβων σε ένα κατευθυνόμενο γράφημα. ..... - 26 -
Εικόνα 3. Παράδειγμα σύνδεσης μεταξύ των κόμβων σε ένα μη κατευθυνόμενο γράφημα ..... - 26 -
Εικόνα 4. Παράδειγμα βρόχου σε γράφημα. ..... - 26 -
Εικόνα 5. Παράδειγμα κύκλων σε κατευθυνόμενο γράφημα. ..... - 27 -
Εικόνα 6. Παράδειγμα DAG γραφήματος. ..... - 27 -
Εικόνα 7. Παράδειγμα σύνδεσης σε σειρά. ..... - 29 -
Εικόνα 8. Παράδειγμα σύνδεσης σε σειρά με κόμβους Καπνός (Κ), Φωτιά (Φ) και Πυροσβεστική (Π). ..... - 30 -
Εικόνα 9. Παράδειγμα αποκλίνουσας σύνδεσης. ..... - 31 -
Εικόνα 10. Παράδειγμα αποκλίνουσας σύνδεσης με κόμβους Κάπνισμα (Κ), Πνευμονοπάθια (Π) και Εγκεφαλικό επεισόδιο (Ε). ..... - 32 -
Εικόνα 11. Παράδειγμα συγκλίνουσας σύνδεσης. ..... - 33 -
Εικόνα 12. Παράδειγμα συγκλίνουσας σύνδεσης με απογόνους. ..... - 34 -
Εικόνα 13. Παράδειγμα συγκλίνουσας σύνδεσης με κόμβους Μηχανική βλάβη (Μ), Διαρροή αερίου (Δ) και Ένδειξη κινδύνου (Ε). ..... - 34 -
Εικόνα 14. Πιθανές συνδέσεις μεταξύ τριών κόμβων σε ένα γράφημα DAG. ..... - 37 -
Εικόνα 15. Παράδειγμα DAG με evidence node. ..... - 38 -
Εικόνα 16. Παράδειγμα DAG για την αποσαφήνιση του Markov Blanket. ..... - 41 -
Εικόνα 17. Παράδειγμα DAG για την εξαγωγή της από κοινού συνάρτησης πιθανότητας του δικτύου. ..... - 45 -
Εικόνα 18. Παράδειγμα τριών κόμβων σε σύνδεση σε σειρά. ..... - 62 -
Εικόνα 19 Σχηματική αναπαράσταση μεταδικτύου (Koller and Friedman 2009). ..... - 77 -
Εικόνα 20. Σχηματική αναπαράσταση μεταδικτύου με χρήση στιγμιοτύπων (Koller and Friedman 2009). ..... - 77 -
Εικόνα 21. Απομόμωση συγκεκριμένου στιγμιοτύπου από το μεταδίκτυο. ..... - 79 -
Εικόνα 22. Περαιτέρω ανάλυση του μεταδικτύου της Εικόνας 19 (Σχηματική αναπαράσταση μεταδικτύου (Koller and Friedman 2009). ..... - 81 -
Εικόνα 23. Παράδειγμα για την διασαφήνιση της τοπολογικής διάταξης. ..... - 117 -
Εικόνα 24. Τελικό μπεϋζιανό δίκτυο της έρευνας των Marvin *et al.* (2017). ..... - 125 -

Εικόνα 25. Διάγραμμα ροής εργασίας. ..... - 126 -
Εικόνα 26. Στρατηγική εκτίμησης δεδομένων με βάση το γνωστό δίκτυο των Marvin *et al*. (2017). ..... - 129 -
Εικόνα 27. Στρατηγική κατασκευής δικτύου από τα δεδομένα με constrained based αλγόριθμους. ..... - 130 -
Εικόνα 28. Στρατηγική κατασκευής δικτύου από τα δεδομένα με score based αλγόριθμους και τυχαίο αρχικό δίκτυο. ..... - 131 -
Εικόνα 29. Στρατηγική κατασκευής δικτύου από τα δεδομένα με score based αλγόριθμους και εκκίνηση με το δίκτυο των Marvin *et al*. (2017). ..... - 132 -
Εικόνα 30. Το μπεϋζιανό δίκτυο της μεθόδου με τις καλύτερες εκτιμήσεις δεδομένων. - 139 -
Εικόνα 31. Ο πίνακας σύγχυσης της μεθόδου με τις καλύτερες εκτιμήσεις δεδομένων. - 140 -
Εικόνα 32. Η δομή του μπεϋζιανού δικτύου με εφαρμογή της βέλτιστης στρατηγικής και χρήση ίδιων δεδομένων εκπαίδευσης με τους Marvin et al. (2017). ..... - 146 -
Εικόνα 33: Ο πίνακας σύγχυσης που προέκυψε με εφαρμογή της βέλτιστης στρατηγικής και χρήση ίδιων δεδομένων εκπαίδευσης με τους Marvin et al. (2017). ... - 147 -
Εικόνα 34. Δομή νέου αρχικού δικτύου με τη μεταβλητή NM Hazard ως παιδί... - 151 -
Εικόνα 35. Η τελική δομή δικτύου όταν το δίκτυο εκκίνησης αυτό που είχε την NM Hazard ως παιδί. ..... - 152 -
Εικόνα 36. Ο πίνακας σύγχυσης για τη βέλτιστη στρατηγική με δίκτυο εκκίνησης αυτό που είχε την NM Hazard ως παιδί. ..... - 153 -

# Κατάλογος Πινάκων

Πίνακας 1. Πίνακας CPT για τον κόμβο $X1$του παραδείγματος $σύνδεσης$ $X1 \rightarrow X2 \rightarrow X3$. ..... - 47 -
Πίνακας 2. Πίνακας CPT για τον κόμβο $X2$του παραδείγματος $σύνδεσης$ $X1 \rightarrow X2 \rightarrow X3$. ..... - 47 -
Πίνακας 3. Πίνακας CPT για τον κόμβο $X3$του παραδείγματος $σύνδεσης$ $X1 \rightarrow X2 \rightarrow X3$. ..... - 47 -
Πίνακας 4. Πίνακας CPT για τον κόμβο $X1$του παραδείγματος $σύνδεσης$ $X1 \rightarrow X2 \rightarrow X3$. ..... - 51 -
Πίνακας 5. . Πίνακας CPT για τον κόμβο $X2$του παραδείγματος $σύνδεσης$ $X1 \rightarrow X2 \rightarrow X3$. ..... - 51 -
Πίνακας 6. . Πίνακας CPT για τον κόμβο $X3$του παραδείγματος $σύνδεσης$ $X1 \rightarrow X2 \rightarrow X3$. ..... - 51 -
Πίνακας 7. Παρουσίαση της λογικής δεδομένων πολλών μεταβλητών και πολλών στιγμιοτύπων. ..... - 56 -
Πίνακας 8. Εννοιολογική παρουσίαση CPT. ..... - 59 -
Πίνακας 9. Πίνακας στιγμιοτύπων του παραδείγματος της Εικόνας 18. ..... - 62 -
Πίνακας 10. Πίνακας CPT για τον κόμβο $X1$του παραδείγματος της Εικόνας 18. . - 63 -
Πίνακας 11. Πίνακας CPT για τον κόμβο $X2$του παραδείγματος της Εικόνας 18. . - 63 -
Πίνακας 12. Πίνακας CPT για τον κόμβο $X3$του παραδείγματος της Εικόνας 18. . - 63 -
Πίνακας 13. Συμπληρωμένος πίνακας CPT για τον κόμβο $X1$του παραδείγματος της Εικόνας 18. ..... - 67 -
Πίνακας 14. Συμπληρωμένος πίνακας CPT για τον κόμβο $X2$του παραδείγματος της Εικόνας 18. ..... - 68 -
Πίνακας 15. Συμπληρωμένος πίνακας CPT για τον κόμβο $X3$του παραδείγματος της Εικόνας 18. ..... - 68 -
Πίνακας 16. Παράδειγμα μη πλήρους πίνακα δεδομένων. ..... - 88 -
Πίνακας 17. Αρχικοποίηση πίνακα CPT για τον κόμβο $X1$. ..... - 88 -
Πίνακας 18. Αρχικοποίηση πίνακα CPT για τον κόμβο $X2$ ..... - 89 -
Πίνακας 19. Αρχικοποίηση πίνακα CPT για τον κόμβο $X3$ ..... - 89 -
Πίνακας 20. Πίνακας ψευδοστιγμιοτύπων πρώτης επανάληψης. ..... - 92 -
Πίνακας 21. Μεταβλητές εισόδου στο μπεϋζιανό δίκτυο. ..... - 122 -

Πίνακας 22. Μεταβλητές εξόδου στο μπεϋζιανό δίκτυο. ..... - 124 -
Πίνακας 23. Παράδειγμα πίνακα σύγχυσης για πρόβλημα κατηγοριοποίησης δύο κλάσεων(binary classification)..... - 133 -
Πίνακας 24. Αποτελέσματα των accuracy και MCC για την 1η περίπτωση. ..... - 140 -
Πίνακας 25. Αποτελέσματα των accuracy και MCC για την 2η περίπτωση. ..... - 142 -
Πίνακας 26. Αποτελέσματα των accuracy και MCC για την 3η περίπτωση. ..... - 143 -
Πίνακας 27. Αποτελέσματα των accuracy και MCC για την 4η περίπτωση. ..... - 144 -
Πίνακας 28. Αναγωγή του προβλήματος κατηγοριοποίησης τεσσάρων κλάσεων σε πρόβλημα δύο κλάσεων. ..... - 149 -

# 1 Εισαγωγή

Η σημερινή εποχή χαρακτηρίζεται από την η 3η Βιομηχανική Επανάσταση που βρίσκεται ακόμα σε εξέλιξη, είναι γνωστή ως «η Εποχή της Πληροφορίας» και αποτελεί συνέπεια της ψηφιακής επανάστασης που ξεκίνησε στα τέλη του 20ου αιώνα. Όμως, χωρίς ακόμα να έχει παρέλθει η εποχή αυτή, γίνεται ήδη λόγος για την 4η Βιομηχανική Επανάσταση, με την επιστήμη των δεδομένων (Data Science) και την τεχνητή νοημοσύνη (Artificial Intelligence) να βρίσκονται στο επίκεντρό της. Στην τομή αυτών των δύο πεδίων βρίσκεται η μηχανική μάθηση (Machine Learning), βασικό αντικείμενο της οποίας είναι η κατασκευή αλγορίθμων που μαθαίνουν μέσα από τα δεδομένα και στη συνέχεια λαμβάνουν βέλτιστες αποφάσεις ή κάνουν προβλέψεις. Η μηχανική μάθηση έχει κεντρικό ρόλο στην εξέλιξη της 4ης Βιομηχανικής Επανάστασης καθώς παρουσιάζει πολλές εφαρμογές, όπως για παράδειγμα στην αξιολόγηση ρίσκου, στις εφαρμογές των μέσων κοινωνικής δικτύωσης, στην παραγωγή αυτόματων οχημάτων, στις μηχανές αναζήτησης, στη ρομποτική και σε πολλούς άλλους τομείς.

Αναλυτικότερα, μηχανική μάθηση είναι μια αυτοματοποιημένη διαδικασία κατά την οποία εξάγονται διάφορες σχέσεις και μοτίβα από ένα σύνολο δεδομένων (data set). Αφού γίνει η συλλογή του συνόλου των δεδομένων, με αλγοριθμική διαδικασία χτίζεται ένα (στατιστικό) μοντέλο το οποίο ενσωματώνει τις σχέσεις και τα μοτίβα που υπάρχουν στο σύνολο των δεδομένων. Η αυτοματοποιημένη διαδικασία εύρεσης των σχέσων και των μοτίβων στα δεδομένα και η επιλογή του στατιστικού μοντέλου που αναπαριστά καλύτερα αυτές τις σχέσεις είναι η διαδικασία της μάθησης που επιτελεί η μηχανή.

Η μηχανική μάθηση χωρίζεται σε τρεις μεγάλες κατηγορίες οι οποίες είναι η μάθηση υπό επιτήρηση (supervised learning), η μάθηση χωρίς επιτήρηση (unsupervised learning) και η ενισχυτική μάθηση (reinforcement learning).

Στο supervised learning τα δεδομένα αποτελούνται από ένα σύνολο παραδειγμάτων ή στιγμιοτύπων (instances) και αφορούν κάποια χαρακτηριστικά (features) τα οποία σχετίζονται με μία ή περισσότερες ετικέτες (label). Για παράδειγμα, αν τα στιγμιότυπα αντιστοιχούν σε διαφορετικούς ανθρώπους, τα χαρακτηριστικά θα μπορούσαν να είναι το ύψος, το βάρος και το χρώμα των ματιών και η ετικέτα το όνομα του ανθρώπου με

τον οποίο σχετίζονται τα χαρακτηριστικά αυτά. Τα χαρακτηριστικά μπορούν να μετρώνται με συνεχή τρόπο, όπως το ύψος και το βάρος, ή με διακριτό τρόπο, όπως το χρώμα των ματιών. Όταν ένα χαρακτηριστικό μετριέται με συνεχή τρόπο, το πλήθος των τιμών του είναι άπειρο, ενώ όταν μετριέται με διακριτό τρόπο οι τιμές του έχουν πεπερασμένο πλήθος. Τα χαρακτηριστικά ή αλλιώς οι μεταβλητές των δεδομένων αποκαλούνται συνήθως επεξηγηματικές μεταβλητές (explanatory variables) ή μεταβλητές εισόδου και οι ετικέτες του κάθε στιγμιότυπου ονομάζονται μεταβλητές απόκρισης (response variables). Ένας supervised learning αλγόριθμος χρησιμοποιεί το σύνολο των δεδομένων ώστε να παράξει ένα στατιστικό μοντέλο το οποίο θα λαμβάνει ώς είσοδο ένα στιγμιότυπο των επεξηγηματικών μεταβλητών και θα δίνει ως έξοδο την προβλεπόμενη κατάσταση της μεταβλητής απόκρισης.Με άλλα λόγια, οι αλγόριθμοι που ανήκουν στην κατήγορία του supervised learning προσπαθούν να εγκαταστήσουν μία σχέση μεταξύ του χώρου που ορίζουν οι επεξηγηματικές μεταβλητές και αυτού που ορίζουν οι μεταβλητές απόκρισης. Το supervised learning χωρίζεται σε δύο γενικές περιπτώσεις ανάλογα με τον τύπο της μεταβλητής απόκρισης. Όταν η μεταβλητή απόκρισης είναι συνεχής έχουμε την περίπτωση της παλινδρόμησης (regression), ενώ όταν λαμβάνει διακριτές τιμές από πεπερασμένες κατηγορίες έχουμε την περίπτωση της κατηγοριοποίησης (classification).

Σε αντίθεση με το supervised learning, στο unsupervised learning το σύνολο των δεδομένων δεν έχει χαρακτηριστικά με ετικέτες. Ένα αλγόριθμος unsupervised learning βρίσκει σχέσεις και μοτίβα που μπορεί να υπάρχουν στα δεδομένα, χωρίς να έχει γίνει υπόδειξη των αντίστοιχων ετικετών για κάθε στιγμιότυπο. Μία μεγάλη κατηγορία unsupervised learning διαδικασιών είναι η ομαδοποίηση (clustering), κατά την οποία μέσα από ένα σύνολο δεδομένων χωρίς ετικέτες ανακαλύπτονται μοτίβα ικανά να χωρίσουν τα δεδομένα σε ομάδες με κοινά χαρακτηριστικά.

Τέλος, στην κατηγορία του reinforcement learning, ο αλγόριθμος μαθαίνει τα δεδομένα μέσα από διάφορες δράσεις που πραγματοποιεί στο περιβάλλον του. Διαφορετικές δράσεις επιφέρουν διαφορετικό αποτέλεσμα, που κάποιες φορές έχει τη μορφή τιμωρίας (penalty) και άλλες φορές τη μορφή επιβράβευσης (reward). Μέσα από αυτές τις δράσεις σε μια διαδικασία δοκιμής και σφάλματος ο αλγόριθμος μαθαίνει τα δεδομένα για το περιβάλλον του. Η επιβράβευση λειτουργεί ως ανάδραση που δίνει την κατάλληλη πληροφορία για την επιλογή της νέας δράσης και κάθε νέα δράση επιλέγεται ως εκείνη που μεγιστοποιεί την επιβράβευση. Συνήθεις εφαρμογές

του reinforcement learning βρίσκονται στην αυτοματοποίηση των οχημάτων, στη ρομποτική κτλ. (Keller *et al.*, 2015).

Στο πλαίσιο της παρούσας διπλωματικής θα αναλυθούν τα μπεϋζιανά δίκτυα, τα οποία εντάσσονται στην περιοχή του supervised learning, κι έτσι κρίνεται σκόπιμη μία εκτενέστερη αναφορά σε αυτή την κλάση αλγορίθμων μηχανικής μάθησης. Γενικά, στο supervised learning το σύνολο των δεδομένων χωρίζεται σε δύο μέρη, με το πρώτο να είναι το σύνολο εκπαίδευσης (training set) και το δέυτερο να είναι το σύνολο εξέτασης (test set). Το training set είναι το σύνολο που χρησιμοποιείται για την εκπαίδευση του μοντέλου, ενώ το test set χρησιμοποιείται για να εξεταστεί η προβλεπτική ικανότητα του μοντέλου σε νέα δεδομένα, των οποίων η πληροφορία δεν έχει χρησιμοποιηθεί από το μοντέλο.Ένα σύνηθες πρόβλημα στη διαδικασία supervised learning προκύπτει όταν κατά τη διαδικασία της μάθησης το μοντέλο μπορεί να προβλέπει επιτυχώς τις καταστάσεις των μεταβλητών απόκρισης του training set αλλά δεν γενικεύει, δηλαδή προβλέπει με πολύ μικρή ακρίβεια τις μεταβλητές απόκρισης νέων στιγμιοτύπων. Η κατάσταση αυτή αποκαλείται υπερβολική προσαρμογή (overfitting) του μοντέλου στα δεδομένα εκπαίδευσης. Μια λύση αποτελεί η αύξηση του πλήθους των δεδομένων του training set, ώστε να καλύπτεται μεγαλύτερος χώρος . Κάτι τέτοιο όμως δεν είναι πάντα εφικτό, καθώς σε πολλές περιπτώσεις παρατηρείται έλλειψη δεδομένων για ένα υπό μελέτη πρόβλημα. Στη συγκεκριμένη εργασία, για να αποφευχθεί όσο είναι δυνατόν αυτό χρησιμοποιείται η μέθοδος της διασταυρούμενης αξιολόγησης (cross validation), η οποία θα αναλυθεί στο κεφάλαιο 5 (Witten *et al.*, 2011).

Τα μπεϋζιανά δίκτυα χωρίζονται στα συνεχή δίκτυα, όταν το σύνολο των δεδομένων μετριέται με συνεχή τροπο, και στα διακριτά δίκτυα, όταν το σύνολο των δεδομένων μετριέται με διακριτό τρόπο. Τα διακριτά μπεϋζιανά δίκτυα χρησιμοποιούνται πολύ συχνά σε προβλήματα κατηγοριοποίησης. Η παρούσα διπλωματική αποτελεί μια προσπάθεια περιγραφής και ανάλυσης των βασικών εννοιών των μπεϋζιανών δικτύων, με έμφαση στα διακριτά δίκτυα. Μετά την αναλυτική περιγραφή των μπεϋζιανών δικτύων παρουσιάζεται μία εφαρμογή τους. Πιο συγκεκριμένα, αναλύεται η χρήση μπεϋζιανών δικτύων στην πρόβλεψη κινδύνου σε νανοϋλικά.

Αρχικά, παρουσιάζονται βασικά στοιχεία των Πιθανοτήτων και Στατιστικής που θα χρειαστούν για την ανάλυση των μπεϋζιανών δικτύων. Γίνεται αναφορά στις διαφορετικές προσεγγίσεις της έννοιας της πιθανότητας, στην έννοια της στοχαστικής

ανεξαρτησίας καθώς και στο θεώρημα του Bayes. Επιπλέον, αναφέρονται βασικά στοιχεία που αφορούν τις μονοδιάστατες και πολυδιάστατες τυχαίες μεταβλητές όπως η στοχαστική ανεξαστησία και η υπό συνθήκη ανεξαρτησία τυχαίων μεταβλητών, που θεωρούνται χρήσιμα για την ομαλή εισαγωγή του αναγνώστη στα θέματα που αναλύονται σε επόμενα κεφάλαια.

Στη συνέχεια, αναλύονται κάποιες βασικές έννοιες της θεωρίας των γραφημάτων που χρειάζονται για τον ορισμό και την κατανόηση των μπεϋζιανών δικτύων. Συγκεκριμένα, γίνεται ειδική αναφορά στα κατευθυνόμενα ακυκλικά γραφήματα (directed acyclic graphs), στις βασικές συνδέσεις μεταξύ των κόμβων σε ένα κατευθυνόμενο ακυκλικό γράφημα, στο κριτήριο του διαχωρισμού κατεύθυνσης (direction separation) μεταξύ των κόμβων καθώς και στην κουβέρτα του Markov (Markov blanket).

Στο επόμενο κεφάλαιο γίνεται η βασική ανάλυση των μπεϋζιανών δικτύων ξεκινώντας με τον ορισμό των δικτύων. Έπειτα, για την περίπτωση συνόλων δεδομένων που είναι πλήρη, πραγματοποιείται η ανάλυση δύο μεθόδων εύρεσης παραμέτρων με πρώτη τη μέθοδο εκτίμησης μέγιστης πιθανοφάνειας (ΕΜΠ) και δεύτερη τη μπεϋζιανή προσέγγιση εύρεσης παραμέτρων. Ύστερα γίνεται αναφορά στις μεθόδους εκτίμησης στην περίπτωση παρουσιάσης νέων δεδομένων που δεν εμφανίστηκαν στο training set. Επιπλέον, αναλύεται ο αλγόριθμος EM (Expectation Maximization) μέσω του οποίου βρίσκονται οι παράμετροι του στατιστικού μοντέλου στην περίπτωση που το σύνολο των δεδομένων δεν είναι πλήρες. Στη συνέχεια αναφέρονται μέθοδοι για την εύρεση της δομής του δικτύου από τα δεδομένα καθώς και ο δομικός αλγόριθμος EM (structural EM) που βρίσκει τη δομή του δικτύου με βάση μη πλήρες σύνολο δεδομένων. Το κεφάλαιο κλείνει με αναφορά στους τρόπους πρόβλεψης με μπεϋζιανά δίκτυα.

Ακολούθως, γίνεται εφαρμογή των προηγούμενων μεθόδων με βάση συγκεκριμένη έρευνα και συγκεκριμένο σύνολο δεδομένων που αφορούν την δυνατότητα πρόβλεψης της επικινδυνότητας που μπορεί να παρουσιάσει ένα. Στο κεφάλαιο αυτό παρουσιάζονται τα στοιχεία της έρευνας προς μελέτη ενώ στη συνέχεια αναλύεται η μεθοδολογία που ακολουθήθηκε στην παρούσα διπλωματική για την πρόβλεψη της επικινδυνότητας των νανοϋλικών. Έπειτα παρουσιάζονται τα αποτελέσματα της εφαρμογής καθώς και ο σχολιασμός των αποτελεσμάτων.

Στο τελευταίο κεφάλαιο γίνεται μια γενική αποτίμηση των αποτελεσμάτων και αναφέρονται τα τελικά συμπεράσματα της διπλωματικής.

# 2 Στοιχεία Πιθανοτήτων

Για την πληρέστερη κατανόηση των μεθόδων που θα αναλυθούν στα επόμενα κεφάλαια, στο κεφάλαιο αυτό θα παρουσιαστούν βασικά στοιχεία πιθανοτήτων. Η ανάλυση των εννοιών βασίζεται σε μεγάλο βαθμό στους Κοκολάκη και Σπηλιώτη (2002) καθώς και Χαραλαμπίδη (2009). Αρχικά, παρουσιάζονται οι διαφορετικές προσεγγίσεις της έννοιας της πιθανότητας και η αξιωματική θεμελίωση της πιθανότητας. Στη συνέχεια, αναφέρονται βασικά στοιχεία που αφορούν τη δεσμευμένη πιθανότητα, το θεώρημα του Bayes και την ανεξαρτησία ενδεχομένων. Τέλος, αναλύονται οι έννοιες της τυχαίας μεταβλητής, των πολυδιάστατων τυχαίων μεταβλητών καθώς και της μέσης τιμής.

## 2.1 Βασικές Έννοιες Πιθανοτήτων

Βασικό ρόλο στην ανάπτυξη του ορισμού της πιθανότητας διαδραματίζει η έννοια του πειράματος τύχης, αφού τα τυχερά παιχνίδια είναι εκείνα που αποτέλεσαν τις πρώτες εφαρμογές των πιθανοτήτων. Ένα πείραμα καλείται πείραμα τύχης όταν εκτελείται κάτω από τις ίδιες συνθήκες αλλά μπορεί να δώσει διαφορετικά αποτελέσματα. Γενικότερα, η έννοια της πιθανότητας συνδέεται με την αβεβαιότητας του αποτελέσματος ενός πειράματος ή φαινομένου.

Το σύνολο των δυνατών αποτελεσμάτων ενός πειράματος ή φαινομένου καλείται δειγματικός χώρος (δ.χ.) και συμβολίζεται με Ω. Ένα στοιχείο ω του δειγματικού χώρου Ω ονομάζεται δειγματικό σημείο και κάθε υποσύνολο του Ω ονομάζεται ενδεχόμενο. Ένα ενδεχόμενο $A \in \Omega$ λέγεται ότι πραγματοποιείται όταν το αποτέλεσμα ενός πειράματος ή φαινομένου είναι στοιχείο του $A$.

Επιπλέον, δύο ενδεχόμενα $A, B$ καλούνται ξένα ή ασυμβίβαστα όταν $A \cap B = \emptyset$ και γενικότερα τα ενδεχόμενα $A_1, A_2, \ldots, A_n$ καλούνται ανά δύο ξένα ή ασυμβίβαστα όταν $A_i \cap A_j = \emptyset$ για κάθε $i \neq j$ με $i, j = 1, \ldots, n$ (Κοκολάκης και Σπηλιώτης, 2002; Χαραλαμπίδης, 2009).

### *2.1.1 Στατιστική Πιθανότητα*

Ας υποτεθεί ότι ένα πείραμα με δειγματικό χώρο Ω, μπορεί να επαναληφθεί άπειρες φορές κάτω από τις ίδιες συνθήκες. Επιπλέον, έστω ένα ενδεχόμενο $A \in \Omega$, το οποίο πραγματοποιείται $f_n(A)$ φορές σε n το πλήθος επαναλήψεις του πειράματος. Τότε, αν υπάρχει το όριο της σχετικής συχνότητας:

$$\lim_{n\to\infty} p_n(A) = \lim_{n\to\infty} \frac{f_n(A)}{n}$$

αποτελεί την τιμή της πιθανότητας του ενδεχομένου Α. Δηλαδή:

$$P(A) = \lim_{n\to\infty} \frac{f_n(A)}{n}$$

Ωστόσο, η ύπαρξη του ορίου δεν εξασφαλίζεται πάντα και αυτός είναι ο λόγος που η σύγκλιση μπορεί να διαπιστωθεί εμπειρικά και μέχρι ενός βαθμού, αφού το πλήθος

εμφάνισης του ενδεχομένου Α είναι γνωστό μόνο μετά από την επανάληψη του πειράματος n φορές.

### *2.1.2 Υποκειμενική Πιθανότητα*

Μια διαφορετική προσέγγιση της έννοιας της πιθανότητας, είναι εκείνη που βασίζεται στην άποψη ότι η πιθανότητα να πραγματοποιηθεί ένα ενδεχόμενο συνδέεται με το προσωπικό βαθμό βεβαιότητας ή αβεβαιότητας που έχουμε ως προς την εμφάνιση του ενδεχομένου. Συνεπώς, δυο διαφορετικοί παρατηρητές ενός φαινομένου μπορούν να δώσουν διαφορετική πιθανότητα στην εμφάνιση του, βασιζόμενοι στις γνώσεις τους , την εμπειρία τους και γενικότερα στην πεποίθησή τους. Για το λόγο αυτό η συγκεριμένη πιθανότητα καλείται υποκειμενική πιθανότητα.

### *2.1.3 Αξιωματικός Ορισμός Πιθανότητας*

Έστω Ω ο δειγματικός χώρος, τότε το F αποτελεί ένα σ-πεδίο του Ω όταν ικανοποιούνται τα παρακάτω:

α) $\Omega \in F$

β) αν $A \in F$ τότε $A' = \Omega - A \in F$

γ) αν $A_i \in F$ τότε $\bigcup_{i=1}^{\infty} A_i \in F$ όπου $i = 1,2, \ldots$

Τα στοιχεία του F ονομάζονται ενδεχόμενα ή μετρήσιμα σύνολα και το ζεύγος (Ω,F) ονομάζεται μετρήσιμος χώρος.

Ένα υποσύνολο του Ω ονομάζεται ενδεχόμενο μόνο εάν ανήκει στο F και η πιθανότητα αφορά τα ενδεχόμενα-υποσύνολα του Ω που είναι στοιχεία του F.

Έστω ο μετρήσιμος χώρος (Ω,F) και P μια συνολοσυνάρτηση με $P: F \to \mathbb{R}$ . Η $P$ ονομάζεται μέτρο πιθανότητας αν ικανοποιεί τα παρακάτω τρία αξιώματα:

α) $P(A) \geq 0$ , $A \in F$

β) αν $A_i \in F$ με $i = 1,2, \ldots$ και $A_i \cap A_j = \emptyset \; \forall \; i \neq j$ (ανά δύο ξένα ενδεχόμενα) τότε

$$P(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} P(A_i)$$ (αριθμήσιμη αθροιστικότητα)

γ) $P(\Omega) = 1$

Η τριάδα (Ω,F,P) ονομάζεται χώρος πιθανότητας.

### *2.1.4 Δεσμευμένη Πιθανότητα και Αποτελέσματα*

Έστω (Ω,F,P) χώρος πιθανότητας και A, B ∈ F ενδεχόμενα, με P(B)>0. Τότε η δεσμευμένη πιθανότητα του A δεδομένου του B ορίζεται ως εξής:

$$P(A|B) = \frac{P(AB)}{P(B)}$$

Δηλαδή, η πιθανότητα να συμβεί το A όταν γνωρίζουμε ότι έχει συμβεί το B ή γνωρίζουμε ότι θα συμβεί το B. Η P(A|B) ικανοποιεί τα αξιώματα του μέτρου πιθανότητας.

*Πολλαπλασιαστικός τύπος*

Άμεση συνέπεια της δεσμευμένης πιθανότητας είναι ο πολλαπλασιαστικός τύπος όπου για ενδεχόμενα A, B ισχύει ότι:

$P(AB) = P(A|B) * P(B)$ όταν το $P(B) > 0$

ή

$P(AB) = P(B|A) * P(A)$ όταν το $P(A) > 0$

Επιπλέον, ο πολλαπλασιαστικός τύπος γενικεύεται ως εξής:
Έστω $A_i \in F, i = 1,2,\dots n$ ενδεχόμενα με $P(A_1A_2 \dots A_{n-1}) > 0$ τότε:

$$P(A_1A_2 \dots A_n) = P(A_1)P(A_2|A_1)P(A_3|A_1A_2) \dots P(A_n|A_1A_2 \dots A_{n-1})$$

*Θεώρημα ολικής πιθανότητας*

Σημαντικό αποτέλεσμα της δεσμευμένης πιθανότητας είναι το θεώρημα ολικής πιθανότητας που διατυπώνεται ως εξής:
Εάν $B_1, B_2, \dots$ είναι μια πεπερασμένη ή άπειρη ακολουθία ξένων ενδεχομένων με $\bigcup_{i=1}^{\infty} B_i = \Omega$ και $P(B_i) > 0 \ \ (i = 1,2,\dots)$, τότε για κάθε ενδεχόμενο A ισχύει:

$$P(A) = \sum_{i=1}^{\infty} P(A|\, B_i)P(B_i)$$

### *2.1.5 Θεώρημα του Bayes*

Εάν $B_1, B_2, \ldots$ είναι μια πεπερασμένη ή άπειρη ακολουθία ξένων ενδεχομένων με $\cup_{i=1}^{\infty} B_i = \Omega$, $P(B_i) > 0$ $(i = 1,2, \ldots)$, και Α ένα ενδεχόμενο με $P(A) > 0$ τότε:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{\sum_{i=1}^{\infty} P(A|B_i)P(B_i)} \quad k = 1,2, \ldots$$

Από το θεώρημα ολικής πιθανότητας το $P(A) = \sum_{i=1}^{\infty} P(A|\, B_i)P(B_i)$ άρα το θεώρημα του Bayes μπορεί να γραφτεί ως εξής:

$$P(B_k|A) = \frac{P(A|B_k)P(B_k)}{P(A)}$$

### *2.1.6 Ανεξαρτησία Ενδεχομένων*

Έστω Α, Β δύο ενδεχόμενα ενός δειγματικού χώρου Ω και τα Α, Β είναι ανεξάρτητα κάτι που παριστάνεται ως εξής: $A \perp\!\!\!\perp B$.

Τότε έχουμε:

$$A \perp\!\!\!\perp B \Leftrightarrow P(AB) = P(A) * P(B)$$

ισοδύναμα

$$A \perp\!\!\!\perp B \Leftrightarrow P(A|B) = P(A)$$

Δηλαδή, η πραγματοποίηση του ενδεχομένου Β δεν έχει επίδραση στη πραγματοποίηση του Α.

Επειδή η ανεξαρτησία ενδεχομένων είναι συμμετρική θα ισχύει επίσης ότι $P(B|A) = P(B)$.

*Ανά δύο ανεξάρτητα*

Τα ενδεχόμενα $A_1, A_2, \ldots, A_n$ καλούνται ανά δύο ανεξάρτητα όταν:

$$P(A_iA_j) = P(A_i)P(A_j) \ \forall\, i \neq j \text{ με } i,j = 1,2,\dots n$$

*Πλήρως ανεξάρτητα*

Τα ενδεχόμενα $A_1, A_2, \dots, A_n$ καλούνται πλήρως ανεξάρτητα όταν για κάθε σύνολο δεικτών $J \subseteq \{1,2,\dots,n\}$ ισχύει:

$$P(\bigcap_{j\in J} A_j) = \prod_{j\in J} P(A_j)$$

Ενώ για όλα τα ενδεχόμενα θα ισχύει:

$$P(\bigcap_{j=1}^{n} A_j) = \prod_{j=1}^{n} P(A_j) = P(A_1)P(A_2)\cdots P(A_n)$$

*Ανεξαρτησία ενδεχομένων υπό συνθήκη*

Επέκταση της έννοιας της ανεξαρτησίας ενδεχομένων αποτελεί η ανεξαρτησία ενδεχομένων υπό συνθήκη. Αυτό σημαίνει πως υπό τη συνθήκη ότι έχει πραγματοποιηθεί ένα ενδεχόμενο, σταματάει η στοχαστική εξάρτηση μεταξύ άλλων ενδεχομένων.

Έστω ότι έχουμε A,B,C ενδεχόμενα ενός δειγματικού χώρου Ω και τα A, B είναι ανεξάρτητα υπό τη συνθήκη οτι εχει πραγματοποιηθεί το C κάτι που συμβολίζεται ως $(A \perp\!\!\!\perp B)|C$.

Τότε ισχύουν τα εξής:

$$(A \perp\!\!\!\perp B)|C \ \Leftrightarrow P(AB|C) = P(A|C) * P(B|C)$$

ισοδύναμα

$$(A \perp\!\!\!\perp B)|C \ \Leftrightarrow P(A|BC) = P(A|C)$$

και

$$P(B|AC) = P(B|C)$$

λόγω του οτι η ανεξαρτησία είναι συμμετρική.

Δηλαδή αν τα ενδεχόμενα A,B είναι ανεξάρτητα υπό τη συνθήκη ότι έχει πραγματοποιηθεί το C, σημαίνει ότι η πραγματοποίηση του B δεδομένου ότι έχει πραγματοποιηθεί το C δεν επηρεάζει την πιθανότητα πργαματοποίησης του A για αυτό τον λόγο η πιθανότητα $P(A|BC) = P(A|C)$.

## 2.2 Τυχαίες Μεταβλητές και Ιδιότητες τους

### *2.2.1 Τυχαία Μεταβλητή*

Έστω (Ω,F,P) ένας χώρος πιθανότητας. Μια συνάρτηση $X : \Omega \longrightarrow \mathbb{R}$ ονομάζεται τυχαία μεταβλητή (τ.μ.) αν και μόνο εάν το σύνολο:

$$\{\omega \in \Omega: X(\omega) \leq x\} \in F, \qquad \forall x \in \mathbb{R}$$

Η τυχαία μεταβλητή αντιστοιχεί σε κάθε στοιχείο ω ∈ Ω έναν πραγματικό αριθμό $x = X(\omega)$.

### *2.2.2 Συνάρτηση Κατανομής Πιθανότητας*

Έστω (Ω,F,P) ένας χώρος πιθανότητας και X μια τυχαία μεταβλητή ορισμένη στο Ω. Η συνάρτηση F η οποία ορίζεται από τη σχέση :

$$F(x) = P(X \leq x) = P(\{\omega \in \Omega: X(\omega) \leq x\}), x \in \mathbb{R}$$

καλείται συνάρτηση κατανομής πιθανότητας (σ.κ.π) της τ.μ. X.

Μια συνάρτηση F είναι συνάρτηση κατανομής πιθανότητας μιας τ.μ. X αν και μόνο αν:

$i)$ $0 \leq F(x) \leq 1$ , $x \in \mathbb{R}$ (λαμβάνει τιμές στο [0,1])

$ii)$ $F(x_1) \leq F(x_2)$ , $x_1, x_2 \in \mathbb{R}$ με $x_1 < x_2$ (αύξουσα)

$iii)$ $\lim_{x \to x_0^+} F(x) = F(x_0)$ $x_0 \in \mathbb{R}$ (δεξιά συνεχής)

$iv)$ $\lim_{x \to -\infty} F(x) = F(-\infty) = 0$ και $\lim_{x \to +\infty} F(x) = F(+\infty) = 1$

### *2.2.3 Διακριτή Τυχαία Μεταβλητή και Συνάρτηση Μάζας Πιθανότητας*

Μια τυχαία μεταβλητή X καλείται διακριτή αν παίρνει με πιθανότητα 1 αριθμήσιμο (πεπερασμένο ή άπειρο) σύνολο τιμών $R_X = \{x_0, x_1, \dots, x_n, \dots\}$.

Δηλαδή, η $P(X \in R_X) = 1$.

Η συνάρτηση $f(x_k) = P(X = x_k) = P(\{\omega \in \Omega : X(\omega) = x_k\}), k = 0,1,\dots,n,\dots$ η οποία εκχωρεί σε κάθε σημείο $x_k$ την πιθανότητα του καλείται συνάρτηση μάζας πιθανότητας (σ.μ.π.) ή συνάρτηση πιθανότητας (σ.π.) της τυχαίας μεταβλητής X.

Η συνάρτηση f είναι σ.μ.π. αν και μόνο αν:

$i) f(x_k) \geq 0, \forall k \in \mathbb{N}$

$ii) \sum_{k=1}^{\infty} f(x_k) = 1$

Οι σχέσεις που συνδέουν τη συνάρτηση κατανομής πιθανότητας και τη συνάρτηση μάζας πιθανότητας είναι:

$i) F(x) = \sum_{x_k \leq x} f(x_k)\ , x \in \mathbb{R}$

$ii) f(x_k) = F(x_k) - F(x_k^-)$

### *2.2.4 Συνεχής Τυχαία Μεταβλητή και Συνάρτηση Πυκνότητας Πιθανότητας*

Μια τυχαία μεταβλητή X καλείται συνεχής αν υπάρχει μια πραγματική συνάρτηση f με πεδίο ορισμού το $\mathbb{R}$ τέτοια ώστε:

$f(x) \geq 0, \ \ \forall x \in \mathbb{R}$

και

$F(x) = P(X \leq x) = \int_{-\infty}^{x} f(t)\, dt\ ,\ \forall x \in \mathbb{R}$

όπου η συνάρτηση f καλείται συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) της τυχαίας μεταβλητής X.

Το βασικό θεώρημα για να καταλάβουμε αν μια συνάρτηση f είναι συνάρτηση πυκνότητας πιθανότητας είναι το εξής:

Η συνάρτηση f είναι σ.π.π. αν και μόνο αν

$i)\ f(x) \geq 0,\ \ \forall x \in \mathbb{R}$

$ii)\ \int_{-\infty}^{+\infty} f(x)\, dx = 1$

Οι σχέσεις που συνδέουν τη συνάρτηση κατανομής πιθανότητας και τη συνάρτηση πυκνότητας πιθανότητας είναι:

$i)\ F(x) = \int_{-\infty}^{x} f(t)\, dt\ ,\ \forall x \in \mathbb{R}$

$ii)\ f(x) = \frac{dF(x)}{dx}$

## 2.3 Πολυδιάστατες Τυχαίες Μεταβλητές, Κατανομές και Ιδιότητές τους

Έστω ένας χώρος πιθανότητας (Ω,F,P). Ένα ζεύγος συναρτήσεων $(X,Y):\Omega \to \mathbb{R}^2$ το οποίο σε κάθε ω∈Ω αντιστοιχεί ένα ζεύγος πραγματικών αριθμών $(X(\omega),Y(\omega)) = (x,y)$ καλείται δισδιάστατη τυχαία μεταβλητή αν και μόνο αν οι X,Y είναι τυχαίες μεταβλητές, δηλαδή αν το σύνολο

$$\{\omega \in \Omega: X(\omega) \le x, Y(\omega) \le y\} \in F, \ \ \forall\, x,y \in \mathbb{R}$$

Ομοίως, έστω ένας χώρος πιθανότητας (Ω,F,P). Μια διανυσματική συνάρτηση $\boldsymbol{X} = (X_1, X_2, \dots, X_n)^T : \Omega \to \mathbb{R}^n$ η οποία σε κάθε ω∈Ω αντιστοιχεί ένα διάνυσμα $(x_1, x_2, \dots, x_n)$ με $x_i = X_i(\omega),\ i = 1,2,\dots,n$ καλείται ν-διάστατη τυχαία μεταβλητή ή τυχαίο διάνυσμα αν και μόνο εάν οι $X_i: \Omega \to \mathbb{R}$ με $i = 1,2,\dots,n$ είναι τυχαίες μεταβλητές.

### *2.3.1 Συνάρτηση Κατανομής Πιθανότητας Τυχαίου Διανύσματος*

Η συνάρτηση: $F(x_1, x_2, \dots, x_n) = P(X_1 \le x_1, X_2 \le x_2, \dots, X_n \le x_n) = P(\{\omega \in \Omega: X_1(\omega) \le x_1, X_2(\omega) \le x_2, \dots, X_n(\omega) \le x_n\})$ με $(x_1, x_2, \dots, x_n)^T = \boldsymbol{x} \in \mathbb{R}^n$ ονομάζεται από κοινού συνάρτηση κατανομής πιθανότητας των τυχαίων μεταβλητών $X_1, X_2, \dots, X_n$.

### *2.3.2 Διακριτές Τυχαίες Μεταβλητές και η Από Κοινού Συνάρτηση Μάζας Πιθανότητας*

Μια δισδιάστατη τυχαία μεταβλητή (X,Y) καλείται διακριτή αν παίρνει με πιθανότητα 1 αριθμήσιμο (πεπερασμένο ή άπειρο) σύνολο τιμών $R_{X,Y}$. Δηλαδη, η $P((X,Y) \in R_{X,Y}) = 1$.

Η συνάρτηση $f(x_i, y_j) = P(X = x_i, Y = y_j) = P(\{\omega \in \Omega: X(\omega) = x_i, Y(\omega) = y_j\})$ για $i = 1,2,\dots,\ j = 1,2,\dots$ η οποία εκχωρεί σε κάθε σημείο $(x_i, y_j)$ την πιθανότητα του, καλείται από κοινού συνάρτηση μάζας πιθανότητας των τυχαίων μεταβλητών X,Y.

Μια συνάρτηση f είναι από κοινού συνάρτηση μάζας πιθανότητας των τυχαίων μεταβλητών X,Y αν και μόνο αν:

$i)\ f(x_i, y_j) \geq 0$ για $i = 1,2,\ldots j = 1,2,\ldots$

$$ii)\ \sum_{j=1}^{\infty}\sum_{i=1}^{\infty} f(x_i, y_j) = 1$$

Οι σχέσεις που συνδέουν την από κοινού κατανομή πιθανότητας και την από κοινού συνάρτηση μάζας πιθανότητας είναι οι εξής:

$$i)\ F(x,y) = \sum_{y_j \leq y}\sum_{x_i \leq x} f(x_i, y_j)\ \ x, y \in \mathbb{R}$$

$$ii)\ f(x_i, y_j) = F(x_i, y_j) - F(x_i^-, y_j) - F(x_i, y_j^-) + F(x_i^-, y_j^-)\ \ i, j = 0,1,\ldots$$

### *2.3.3 Συνεχείς Τυχαίες Μεταβλητές και η Από Κοινού Συνάρτηση Πυκνότητας Πιθανότητας*

Μια δισδιάστατη τυχαία μεταβλητή (X,Y) καλείται συνεχής αν υπάρχει μια πραγματική συνάρτηση f τέτοια ώστε:

$$i)\ f(x,y) \geq 0\ \ x, y \in \mathbb{R}$$

$$ii)\ F(x,y) = P(X \leq x, Y \leq y) = \int_{-\infty}^{y}\int_{-\infty}^{x} f(t,u)\, dt\, du$$

Η συνάρτηση f στην περίπτωση αυτή, καλείται από κοινού συνάρτηση πυκνότητας πιθανότητας των τυχαίων μεταβλητών X,Y.

Μια συνάρτηση f είναι από κοινού συνάρτηση πυκνότητας πιθανότητας των τυχαίων μεταβλητών X,Y αν και μόνο αν:

$$i)\ f(x,y) \geq 0,\ \ x, y \in \mathbb{R}$$

$$ii)\ \int_{-\infty}^{+\infty}\int_{-\infty}^{+\infty} f(x,y)\, dx\, dy = 1$$

Οι σχέσεις που συνδέουν την από κοινού κατανομή πιθανότητας και την από κοινού συνάρτηση πυκνότητας πιθανότητας είναι οι εξής:

$i)\ F(x,y) = \int_{-\infty}^{y} \int_{-\infty}^{x} f(t,u)\, dt\, du$

$ii)\ f(x,y) = \dfrac{\partial^2 F(x,y)}{\partial y \partial x}$

### *2.3.4 Περιθώριες Κατανομές*

Έστω f η από κοινού συνάρτηση μάζας πιθανότητας των διακριτών τυχαίων μεταβλητών X,Y. Τότε η συνάρτηση $f_X$ με :

$$f_X(x_i) = \sum_{j=0}^{\infty} f(x_i, y_j)\ , i = 0,1,\dots$$

καλείται περιθώρια συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής X.

Αντίστοιχα η συνάρτηση :

$$f_Y(y_j) = \sum_{i=0}^{\infty} f(x_i, y_j)\ , j = 0,1,\dots$$

καλείται περιθώρια συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής Y.

Έστω f η από κοινού συνάρτηση πυκνότητας πιθανότητας των συνεχών τυχαίων μεταβλητών X,Y. Τότε η συνάρτηση $f_X$ με :

$$f_X(x) = \int_{-\infty}^{\infty} f(x,y)\, dy\ , x \in \mathbb{R}$$

καλείται περιθώρια συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής X.

Αντίστοιχα η συνάρτηση :

$$f_Y(y) = \int_{-\infty}^{\infty} f(x,y)\, dx\ , y \in \mathbb{R}$$

καλείται περιθώρια συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Y.

### *2.3.5 Δεσμευμένες Κατανομές*

Έστω (X,Y) μια διακριτή δισδιάστατη τυχαία μεταβλητή με $f_{X,Y}(x_i, y_j),\ i,j = 0,1,\ldots$ την από κοινού συνάρτηση μάζας πιθανότητας των X,Y και $f_Y(y_j),\ j = 0,1,\ldots$ την περιθώρια συνάρτηση μάζας πιθανότητας της τυχαίας μεταβλητής Y. Η συνάρτηση :

$$f_{X|Y}(x_i|y_j) = \frac{f_{X,Y}(x_i, y_j)}{f_Y(y_j)}, i = 0,1,\ldots (j = 0,1,\ldots)$$

καλείται δεσμευμένη συνάρτηση μάζας πιθανότητας της X δεδομένης της $Y = y_j$.

Έστω (X,Y) μια συνεχή δισδιάστατη τυχαία μεταβλητή με $f_{X,Y}(x,y), x,y \in \mathbb{R}$ την από κοινού συνάρτηση πυκνότητας πιθανότητας των X,Y και $f_Y(y), y \in \mathbb{R}$ την περιθώρια συνάρτηση πυκνότητας πιθανότητας της τυχαίας μεταβλητής Y. Η συνάρτηση :

$$f_{X|Y}(x|y) = \frac{f_{X,Y}(x,y)}{f_Y(y)}, x \in \mathbb{R}$$

καλείται δεσμευμένη συνάρτηση πυκνότητας πιθανότητας της X δεδομένης της $Y = y$.

### *2.3.6 Ανεξαρτησία Τυχαίων Μεταβλητών*

Δύο τυχαίες μεταβλητές X, Y καλούνται ανεξάρτητες όταν :

$P(\{X \in A\} \cap \{Y \in B\}) = P(X \in A)P(Y \in B)$ για όλα τα τα διαστήματα $A, B \in \mathbb{R}$

Οι τυχαίες μεταβλητές X,Y καλούνται ανεξάρτητες αν και μόνο αν :

$$F_{X,Y}(x,y) = F_X(x)F_Y(y)\ ,\ \ \forall x,y \in \mathbb{R}$$

όπου $F_{X,Y}, F_X, F_Y$ η από κοινού συνάρτηση κατανομής πιθανότητας των X,Y, η περιθώρια συνάρτηση κατανομής πιθανότητας της X και η περιθώρια συνάρτηση κατανομής πιθανότητας της Y αντίστοιχα.

Αν οι τυχαίες μεταβλητές X,Y είναι διακριτές με από κοινού συνάρτηση μάζας πιθανότητας $f_{X,Y}(x_i, y_j)$ , $i, j = 0,1, \ldots$ και περιθώριες συναρτήσεις $f_X(x_i)$ , $i = 0,1, \ldots$ και $f_Y(y_j)$ , $j = 0,1, \ldots$ τότε είναι ανεξάρτητες αν και μόνο αν :

$$f_{X,Y}(x_i, y_j) = f_X(x_i)\, f_Y(y_j)\,, \quad \forall i, j = 0,1, \ldots$$

Αν οι τυχαίες μεταβλητές X,Y είναι συνεχείς με από κοινού συνάρτηση πυκνότητας πιθανότητας $f_{X,Y}(x, y)$ και περιθώριες συναρτήσεις $f_X(x)$ και $f_Y(y)$ , $x \in \mathbb{R}$ , $y \in \mathbb{R}$ τότε είναι ανεξάρτητες αν και μόνο αν :

$$f_{X,Y}(x, y) = f_X(x)\, f_Y(y)\,, \quad \forall x, y \in \mathbb{R}$$

Οι τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ καλούνται ανά δύο ανεξάρτητες όταν:

$$P(\{X_i \in A\} \cap \{X_j \in B\}) = P(X_i \in A) P(X_j \in B)$$

για όλα τα τα διαστήματα $A, B \subset \mathbb{R}$ με $i \neq j$ και $i, j = 0,1, \ldots, n$.

Οι τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ καλούνται ανεξάρτητες όταν :

$$P(\bigcap_{i=1}^{n} \{X_i \in B_i\}) = \prod_{i=1}^{n} P(X_i \in B_i)$$

για όλα τα διαστήματα $B_1, B_2, \ldots, B_n \subset \mathbb{R}$

## 2.4 Μέση Τιμή Τυχαίων Μεταβλητών

Η μέση τιμή μιας τυχαίας μεταβλητής $X$ συμβολίζεται με $E[X]$ (expected value of random variable X) και είναι ένας αριθμός που εκφράζει με τον καλύτερο δυνατό τρόπο την τιμή γύρω από την οποία τοποθετούνται οι τιμές της τυχαίας μεταβλητής. Στην ουσία η μέση τιμή αποτελεί γενίκευση του αριθμητικού μέσου μιας ακολουθίας αριθμών.

### *2.4.1 Μέση Τιμή τυχαίας μεταβλητής*

*Μέση τιμή διακριτής τυχαίας μεταβλητής*

Έστω μια διακριτή τυχαία μεταβλητή $X$ με συνάρτηση μάζας πιθανότητας:

$$f(x_\kappa) = P(X = x_\kappa), \qquad \kappa = 0,1, \dots$$

τότε μέση τιμή $\mu = E[X]$ της τυχαίας μεταβλητής $X$ δίνεται από τη σχέση:

$$\mu = E[X] = \sum_{\kappa=0}^{\infty} x_\kappa f(x_\kappa)$$

υπό την προϋπόθεση ότι η σειρά συγκλίνει απολύτως, δηλαδή:

$$\sum_{\kappa=0}^{\infty} |x_\kappa| f(x_\kappa) < +\infty$$

*Μέση τιμή συνεχής τυχαίας μεταβλητής*

Έστω μια συνεχής τυχαία μεταβλητή $X$ με συνάρτηση πυκνότητας πιθανότητας:

$$f(x) = P(X = x)$$

τότε μέση τιμή $\mu = E[X]$ της τυχαίας μεταβλητής $X$ δίνεται από τη σχέση:

$$\mu = E[X] = \int_{-\infty}^{+\infty} x f(x)\, dx$$

υπό την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απολύτως, δηλαδή:

$$\int_{-\infty}^{+\infty} |x| f(x)\, dx < +\infty$$

### *2.4.2 Μέση Τιμή συνάρτησης τυχαίας μεταβλητής*

Ας υποτεθεί ότι έχουμε μια πραγματική συνάρτηση $g: \mathbb{R} \to \mathbb{R}$ και μια τυχαία μεταβλητή $X$, τότε η $Y = g(X)$ θα είναι επίσης μια τυχαία μεταβλητή.

Συγκεκριμένα, αν η $X$ είναι μια διακριτή τυχαία μεταβλητή τότε και η $Y = g(X)$ θα είναι μια διακριτή τυχαία μεταβλητή. Ενώ στην περίπτωση που η $X$ είναι μια συνεχής τυχαία μεταβλητή, τότε και η $Y = g(X)$ θα είναι μια συνεχής τυχαία μεταβλητή.

Έστω μια διακριτή τυχαία μεταβλητή $X$ με συνάρτηση μάζας πιθανότητας:

$$f(x_\kappa) = P(X = x_\kappa), \qquad \kappa = 0,1, \dots$$

και $y = g(x)$ να είναι μια πραγματική συνάρτηση, τότε η μέση τιμή της διακριτής τυχαίας μεταβλητής $Y = g(X)$ δίνεται από τη σχέση:

$$E[Y] = E[g(X)] = \sum_{\kappa=0}^{+\infty} g(x_\kappa)\, f(x_\kappa)$$

υπό την προϋπόθεση ότι η σειρά συγκλίνει απολύτως, δηλαδή:

$$\sum_{\kappa=0}^{\infty} |g(x_\kappa)| f(x_\kappa) < +\infty$$

Έστω μια συνεχή τυχαία μεταβλητή $X$ με συνάρτηση πυκνότητας πιθανότητας:

$$f(x) = P(X = x)$$

και $y = g(x)$ να είναι μια πραγματική συνάρτηση, τότε η μέση τιμή της συνεχούς τυχαίας μεταβλητής $Y = g(X)$ δίνεται από τη σχέση:

$$E[Y] = E[g(X)] = \int_{-\infty}^{+\infty} g(x) f(x)\, dx$$

υπό την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απολύτως, δηλαδή:

$$\int_{-\infty}^{+\infty} |g(x)| f(x)\, dx < +\infty$$

### *2.4.3 Μέση Τιμή Τυχαίου Διανύσματος*

Έστω ότι έχουμε ένα τυχαίο διάνυσμα $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, τότε η μέση τιμή του τυχαίου διανύσματος δίνεται ως εξής:

$$E[\boldsymbol{X}] = (E[X_1], E[X_2], \ldots, E[X_n])^T$$

*Μέση Τιμή Συνάρτησης Τυχαίου Διανύσματος*

Έστω μια διακριτή n-διάστατη τυχαία μεταβλητή $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ με συνάρτηση μάζας πιθανότητας $f(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x})$, με $\boldsymbol{x} = (x_{1,i_1}, x_{2,i_2}, \ldots, x_{n,i_n})$, όπου $i_r = 0,1, \ldots$ και $r = 1,2, \ldots, n$. Τότε αν $y = g(\boldsymbol{x}) = g(x_1, x_2, \ldots, x_n)$ μια πραγματική συνάρτηση, η μέση τιμή της διακριτής τυχαίας μεταβλητής $Y = g(\boldsymbol{X})$ δίνεται από τη σχέση:

$$\sum_{i_n=0}^{+\infty} \ldots \sum_{i_2=0}^{+\infty} \sum_{i_1=0}^{+\infty} g(x_{1,i_1}, x_{2,i_2}, \ldots, x_{n,i_n})\, f(x_{1,i_1}, x_{2,i_2}, \ldots, x_{n,i_n})$$

υπό την προϋπόθεση ότι η σειρά συγκλίνει απολύτως.

Έστω μια συνεχή n-διάστατη τυχαία μεταβλητή $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ με συνάρτηση πυκνότητας πιθανότητας $f(\boldsymbol{x}) = P(\boldsymbol{X} = \boldsymbol{x})$ με $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ όπου $x_i \in \mathbb{R}$ και $i = 1,2, \ldots, n$. Τότε αν $y = g(\boldsymbol{x}) = g(x_1, x_2, \ldots, x_n)$ μια πραγματική συνάρτηση, η μέση τιμή της συνεχούς τυχαίας μεταβλητής $Y = g(\boldsymbol{X})$ δίνεται από τη σχέση:

$$\int_{-\infty}^{+\infty} \ldots \int_{-\infty}^{+\infty} \int_{-\infty}^{+\infty} g(x_1, x_2, \ldots, x_n) f(x_1, x_2, \ldots, x_n)\, dx_1\, dx_2 \ldots dx_n$$

υπό την προϋπόθεση ότι το ολοκλήρωμα συγκλίνει απολύτως.

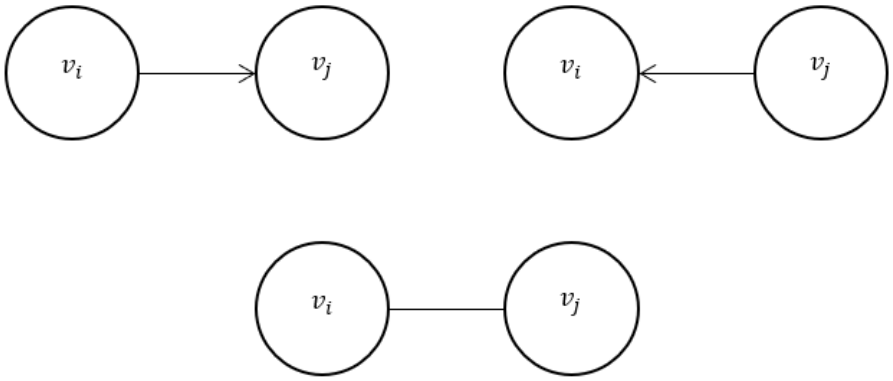
# 3 Στοιχεία Γραφημάτων

Στο κεφάλαιο αυτό θα παρουσιαστούν βασικές έννοιες των γραφημάτων και θα οριστεί η έννοια του κατευθυνόμενου ακυκλικού γραφήματος (DAG). Στη συνέχεια θα αναλυθούν οι βασικές συνδέσεις μεταξύ των κόμβων σε ένα τέτοιου είδους γράφημα και θα οριστεί η έννοια του d-separation μεταξύ των κόμβων του γραφήματος. Επιπλέον, θα γίνει αναφορά στις ιδιότητες του Markov και τη σχέση του d-separation με τη στοχαστική ανεξαρτησία μεταβλητών. Οι παραπάνω πληροφορίες είναι ουσιαστικές για τον ορισμό και την καλύτερη κατανόηση των μπεϋζιανών δικτύων.

## 3.1 Βασικές Έννοιες Γραφημάτων

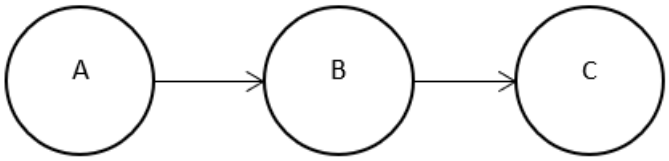
Ένα μπεϋζιανό δίκτυο βασίζεται σε ένα γράφημα από κόμβους που αντιστοιχούν σε τυχαίες μεταβλητές, το οποίο μάλιστα παρουσιάζει συγκεκριμένες ιδιότητες. Πριν γίνει η ανάλυση των ιδιοτήτων θα πρέπει αρχικά να εισαχθεί η έννοια του γραφήματος.

Ένα γράφημα G αποτελείται από ένα σύνολο κόμβων (ή κορυφών) V και ένα σύνολο ακμών (ή τόξων) E και συμβολίζεται ως G=(V,E). Ένα ζευγάρι κόμβων $v_i$ και $v_j$ με $v_i, v_j \in V$ μπορούν να συνδεόνται με ακμή που έχει κατεύθυνση $v_i \rightarrow v_j$ , $v_i \leftarrow v_j$ ή με ακμή χωρίς κατεύθυνση $v_i - v_j$ (Εικόνα 1). Άρα, τα γραφήματα μπορούν να χωριστούν σε κατευθυνόμενα (directed graphs) στη περίπτωση που οι ακμές τους έχουν κατεύθυνση και σε μη κατευθυνόμενα (undirected graphs) στη περίπτωση που οι ακμές τους δεν έχουν κατεύθυνση.



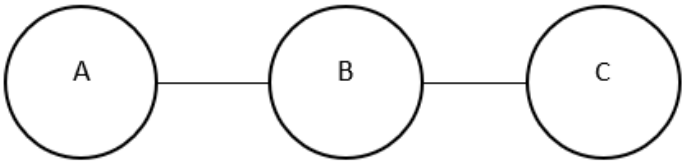

**Εικόνα 1. Είδη συνδέσεων μεταξύ δύο μεταβλητών $v_i - v_j$.**

Για να περιγραφεί η σύνδεση μεταξύ των κόμβων σε ένα κατευθυνόμενο γράφημα μπορεί να χρησιμοποιηθεί συγκεκριμένη ορολογία που παραπέμπει σε συγγενικές σχέσεις. Στην Εικόνα 2, ο κόμβος A είναι ο γονέας (parent) του κόμβου B και ο B το παιδί (child) του κόμβου A. Ο κόμβος C είναι απόγονος (descendant) του A και ο A είναι πρόγονος (ancestor) του κόμβου C. Επιπλέον, οι κόμβοι που δεν έχουν γονείς λέγονται ρίζες (roots), ενώ η οικογένεια (family) ενός κόμβου αποτελείται από τον ίδιο και τους γονείς του. Για παράδειγμα, ο κόμβος A είναι ρίζα, ενώ η οικογένεια του κόμβου B είναι ο ίδιος και ο A.



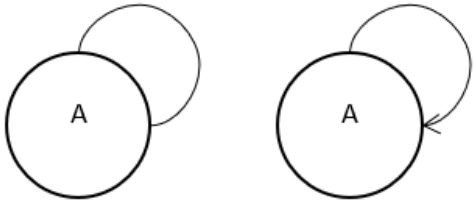

**Εικόνα 2. Παράδειγμα σύνδεσης μεταξύ των κόμβων σε ένα κατευθυνόμενο γράφημα.**

Στην περίπτωση που δεν υπάρχει κατεύθυνση στις ακμές, οι διπλανοί κόμβοι που συνδέονται με έναν συγκεκριμένο κόμβο ονομάζονται γείτονες (neighbors). Στην Εικόνα 3, οι κόμβοι A και C είναι οι γείτονες του κόμβου B.



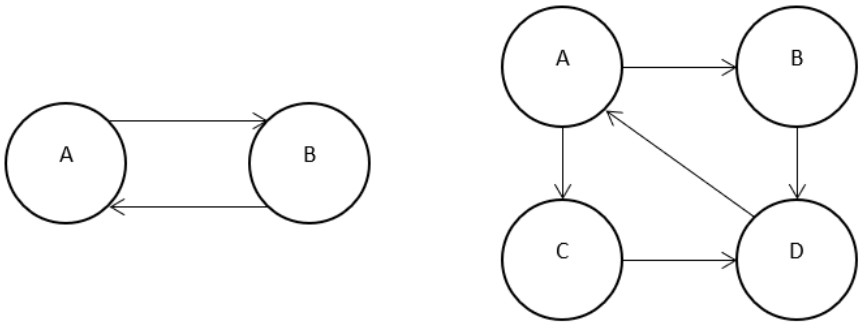

**Εικόνα 3. Παράδειγμα σύνδεσης μεταξύ των κόμβων σε ένα μη κατευθυνόμενο γράφημα**

Επιπλέον, σε ένα γράφημα βρόχος (loop) λέγεται η σύνδεση ενός κόμβου με τον εαυτό του μέσω μιας ακμής με ή χωρίς κατεύθυνση (Εικόνα 4).



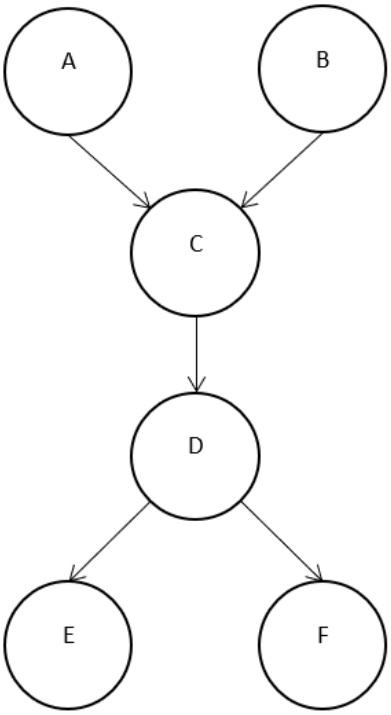

**Εικόνα 4. Παράδειγμα βρόχου σε γράφημα.**

Σε ένα κατευθυνόμενο γράφημα, κύκλος ονομάζεται μια διαδρομή από ακμές με κατεύθυνση που ξεκινούν και καταλήγουν στον ίδιο κόμβο. Για παράδειγμα, στην Εικόνα 5 παρουσιάζονται δύο περιπτώσεις κύκλου, αφού ξεκινώντας από τον κόμβο A η διαδρομή μπορεί να καταλήξει πάλι στον κομβο A.




**Εικόνα 5. Παράδειγμα κύκλων σε κατευθυνόμενο γράφημα.**

Μπορεί πλέον με βάση τις έννοιες που έχουν αναφερθεί, να οριστεί η έννοια του κατευθυνόμενου ακυκλικού γραφήματος. Πιο συγκεκριμένα, ένα γράφημα που κάθε ακμή του έχει κατεύθυνση, ενώ παράλληλα δεν έχει βρόχους και κύκλους, ονομάζεται κατευθυνόμενο ακυκλικό γράφημα (Directed Acyclic Graph-DAG) (Koller and Friedman, 2009; Pearl, 2009).




**Εικόνα 6. Παράδειγμα DAG γραφήματος.**

Αν με $Pa(X)$ συμβολίζεται το σύνολο των γονιών ενός κόμβου $X$, τότε στο παραπάνω γράφημα (Εικόνα 6) μπορούν να παρατηρηθούν τα εξής:

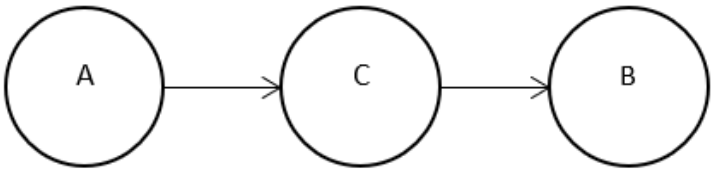
- Το γράφημα είναι ένα DAG, αφού κάθε ακμή έχει κατεύθυνση και δεν υπάρχουν βρόχοι και κύκλοι.
- Οι κόμβοι $A, B$ δεν έχουν κόμβους γονείς, και άρα είναι ρίζες.
- Οι γονείς του C είναι: $Pa(C) = \{A, B\}$ και $A, B, C$ αποτελούν οικογένεια.
- Ο γονέας του D είναι: $Pa(D) = \{C\}$ και $C, D$ αποτελούν οικογένεια.
- Ο γονέας του E είναι: $Pa(E) = \{D\}$ και $D, E$ αποτελούν οικογένεια.
- Ο γονέας του F είναι: $Pa(F) = \{D\}$ και $D, F$ αποτελούν οικογένεια.

## 3.2 Βασικές Συνδέσεις Μεταβλητών/Κόμβων σε DAG

Για να γίνει η ανάλυση των βασικών συνδέσεων των κόμβων σε γραφήματα τύπου DAG θα πρέπει προηγουμένως να αναφερθούν οι έννοιες της διαδρομής (path) και του μπλοκαρίσματος μιας διαδρομής (blocking). Επιπλέον, θα πρέπει να γίνει αναφορά στη d-separation (directed separation) που είναι μια γραφική έννοια διαχωρισμού των κόμβων ή μεταβλητών σε ένα κατευθυνόμενο γράφημα.

Γενικά, διαδρομή καλείται μια ακολουθία διαδοχικών ακμών στο γράφημα, είτε έχουν κατεύθυνση είτε δεν έχουν. Το μπλοκάρισμα διαδρομής αναφέρεται στη διαδικασία κατά την οποία σταματάει η ροή πληροφορίας ή αλλιώς η εξάρτηση, ανάμεσα στους κόμβους και άρα στις μεταβλητές που συνδέονται στη διαδρομή.

Ο πρώτος τρόπος σύνδεσης των μεταβλητών είναι η κατά σειρά σύνδεση (serial connection) ή αλλιώς αιτιακή διαδρομή (causal chain). Για τρεις μεταβλητές, A, B, C, η σύνδεση σε σειρά παρουσιάζεται στην Εικόνα 7.

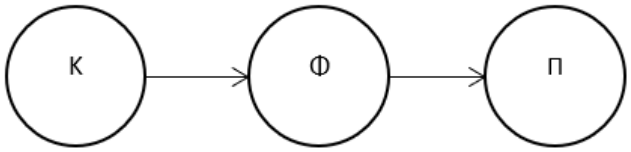



**Εικόνα 7. Παράδειγμα σύνδεσης σε σειρά.**

Στην συγκεκριμένη περίπτωση, οποιαδήποτε πληροφορία για την κατάσταση της μεταβλητής A θα επηρεάσει την πεποίθηση που έχουμε για την κατάσταση της μεταβλητής C, η οποία με τη σειρά της θα επηρεάσει την πεποίθηση μας για την μεταβλητή B. Έτσι, μπορούμε να πούμε ότι η ροή της πληροφορίας στη διαδρομή από την A στην B δεν είναι μπλοκαρισμένη. Αν όμως η κατάσταση της μεταβλητής C είναι γνωστή, τότε αν είναι γνωστή και η κατάσταση της A δεν θα επηρεάσει την πεποίθηση μας για την B και αντίστροφα. Δηλαδή, υπό τη συνθήκη ότι γνωρίζουμε την κατάσταση της C, οποιαδήποτε πληροφορία για την κατάσταση της A μπλοκάρεται με αποτέλεσμα να μην επηρεάζει την πεποίθηση μας για την κατάσταση της B. Σε αυτήν

τη περίπτωση η διαδρομή μεταξύ της Α και της Β έχει μπλοκαριστεί και λέμε ότι οι μεταβλητές Α και Β είναι d-separated δεδομένης της C.

Για παράδειγμα, έστω ότι έχουμε την σύνδεση σε σειρά των μεταβλητών του καπνού (Κ), της φωτιάς (Φ) και της πυροσβεστικής (Π). Η αναπαράσταση αυτής της σύνδεσης σε δίκτυο παρουσιάζεται στην Εικόνα 8.

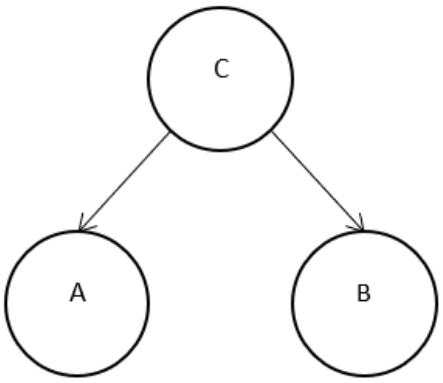



**Εικόνα 8. Παράδειγμα σύνδεσης σε σειρά με κόμβους Καπνός (Κ), Φωτιά (Φ) και Πυροσβεστική (Π).**

Υποθέτοντας ότι δεν γνωρίζουμε τίποτα για το αν υπάρχει φωτιά, τότε, αν μας δοθεί η πληροφορία ότι υπάρχει καπνός θα οδηγήσει σε ανανέωση της πεποίθησής μας για την ύπαρξη της φωτιάς, δηλαδή θα είναι λογικό να θεωρούμε περισσότερο πιθανή την ύπαρξη φωτιάς. Αυτό με τη σειρά του θα οδηγήσει σε ανανέωση της πεποίθησής μας για το αν έχει έρθει η πυροσβεστική, αφού όταν καθίσταται πιθανότερη η ύπαρξη της φωτιάς γίνεται και πιθανότερη η επέμβαση της πυροσβεστικής. Αντίστροφα, αν ξέρουμε ότι έχει έρθει η πυροσβεστική, αυτό καθιστά περισότερο πιθανή την ύπαρξη φωτιάς, η οποία με τη σειρά της καθιστά πιθανότερη την ύπαρξη καπνού. Μπορεί λοιπόν να παρατηρήσει κανείς ότι η πληροφορία μεταδίδεται στη σύνδεση εφόσον δεν ξέρουμε τι γίνεται με την κεντρική μεταβλητή και άρα υπονοείται ότι οι μεταβλητές είναι εξαρτημένες η μια από την άλλη.

Στην περίπτωση που γνωρίζουμε ότι έχει ξεσπάσει φωτιά, οποιαδήποτε πληροφορία σχετίζεται με τον καπνό δεν επηρεάζει την πεποίθηση μας για το αν έχει έρθει η πυροσβεστική. Αντίστροφα, αν δοθεί η πληροφορία ότι έχει έρθει η πυροσβεστική δεν μεταβάλλει την πεποίθηση μας για την ύπαρξη καπνού, από τη στιγμή που γνωρίζουμε ότι υπάρχει φωτιά. Δηλαδή, οποιαδήποτε νέα πληροφορία για τον καπνό μπλοκάρεται υπό τη συνθήκη της φωτιάς και αντίστροφα. Άρα οι μεταβλητές Κ, Π είναι d-separated δεδομένου της μεταβλητής Φ, γεγονός που υπονοεί ότι είναι και ανεξάρτητες υπό τη συνθήκη της Φ.

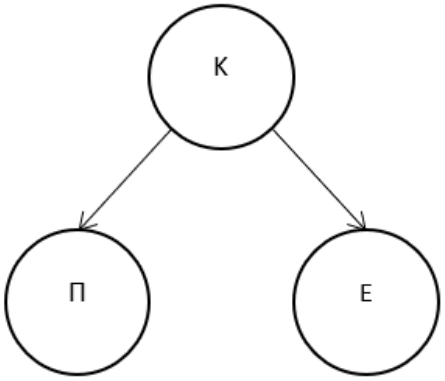
Ο δεύτερος τρόπος σύνδεσης είναι η αποκλίνουσα σύνδεση (diverging connection) ή αλλιώς η κοινή αιτία (common cause). Για τρεις μεταβλητές A, B, C, αποκλίνουσα σύνδεση παρουσιάζεται στην Εικόνα 9




**Εικόνα 9. Παράδειγμα αποκλίνουσας σύνδεσης.**

Στην συγκεκριμένη περίπτωση, αν αρχικά δεν υπάρχει κάποια πληροφορία για την κατάσταση της μεταβλητής C, τότε γνώση της κατάστασης της μεταβλητής A θα επηρεάσει την πεποίθηση μας για την κατάσταση της C. Η νέα πεποίθηση που έχουμε για την C θα επηρεάσει με τη σειρά της το τι πιστεύουμε για τη μεταβλητή B. Δηλαδή, η πληροφορία μεταδίδεται από την A στην B και άρα διαδρομή δεν είναι μπλοκαρισμένη. Αν όμως η κατάσταση της μεταβλητής C είναι γνωστή, τότε η γνώση της κατάστασης της A δεν θα επηρεάσει την πεποίθηση μας για την κατάσταση της B και αντίστροφα. Τότε, η διαδρομή μεταξύ της A και της B έχει μπλοκαριστεί και λέμε ότι οι μεταβλητές A και B είναι d-separated δεδομένης της C.

Για παράδειγμα, έστω ότι έχουμε τις μεταβλητές του καπνίσματος (Κ), της πνευμονοπάθειας (Π) και του εγκεφαλικού επισοδείου (Ε) σε αποκλίνουσα σύνδεση, όπως παρουσιάζεται στην Εικόνα 10.



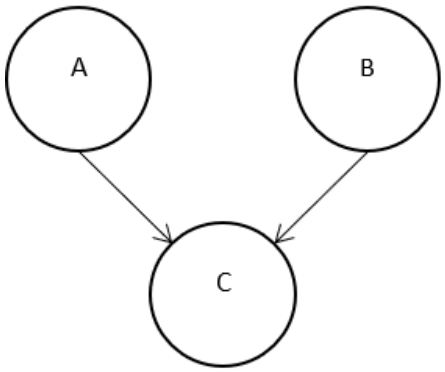

**Εικόνα 10. Παράδειγμα αποκλίνουσας σύνδεσης με κόμβους Κάπνισμα (Κ), Πνευμονοπάθια (Π) και Εγκεφαλικό επεισόδιο (Ε).**

Στην περίπτωση που δεν γνωρίζουμε αν κάποιος καπνίζει, η πληροφορία ότι πάσχει από κάποια πνευμονοπάθεια θα μεταβάλλει την εκτίμηση μάς για το αν καπνίζει και συγκεκριμένα την καθιστά πιθανότερη σε σχέση με προηγουμένως που δεν γνωρίζαμε κάτι για την πνευμονοπάθεια. Έχοντας περισσότερο πιθανό το ενδεχόμενο του καπνίσματος με τη σειρά του θα ανανεώσει την πεποίθηση μας για το ενδεχόμενο εγκαφαλικού επισοδείου. Πράγματι, είναι λογικό να θεωρήσουμε ότι εφόσον είναι πιθανότερο να καπνίζει σε σχέση με το παρελθόν, τότε είναι και πιθανότερο να υποστεί κάποιο εγκεφαλικό. Με την ίδια λογική, αν ακολουθήσουμε την αντίστροφη πορεία ξεκινώντας από την πληροφόρια ότι υπέστει εγκεφαλικό επισόδειο θα αλλάξουμε την πεποίθηση μας για το ενδεχόμενου να καπνίζει και ύστερα για το ενδεχόμενο να πάσχει από κάποια πνευμονοπάθεια. Άρα αν δεν γνωρίζουμε κάτι για το κοινό αίτιο, η ροή της πληροφορίας δεν μπλοκάρεται και η πληροφορία για το ένα ενδεχόμενο θα επηρεάσει τη πεποίθηση μας για το άλλο, δηλαδή υπονοείται η ύπαρξη εξάρτησης μεταξύ των μεταβλητών Ε και Π ή αλλιώς μεταξύ των αποτελεσμάτων της κοινής αιτίας.

Αν γνωρίζουμε όμως ότι κάποιος καπνίζει, τότε αν δοθεί η πληροφορία ότι πάσχει από κάποια πνευμονοπάθεια δεν θα μεταβάλλει την πεποίθηση μας για το αν υπέστει εγκεφαλικό επισόδειο. Αντίστροφα, αν υπέστει εγκεφαλικό επισόδειο δεν αλλάζει την πεποίθηση μας για το αν πάσχει από κάποια πνευμονοπάθεια, δεδομένου ότι γνωρίζουμε ότι καπνίζει. Άρα η ροή της πληροφορίας μπλοκάρεται και μπορούμε να ισχυριστούμε ότι οι μεταβλητές Π, Ε είναι d-separated υπό τη συνθήκη της Κ, δηλαδή

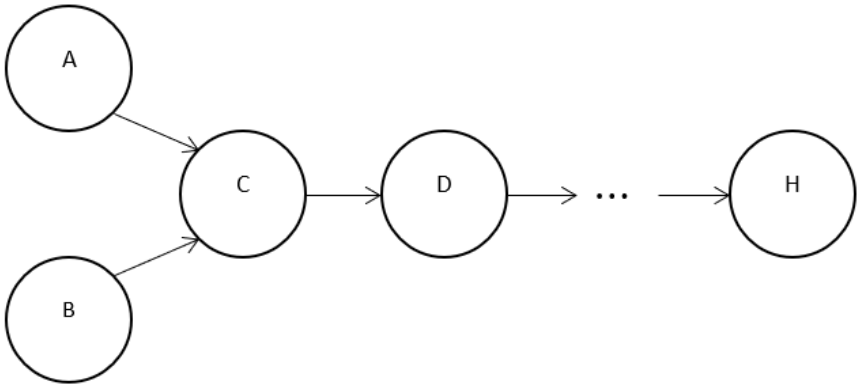
υπονοείται ότι τα αποτελέσματα καθίστανται ανεξάρτητα υπό τη συνθήκη της κοινής αιτίας.

Ο τρίτος τρόπος σύνδεσης καλείται συγκλίνουσα σύνδεση (converging connections) ή αλλιώς κοινό αποτέλεσμα (common effect) ή v-δομή (v-structure). Για τρεις μεταβλητές A, B, C, η συγκλίνουσα σύνδεση παρουσιάζεται στην Εικόνα 11.




**Εικόνα 11. Παράδειγμα συγκλίνουσας σύνδεσης.**

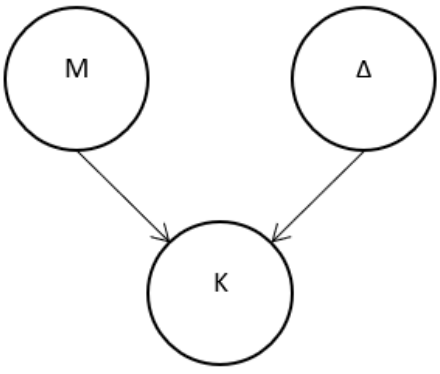
Σε αντίθεση με τις προηγούμενες συνδέσεις, στη συγκλίνουσα σύνδεση όταν δεν γνωρίζουμε τίποτα για την κατάσταση της C, η γνώση της κατάστασης της A δεν επηρεάζει την πεποίθηση μας για την κατάσταση της B και αντίστροφα. Αν όμως είναι γνωστή η κατάσταση της C, τότε η γνώση της κατάστασης μίας εκ των δύο μεταβλητών A ή B θα επηρέασει την πεποίθησή μας για την άλλη. Επίσης, αν είναι γνωστή η κατάσταση ενός απογόνου της C και όχι της ίδιας, τότε πάλι γνώση της κατάστασης της A θα επηρεάσει την πεποίηση μας για την κατάσταση της B και αντίστροφα. Όπως φαίνεται και στην Εικόνα 12, αυτό συμβαίνει διότι η γνώση της κατάστασης της H μπορεί να μας δώσει πληροφορία για την κατάσταση της C.




**Εικόνα 12. Παράδειγμα συγκλίνουσας σύνδεσης με απογόνους.**

Συνεπώς, η γνώση της κατάστασης της C ή κάποιας απογόνου της θα μπλοκάρει τη ροή της πληροφορίας στη διαδρομή και άρα οι A, B παρουσιάζουν εξάρτηση μεταξύ τους. Το αντίθετο θα συμβεί σε περίπτωση που δεν γνωρίζουμε τίποτα για την C ή για κάποια απόγονο της, δηλαδή οι A, B θα είναι ανεξάρτητες.

Υποθέτουμε ότι σε ένα εργοστάσιο φυσικού αερίου έχουμε τις μεταβλητές της μηχανικής βλάβης (M) σε κάποιο από τα μηχανήματα, της διαρροής αερίου (Δ) σε κάποιο σωλήνα και της ένδειξης κινδύνου (K) σε έναν πίνακα ελέγχου, σε συγκλίνουσα σύνδεση, όπως παρουσιάζεται στην Εικόνα 13.




**Εικόνα 13. Παράδειγμα συγκλίνουσας σύνδεσης με κόμβους Μηχανική βλάβη (M), Διαρροή αερίου (Δ) και Ένδειξη κινδύνου (E).**

Σε αντίθεση με τις προηγούμενες συνδέσεις, αν δεν γνωρίζουμε τίποτα για την ένδειξη κινδύνου στην περίπτωση που πληροφορηθούμε ότι υπάρχει μηχανική βλάβη σε κάποιο μηχάνημα δεν επηρεάζει την πεποίθηση μας για την ύπαρξη διαρροής αερίου σε κάποιο σωλήνα. Αντίστροφα, η πληροφορία της διαρροής σε κάποιο σωλήνα του εργοστασίου δεν επηράζει με κάποιο τρόπο την πεποίθηση μας για την ύπαρξη μηχανικής βλάβης σε κάποιο μηχάνημα. Δηλαδή η ροή της πληροφορίας μπλοκάρεται από το γεγονός ότι δεν ξέρουμε τίποτα για την ένδειξη κινδύνου. Άρα, στην περίπτωση που δεν γνωρίζουμε τίποτα για την μεταβλητή Κ, τότε μπορούμε να ισχυριστούμε ότι οι μεταβλητές Μ, Δ είναι ανεξάρτητες, δηλαδή αν δεν γνωρίζουμε κάτι για το κοινό αποτέλεσμα τότε τα αίτια φέρονται ως ανεξάρτητα μεταξύ τους.

Αν όμως γνωρίζουμε ότι υπάρχει ένδειξη κινδύνου στον πίνακα ελέγχου τότε αν μας δοθεί η πληροφορία ότι υπάρχει μηχανική βλάβη επηρεάζει την πεποίθηση που έχουμε για την διαρροή και συγκεκριμένα θα μειώσει την όποια εκτίμηση είχαμε να είναι η διαρροή το αίτιο της ένδειξης κινδύνου. Αντίστροφα, η πληροφορία της διαρροής σε κάποιο σωλήνα, υπό τη συνθήκη ότι έχουμε ένδειξη κινδύνου, θα μειώσει την εκτίμηση μας στο να είναι η μηχανική βλάβη το αίτιο. Άρα, γνωρίζοντας την κατάσταση της μεταβλητής Κ η ροή της πληροφορίας δεν μπλοκάρεται και οι μεταβλητές Μ, Δ εξαρτώνται μεταξύ τους. Συνεπώς, αν γνωρίζουμε το κοινό αποτέλεσμα τότε τα αίτια φέρονται ως εξαρτώμενα μεταξύ τους.

Με βάση την προηγούμενη ανάλυση διαπιστώνει κανείς ότι υποβόσκει μια σχέση μεταξύ της έννοιας του d-separation και της υπό συνθήκη στοχαστικής ανεξαρτησίας. Για να γίνει καλύτερα κατανοήτη η συγκεκριμένη σχέση θα πρέπει πρώτα να οριστεί η έννοια του d-separation (Jensen and Nielsen, 2007; Kjaerulff and Madsen, 2013; Koller and Friedman, 2009; Pearl, 2009).

## 3.3 Κριτήριο D-Separation και η Κουβέρτα του Markov

Γνωρίζοντας τις συνδέσεις των μεταβλητών και έχοντας κάνει ήδη μια μικρή αναφορά στη γραφική έννοια του d-separation (Directed Separation Criterion), μπορεί πλέον να οριστεί αυτή η έννοια με αυστηρό τρόπο. Έστω **A**, **B** και **C** τρία ασύνδετα σύνολα μεταβλητών ή αλλιώς ασύνδετα σύνολα κόμβων ή μεταβλητών ενός DAG. Μια διαδρομή p μπλοκάρεται από το σύνολο κόμβων **C** εάν και μόνο εάν ισχύει ένα από τα παρακάτω:

1. Η διαδρομή p να περιέχει μια κατά σειρά σύνδεση i ⟶ m ⟶ j ή μια αποκλίνουσα σύνδεση i ⟵ m ⟶ j, όπου m να είναι κόμβος που ανήκει στο **C**.
2. Η διαδρομή p να περιέχει μια συγκλίνουσα σύνδεση i ⟶ m ⟵ j, όπου m να είναι κόμβος που δεν ανήκει στο **C** και κανένας από τους απογόνους του δεν ανήκει στο **C**.

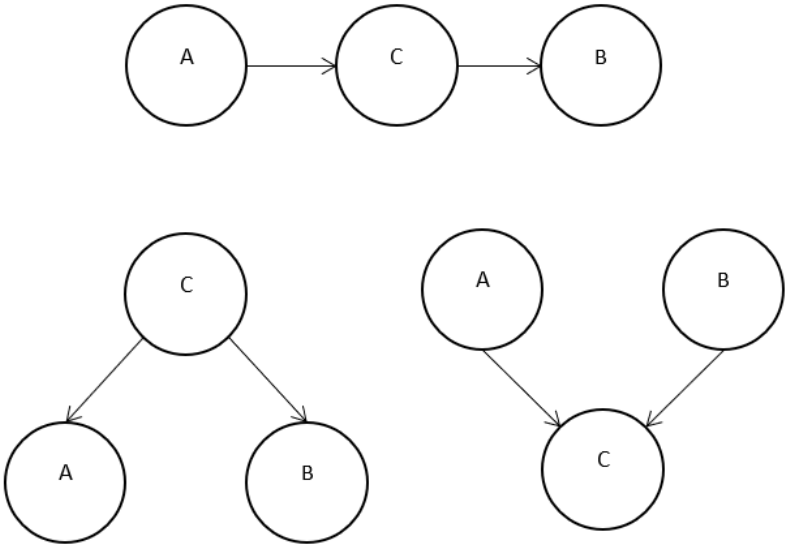
Τότε, λέμε ότι τα **A** και **B** είναι d-separated δεδομένου του **C** δηλαδή $(\boldsymbol{A} \perp\!\!\!\perp \boldsymbol{B}|\boldsymbol{C})_G$, εάν και μόνο εάν το **C** μπλοκάρει κάθε διαδρομή από έναν κόμβο στο **A** σε έναν κόμβο στο **B** (Pearl, 2009).

Σημαντικό αποτέλεσμα του κριτηρίου αποτελεί το γεγονός ότι το d-separation συνεπάγεται και υπό συνθήκη στοχαστική ανεξαρτησία. Συγκεκριμένα, έστω ότι έχουμε τρία ασύνδετα υποσύνολα κόμβων ή μεταβλητών $\boldsymbol{X}, \boldsymbol{Y}, \boldsymbol{Z}$ ενός γραφήματος G τύπου DAG. Τότε:

$$(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G \Rightarrow (\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_P$$

όπου το $()_P$ συμβολίζει στοχαστική ανεξαρτησία. Δηλαδή αν οι $\boldsymbol{X}, \boldsymbol{Y}$ είναι d-separated δεδομένου του $\boldsymbol{Z}$ αυτό σημάινει ότι θα είναι και στοχαστικά ανεξάρτητοι υπό την συνθήκη του $\boldsymbol{Z}$. Ισχύει όμως και το αντίστροφο αποτέλεσμα, δηλαδή αν σε ένα DAG οι $\boldsymbol{X}, \boldsymbol{Y}$ είναι ανεξάρτητοι υπό τη συνθήκη του $\boldsymbol{Z}$ τότε θα είναι και d-separated δεδομένου του $\boldsymbol{Z}$. Δηλαδή:

$$(\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_P \Rightarrow (\boldsymbol{X} \perp\!\!\!\perp \boldsymbol{Y}|\boldsymbol{Z})_G$$



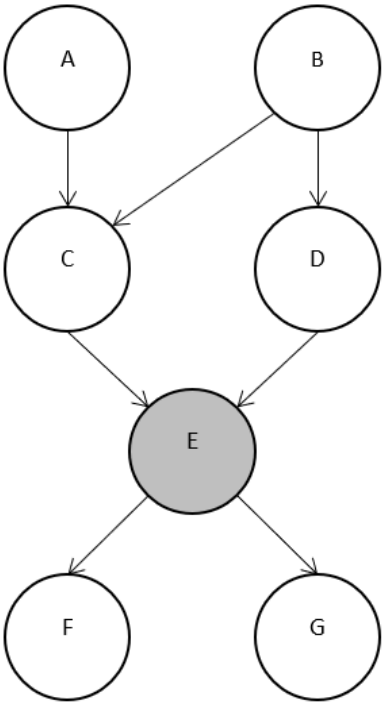

**Εικόνα 14. Πιθανές συνδέσεις μεταξύ τριών κόμβων σε ένα γράφημα DAG.**

Υπό το πρίσμα του συγκεκριμένου αποτελέσματος αν ξαναγυρίσουμε στις συνδέσεις των μεταβλητών, όπως αυτές παρουσιάζονται συνολικά στην Εικόνα 14, μπορούμε να παρατηρήσουμε ότι στις περιπτώσεις της σύνδεσης σε σειρά και της αποκλίνουσας σύνδεσης, αν γνωρίζουμε την κατάσταση της C, τότε οι μεταβλητές A, B είναι d-separated δεδομένου της C και άρα οι A, B είναι στοχαστικά ανεξάρτητες δεδομένου της C.

$$(A \perp\!\!\!\perp B|C)_G \Rightarrow (A \perp\!\!\!\perp B|C)_P \Rightarrow P(A \mid B, C) = P(A \mid C)$$

Αντίθετα στην συγκλίνουσα σύνδεση όταν δεν γνωρίζουμε την κατάσταση της C η κάποιου απογόνου της τότε οι A, B δεν εξαρτώνται μεταξύ τους και άρα είναι στοχαστικά ανεξάρτητες.

$$(A \perp\!\!\!\perp B)_P \Rightarrow P(A \mid B) = P(A) \text{ και } P(B \mid A) = P(B)$$




**Εικόνα 15. Παράδειγμα DAG με evidence node.**

Για παράδειγμα, έστω ότι έχουμε το DAG της Εικόνας 15 και θέλουμε να απαντήσουμε στο ερώτημα αν οι μεταβλητές $A, B$ είναι ανεξάρτητες υπό τη συνθήκη της $E$. Η μεταβλητή $E$ είναι εκείνη η μεταβλητή της οποίας την κατάσταση γνωρίζουμε ή αλλιώς έχουμε την έχουμε παρατηρήσει (observed variable/ evidence variable).

Αρχικά παρατηρούμε ότι από την $A$ στη $B$ υπάρχουν δύο διαδρομές η $A \rightarrow C \leftarrow B$ και η $A \rightarrow C \rightarrow E \leftarrow D \leftarrow B$.

Στη πρώτη διαδρομή έχουμε μια τριάδα μεταβλητών $A \rightarrow C \leftarrow B$ σε σύνδεση κοινού αποτελέσματος χωρίς να γνωρίζουμε κάτι για την κατάσταση της μεταβλητής $C$ που είναι το αποτέλεσμα. Άρα, από προηγούμενη ανάλυση γνωρίζουμε ότι στην περίπτωση αυτή η διαδρομή μπλοκάρεται και δεν υπάρχει ροή της πληροφορίας από $A$ σε $B$.

Η δεύτερη διαδρομή αποτελείται από τρεις το πλήθος τριάδες μεταβλητών που είναι οι εξής:

$$A \rightarrow C \rightarrow E$$

$$C \rightarrow E \leftarrow D$$

$$E \leftarrow D \leftarrow B$$

Στην πρώτη τριάδα $A \rightarrow C \rightarrow E$ οι μεταβλητές $A, C, E$ συνδέονται σε σειρά και δεν γνωρίζουμε κάτι για την κατάσταση της $C$ ή αλλιώς δεν έχουμε παρατηρήσει την μεταβλητή (unobserved variable), που σημαίνει ότι η πρώτη τριάδα δεν είναι μπλοκαρισμένη. Η επόμενη τριάδα $C \rightarrow E \leftarrow D$ είναι μια σύνδεση κοινού αποτελέσματος και συγχρόνως γνωρίζουμε την κατάσταση του αποτελέσματος που είναι η $E$. Άρα από προηγούμενη ανάλυση γνωρίζουμε ότι η τριάδα δεν είναι μπλοκαρισμένη. Η τελευταία τριάδα $E \leftarrow D \leftarrow B$ της δεύτερης διαδρομής είναι μια σύνδεση σε σειρά με την κατάσταση της κεντρικής μεταβλητής να είναι άγνωστη. Άρα πάλι όπως ξέρουμε, η τριάδα αυτή δεν είναι μπλοκαρισμένη.

Συνεπώς, η δέυτερη διαδρομή δεν είναι μπλοκαρισμένη αφού καμία από τις τριάδες που αποτελείται δεν μπλοκάρει τη ροή της πληροφορίας από $A$ σε $B$.Έχοντας τη πρώτη διαδρομή μπλοκαρισμένη και τη δεύτερη να μην είναι μπλοκαρισμένη, ξέρουμε ότι οι $A, B$ δεν είναι d-separated δεδομένου της $E$ και άρα δεν είναι ανεξάρτητες υπό τη συνθήκη της $E$.

Ένα ακόμα ερώτημα μπορεί να είναι αν οι $A, G$ είναι ανεξάρτητες υπό τη συνθήκη της $E$. Στο πλάισο αυτό παρατηρούμε ότι υπάρχουν δύο διαδρομές η, $A \rightarrow C \rightarrow E \rightarrow G$ και η $A \rightarrow C \leftarrow B \rightarrow D \rightarrow E \rightarrow G$.

Η πρώτη διαδρομή αποτελείται από δύο τριάδες που είναι οι εξής:

$$A \rightarrow C \rightarrow E$$

$$C \rightarrow E \rightarrow G$$

Η πρώτη τριάδα είναι μια σύνδεση σε σειρά χωρίς να έχουμε πληροφορία για την κεντρική μεταβλητή $C$, κάτι που σημαίνει ότι η πρώτη τριάδα δεν είναι μπλοκαρισμένη. Η δεύτερη τριάδα είναι πάλι μια σύνδεση σε σειρά όπου όμως γνωρίζουμε την κατάσταση της κεντρικής μεταβλητής $E$ και άρα στη δεύτερη τριάδα η ροή της πληροφορίας μπλοκάρεται. Το συμπέρασμα είναι ότι η πρώτη διαδρομή είναι μπλοκαρισμένη, αφού στην δεύτερη τριάδα μπλοκάρεται η ροή της πληροφορίας.

Η δεύτερη διαδρομή αποτελείται από τέσσερις τριάδες που είναι οι εξής:

$$A \rightarrow C \leftarrow B$$

$$C \leftarrow B \rightarrow D$$

$$B \rightarrow D \rightarrow E$$

$$D \rightarrow E \rightarrow G$$

Εξετάζωντας την πρώτη τριάδα παρατηρούμε ότι έχουμε σύνδεση κοινού αποτελέσματος όπου δεν είναι γνωστή η κατάσταση της $C$, κάτι που σημαίνει ότι η τριάδα είναι μπλοκαρισμένη. Επειδή όμως έχουμε τουλάχιστον ένα κομμάτι της διαδρομής μπλοκαρισμένο συμπαιράνουμε ότι δεν χρειάζεται να εξετάσουμε τις υπόλοιπες τριάδες αφού όλη η διαδρομή θα είναι μπλοκαρισμένη.

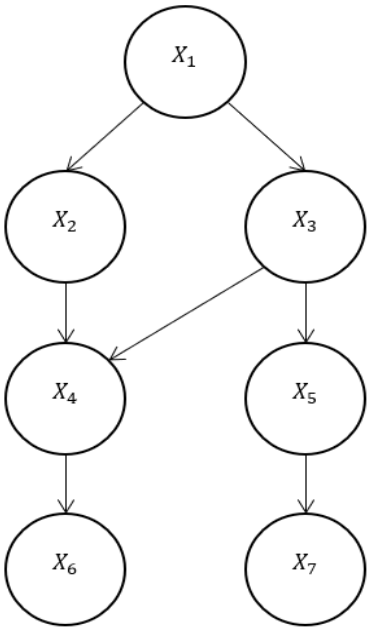
Κατά συνέπεια, επειδή κάθε διαδρομή από την $A$ στην $G$ είναι μπλοκαρισμένη δεδομένου της $E$ λέμε ότι οι $A, G$ είναι d-separated δεδομένου της $E$. Αυτό σημάινει ότι οι $A, G$ είναι ανεξάρτητες υπό τη συνθήκη της $E$.

Στη συνέχεια είναι σκόπιμο να γίνει αναφορά στην έννοια της κουβέρτας του Markov (Markov blanket), αφού αποτελεί κομβικό σημείο για την ανάλυση των μπεϋζιανών δικτύων. Γενικά σε ένα σύνολο τυχαίων μεταβλητών, η κουβέρτα του Markov μιας μεταβλητής είναι το ελάχιστο υποσύνολο που υπό τη συνθήκη του καθιστά την μεταβλητή ανεξάρτητη απο τις υπόλοιπες μεταβλητές του συνόλου. Έ

Έστω $V$ ένα σύνολο κόμβων ή τυχαίων μεταβλητών και έστω μια μεταβλητή $A$, με $A \in V$. Τότε το Markov Blanket της $A$ είναι το ελάχιστο υποσύνολο $S$ του $V$ τέτοιο ώστε:

$$\left(A \perp\!\!\!\perp \left(V - (S + A)\right)\right)_P \mid S$$

Συνεπώς, το Α θα είναι ανεξάρτητο από τους υπόλοιπους κόμβους-μεταβλητές του συνόλου υπό τη συνθήκη του Markov blanket του. Σε κάθε Μπεϋζιανό δίκτυο, η κουβέρτα του Markov μιας μεταβλητής $A$ είναι το σύνολο των γονέων της $A$, των παιδιών της $A$ και των μεταβλητών που έχουν κοινά παιδιά με την $A$.




**Εικόνα 16. Παράδειγμα DAG για την αποσαφήνιση του Markov Blanket.**

Άρα αν υποθέσουμε ότι έχουμε μεταβλητές $X_1, X_2, \ldots, X_7$ που συνδέονται με τον τρόπο που παρουσιάζεται στην Εικόνα 16, τότε η κουβέρτα του Markov (Mb) της $X_2$ θα είναι οι μεταβλητές:

$$Mb(X_2) = (X_1, X_3, X_4)$$

και άρα υπό τη συνθήκη του συνόλου αυτού των μεταβλητών η $X_2$ είναι ανεξάρτητη των υπόλοιπων μεταβλητών. Όμοια, η κουβέρτα του Markov της $X_4$ θα είναι:

$$Mb(X_4) = (X_2, X_3, X_6)$$

Μια ακόμα σημαντική ιδιότητα που θα χρειαστεί στην ανάλυση των μπεϋζιανών δικτύων είναι η λεγόμενη τοπική ιδιότητα του Markov. Σύμφωνα με τη συγκεκριμένη ιδιότητα, σε ένα γράφημα τύπου DAG, με $X_1, X_2, \ldots, X_n$ τυχαίες μεταβητές, κάθε μεταβλητή $X_i$ θα είναι ανεξάρτητη από τους μη απογόνους της (nondescendants) υπό τη συνθήκη των γονιών της. Δηλαδή, αν $pa_{X_i}$ συμβολίζει το σύνολο των γονέων της $X_i$ και $nondescendants_{X_i}$το σύνολο των μεταβλητών που δεν είναι απόγονοι της $X_i$ τότε για κάθε $X_i$:

$$\left(X_i \perp\!\!\!\perp nondescendants_{X_i} \mid pa_{X_i}\right)_P$$

Η τοπική ιδιότητα του Markov αποτελεί την βάση της ανάλυσης ενός μπεϋζιανού δικτύου που ακολουθεί στο επόμενο κεφάλαιο (Nagarajan *et al*., 2013; Koller and Friedman, 2009; Pearl, 2009).

# 4 Μπεϋζιανά Δίκτυα

Στο κεφάλαιο αυτό θα γίνει αναφορά των βασικών εννοιών των μπεϋζιανών δικτύων και θα αναλυθούν μέθοδοι εύρεσης παραμέτρων (parameter learning) καθώς και μέθοδοι κατασκευής της δομής ενός μπεϋζιανού δικτύου με βάση τα δεδομένα (structure learning).

Συγκεκριμένα, θα αναλυθούν η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας και η μέθοδος μπεϋζιανής εκτίμησης των παραμέτρων σε πλήρη δεδομένα. Επιπλέον θα αναλυθεί η εύρεση παραμέτρων με τον αλγόριθμο EM (expaxtation-maximization) για την περίπτωση που το σύνολο των δεδομένων δεν είναι πλήρες. Στη συνέχεια θα παρουσιαστούν τεχνικές κατασκευής της δομής ενός μπεϋζιανού δικτύου από πλήρη καθώς και από μη πήρη δεδομένα με τον δομικό αλγόριθμο EM (structural EM). Τέλος, θα γίνει αναφορά στο τρόπο εκτίμησης νέων δεδομένων με ακριβή τρόπο αλλά και με προσεγγιστικό τρόπο.

## 4.1 Μπεϋζιανά Δίκτυα και Πίνακες Δεσμευμένων Πιθανοτήτων

Τα μπεϋζιανά δίκτυα είναι γραφικά μοντέλα τύπου DAG που επιτρέπουν την αναπαράσταση των στοχαστικών εξαρτήσεων ανάμεσα σε ένα δοσμένο σύνολο τυχαίων μεταβλητών, όπου κάθε κόμβος του DAG αντιστοιχεί σε μια τυχαία μεταβλητή (Nagarajan et al., 2013). Συγκεκριμένα, ένα μπεϋζιανό δίκτυο αποτελείται από τα εξής:

i) Ένα σύνολο μεταβλητών και ένα σύνολο κατευθυνόμενων ακμών ανάμεσα στις μεταβλητές.

ii) Κάθε μεταβλητή έχει ένα πεπερασμένο σύνολο διαφορετικών μεταξύ τους καταστάσεων (τιμών).

iii) Οι μεταβλητές μαζί με τις κατευθυνόμενες ακμές αποτελούν ένα DAG.

iv) Σε κάθε μεταβλητή $A$ με γονείς $B_1, B_2, \dots, B_n$ αντιστοιχεί ένας πίνακας της δεσμευμένης συνάρτησης πιθανότητας $P(A|B_1, B_2, \dots, B_n)$.

Στην περίπτωση που μια μεταβλητή $A$ δεν έχει γονέα, τότε ο πίνακας αφορά τη συνάρτηση πιθανότητας $P(A)$ της τυχαίας μεταβλητής A (Jensen and Nielsen, 2007). Επιπλέον, ένα μπεϋζιανό δίκτυο μετουσιώνεται στοχαστικά μέσω μιας από κοινού συνάρτησης μάζας ή πυκνότητας πιθανότητας ανάλογα με το αν οι μεταβλητές είναι διακριτές ή συνεχείς αντίστοιχα.

Αν υποθέσουμε ότι έχουμε ένα μπεϋζιανό δίκτυο με γραφικό μοντέλο G τύπου DAG που δίνει τη δομή του δικτύου και $X_1, X_2, \dots, X_n$ τυχαίες μεταβλητές, τότε η από κοινού συνάρτηση πιθανότητας θα είναι η $P(X_1, X_2, \dots, X_n)$. Σύμφωνα με τον πολλαπλασιαστικό τύπο (chain rule), μπορεί να γραφεί ως εξής:

$$P(X_1, X_2, \dots, X_n) = P(X_1)P(X_2 \mid X_1) \dots P(X_n \mid X_1, X_2, \dots, X_{n-1})$$

$$P(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid X_1, X_2, \dots, X_{i-1})$$

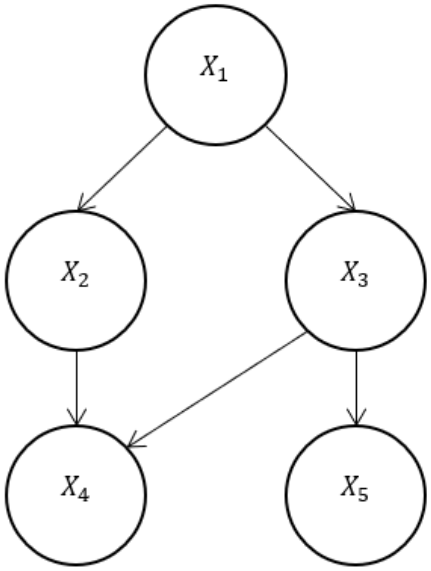
Γνωρίζουμε όμως από την τοπική ιδιότητα του Markov ότι κάθε τυχαία μεταβλητή είναι ανεξάρτητη από εκείνες τις μεταβλητές του δικτύου που δεν είναι απόγονοί της υπό τη συνθήκη των γονιών της, δηλαδή:

$$P(X_i \mid X_1, X_2, \ldots, X_{i-1}) = P(X_i \mid pa_{X_i})$$

όπου $pa_{X_i}$ συμβολίζει το σύνολο των γονέων της $X_i$. Άρα, η από κοινού συνάρτηση πιθανότητας του δικτύου πλέον γράφεται ως εξής:

$$P(X_1, X_2, \ldots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa_{X_i})$$

Για παράδειγμα, έστω ότι έχουμε το γράφημα τύπου DAG της Εικόνας 17, που αποτελεί ένα μπεϋζιανό δίκτυο με τυχαίες μεταβλητές τις $X_1, X_2, \ldots, X_5$.




**Εικόνα 17. Παράδειγμα DAG για την εξαγωγή της από κοινού συνάρτησης πιθανότητας του δικτύου.**

Τότε, αν $P$ είναι η από κοινού συνάρτηση πιθανότητας του μπεϋζιανού δικτύου θα έχουμε:

$$P(X_1, X_2, \ldots, X_5) = \prod_{i=1}^{5} P(X_i \mid X_1, X_2, \ldots, X_{i-1})$$

όπου πλέον μπορεί να γραφτεί ως εξής:

$$P(X_1, X_2, \ldots, X_5) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_1)P(X_4 \mid X_2, X_3)P(X_5 \mid X_3)$$

Συνεπώς, οι υπό συνθήκη ανεξαρτησίες των τυχαίων μεταβλητών $X_i$ που υποδηλώνει η δομή G του μπεϋζιανού δικτύου αποτυπώνονται στη συγκεκριμένη παραγοντοποίηση της από κοινού συνάρτησης πιθανότητας.

Η συγκεκριμένη παραγοντοποίηση απλοποιεί σημαντικά τους υπλογισμούς των πιθανοτήτων που θα έρπεπε να γίνουν σε σχέση με τον αρχικό τύπο. Για παράδειγμα, αν θεωρήσουμε ότι κάθε τυχαία μεταβλητή $X_i$ έχει μόνο δύο δυνατές τιμές (καταστάσεις) τότε θα έπρεπε να βρεθούν $2^n$ το πλήθος πιθανότητες. Αντίθετα, αν λάβουμε υπόψιν την καινούργια παραγοντοποίηση και αν υποθέσουμε ότι κάθε $X_i$ έχει $\kappa$ το πλήθος γονείς τότε θα πρέπει να βρεθούν $n * 2^{\kappa}$ το πλήθος πιθανότητες. Αυτό συνεπάγεται ότι για σχετικά μικρό πλήθος γονέων οι πιθανότητες που θα πρέπει να εκτιμηθούν είναι λιγότερες. Για παράδειγμα, αν είχαμε $n = 10$ και $\kappa = 3$ τότε στην πρώτη περίπτωση θα πρέπει να υπολογιστούν $2^{10} = 1024$ πιθανότητες, ενώ στην δεύτερη περίπτωση μόνο $10 * 2^3 = 80$ πιθανότητες.

Στην περίπτωση των διακριτών μπεϋζιανών δικτύων, σε κάθε τυχαία μεταβλητή $X_i$ με γονείς $pa_{X_i}$ αντιστοιχεί ένας πίνακας δεσμευμένης συνάρτησης πιθανότητας που αφορά τη δεσμευμένη τυχαία μεταβλητή $X_i \mid pa_{X_i}$. Οι συγκεκριμένοι πίνακες δεσμευμένων κατανομών είναι γνωστοί ως conditional probability tables (CPTs). Για παράδειγμα, ας υποθέσουμε ότι έχουμε ένα μπεϋζιανό δίκτυο με τυχαίες μεταβλητές $X_1, X_2, X_3$ και δομή G τέτοια ώστε να συνδεόνται σε σειρά $X_1 \to X_2 \to X_3$. Με τις τυχαίες μεταβλητές να παίρνουν μόνο δύο δυνατές τιμές η κάθεμια όπου $X_1 = \{x_1^0, x_1^1\}$, $X_2 = \{x_2^0, x_2^1\}$ και $X_3 = \{x_3^0, x_3^1\}$, η από κοινού συνάρτηση μάζας πιθανότητας του δικτύου θα είναι:

$$P(X_1, X_2, X_3) = \prod_{i=1}^{3} P(X_i \mid pa_{X_i}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)$$

και τα CPTs για την κάθε μεταβλητή του δικτύου θα είναι αυτά που παρουσιάζονται στους Πίνακες 1-3.

**Πίνακας 1. Πίνακας CPT για τον κόμβο $X_1$του παραδείγματος σύνδεσης $X_1 \rightarrow X_2 \rightarrow X_3$.**

| $X_1$ | $P(X_1)$ |
|---|---|
| $x_1^0$ | $P(x_1^0)$ |
| $x_1^1$ | $P(x_1^1)$ |

**Πίνακας 2. Πίνακας CPT για τον κόμβο $X_2$του παραδείγματος σύνδεσης $X_1 \rightarrow X_2 \rightarrow X_3$.**

| $X_1$ | $X_2$ | $P(X_2 \mid X_1)$ |
|---|---|---|
| $x_1^0$ | $x_2^0$ | $P(x_2^0 \mid x_1^0)$ |
| $x_1^0$ | $x_2^1$ | $P(x_2^1 \mid x_1^0)$ |
| $x_1^1$ | $x_2^0$ | $P(x_2^0 \mid x_1^1)$ |
| $x_1^1$ | $x_2^1$ | $P(x_2^1 \mid x_1^1)$ |

**Πίνακας 3. Πίνακας CPT για τον κόμβο $X_3$του παραδείγματος σύνδεσης $X_1 \rightarrow X_2 \rightarrow X_3$.**

| $X_2$ | $X_3$ | $P(X_3 \mid X_2)$ |
|---|---|---|
| $x_2^0$ | $x_3^0$ | $P(x_3^0 \mid x_2^0)$ |
| $x_2^0$ | $x_3^1$ | $P(x_3^1 \mid x_2^0)$ |
| $x_2^1$ | $x_3^0$ | $P(x_3^0 \mid x_2^1)$ |
| $x_2^1$ | $x_3^1$ | $P(x_3^1 \mid x_2^1)$ |

Όπως μπορεί να παρατηρήσει κανείς, ο πρώτος πίνακας είναι απλά οι πιθανότητες για τις διάφορες τιμές της $X_1$, αφού δεν έχει γονείς και άρα δεν έχει κάποια εξάρτηση από άλλη μεταβλητή. Στις υπόλοιπες μεταβλητές $X_2$, $X_3$ αντιστοιχεί ένας CPT για την κάθεμια, που περιέχει όλους τους συνδυασμούς πιθανοτήτων $P(X_2 \mid X_1)$ και $P(X_3 \mid X_2)$.

Επίσης, θα πρέπει να σημειωθεί ότι σε κάθε CPT, για συγκεκριμένη τιμή των γονιών οι πιθανότητες πρέπει να αθροίζουν στο 1. Δηλαδή, θα πρέπει:

$$P(X_2 \mid X_1 = x_1^0) = P(x_2^0 \mid x_1^0) + P(x_2^1 \mid x_1^0) = 1$$

αφού η $X_2 \mid X_1 = x_1^0$ αποτελεί μια δεσμευμένη κατανομή με συνάρτηση μάζας πιθανότητας της $P(X_2 \mid X_1 = x_1^0)$.

Γενικά, αν $Val(X_i)$ και $Val(pa_{X_i})$ είναι το σύνολο των τιμών της $X_i$ και των γονέων $pa_{X_i}$ της $X_i$ αντίστοιχα, επειδή για κάθε διαφορετικό συνδυασμό τιμών των γονιών έχουμε μια δεσμευμένη κατανομή, θα ισχύει ότι:

$$\forall \alpha \in Val(pa_{X_i}), \qquad P(X_i \mid pa_{X_i} = \alpha) = \sum_{x_i \in Val(X_i)} P(x_i \mid pa_{X_i} = \alpha) = 1$$

Συνεπώς, έχοντας την παραγοντοποίηση της από κοινού συνάρτησης πιθανότητας και γνωρίζοντας τα CPTs, μπορούμε να βρούμε την πιθανότητα κάθε δυνατής κατάστασης των μεταβλητών. Δηλαδή, αν έχουμε $X_1 = x_1^0,\ X_2 = x_2^0$ και $X_3 = x_3^0$, τότε πιθανότητα του να συμβεί αυτό το ενδεχόμενο δίνεται από την από κοινού συνάρτησης πιθανότητας του δικτύου ως εξής:

$$P(X_1 = x_1^0, X_2 = x_2^0, X_3 = x_3^0) = P(x_1^0)P(x_2^0 \mid x_1^0)P(x_3^0 \mid x_2^0)$$

όπου οι πιθανότητες $P(x_1^0),\ P(x_2^0 \mid x_1^0)$ και $P(x_3^0 \mid x_2^0)$ είναι γνωστές από τα CPTs.

Εκμεταλλευόμενοι αυτή τη δυνατότητα των μπεϋζιανών δικτύων μπορούμε γενικά να διατυπώσουμε ερωτήσεις για την πιθανότητα των καταστάσεων διάφορων μεταβλητών υπό την γνώση ότι κάποιες άλλες έχουν ήδη λάβει συγκεκριμένες τιμές. Η συγκεκριμένη διαδικασία ονομάζεται μπεϋζιανή συμπερασματολογία (Jensen and Nielsen, 2007; Koller and Friedman, 2009; Pearl, 2009).

## 4.2 Παραμετροποίηση Μπεϋζιανού Δικτύου

Γνωρίζουμε ότι σε κάθε μεταβλητή, ενός διακριτού μπεϋζιανού δικτύου, αντιστοιχεί και ένα CPT, οι πιθανότητες του οποίου αρχικά είναι άγνωστες. Για να βρεθούν, το πρώτο βήμα είναι η παραμετροποίηση του δικτύου, που σημαίνει ότι για κάθε πιθανότητα που εμφανίζεται στα CPTs θα πρέπει να αντιστοιχεί μια παράμετρος.

Για κάθε μεταβλητή $X$ του δικτύου και των γονιών της $\boldsymbol{U}$, δηλαδή σε κάθε CPT, αντιστοιχεί ένα σύνολο παραμέτρων $\boldsymbol{\theta}_{X \mid \boldsymbol{U}}$. Το σύνολο αυτό με τη σειρά του θα περιέχει τις παραμέτρους $\theta_{x \mid \boldsymbol{u}}$ που αντιστοιχούν στη δεσμευμένη πιθανότητα $P(x \mid \boldsymbol{u})$, με $x$ να είναι μια τιμή από το σύνολο τιμών $Val(X)$ της $X$ και $\boldsymbol{u}$ ένας συνδυασμός τιμών των γονιών της $X$ από το σύνολο τιμών των γονιών $Val(\boldsymbol{U})$. Επιπλέον, θα πρέπει για κάθε συνδυασμό $\boldsymbol{u}$ των γονιών:

$$\sum_{x} \theta_{x \mid \boldsymbol{u}} = 1$$

αφού αντιστοιχούν στη δεσμευμένη πιθανότητα $P(x \mid \boldsymbol{u})$. Αν $\boldsymbol{\theta}$ είναι το σύνολο όλων των $\boldsymbol{\theta}_{X \mid \boldsymbol{U}}$, τότε λέγεται ότι αποτελεί την παραμετροποίηση του μπεϋζιανού δικτύου και περιέχει το σύνολο των παραμέτρων του δικτύου.

Με τα παραπάνω δεδομένα, αν $\boldsymbol{x}$ είναι ένα συνδυασμός τιμών του συνόλου των μεταβλητών $\boldsymbol{X}$ του δικτύου, η πιθανότητα του $\boldsymbol{x}$ μέσα από την από κοινού συνάρτηση πιθανότητας του δικτύου μπορεί να γραφτεί ως γινόμενο των παραμέτρων $\boldsymbol{\theta}_{x \mid \boldsymbol{u}}$ για τα $x \mid \boldsymbol{u}$ που αντιστοιχούν στον συνδυασμό $\boldsymbol{x}$ :

$$P(\boldsymbol{x}) = \prod_{x \mid \boldsymbol{u}} \boldsymbol{\theta}_{x \mid \boldsymbol{u}}$$

Επαναφέροντας πάλι το απλό παράδειγμα ενός μπεϋζιανού δικτύου με δομή $X_1 \rightarrow X_2 \rightarrow X_3$, με τις τυχαίες μεταβλητές να παίρνουν τις τιμές $X_1 = \{x_1^0, x_1^1\}$, $X_2 = \{x_2^0, x_2^1\}$, $X_3 = \{x_3^0, x_3^1\}$ και την από κοινού συνάρτηση πιθανότητας να είναι:

$$P(X_1, X_2, X_3) = \prod_{i=1}^{3} P(X_i \mid pa_{X_i}) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)$$

μπορούμε πλέον να το παραμετροποιήσουμε. Στη περίπτωση αυτή, τα CPTs του δικτύου πλέον θα είναι αυτά που δίνονται στους Πίνακες 4-6.

**Πίνακας 4. Πίνακας CPT για τον κόμβο $X_1$του παραδείγματος σύνδεσης $X_1 \rightarrow X_2 \rightarrow X_3$.**

| $X_1$ | $\boldsymbol{\theta}_{X_1}$ |
|---|---|
| $x_1^0$ | $\theta_{x_1^0}$ |
| $x_1^1$ | $\theta_{x_1^1}$ |

**Πίνακας 5. . Πίνακας CPT για τον κόμβο $X_2$του παραδείγματος σύνδεσης $X_1 \rightarrow X_2 \rightarrow X_3$.**

| $X_1$ | $X_2$ | $\boldsymbol{\theta}_{X_2 \mid X_1}$ |
|---|---|---|
| $x_1^0$ | $x_2^0$ | $\theta_{x_2^0 \mid x_1^0}$ |
| $x_1^0$ | $x_2^1$ | $\theta_{x_2^1 \mid x_1^0}$ |
| $x_1^1$ | $x_2^0$ | $\theta_{x_2^0 \mid x_1^1}$ |
| $x_1^1$ | $x_2^1$ | $\theta_{x_2^1 \mid x_1^1}$ |

**Πίνακας 6. . Πίνακας CPT για τον κόμβο $X_3$του παραδείγματος σύνδεσης $X_1 \rightarrow X_2 \rightarrow X_3$.**

| $X_2$ | $X_3$ | $\boldsymbol{\theta}_{X_3 \mid X_2}$ |
|---|---|---|
| $x_2^0$ | $x_3^0$ | $\theta_{x_3^0 \mid x_2^0}$ |
| $x_2^0$ | $x_3^1$ | $\theta_{x_3^1 \mid x_2^0}$ |
| $x_2^1$ | $x_3^0$ | $\theta_{x_3^0 \mid x_2^1}$ |
| $x_2^1$ | $x_3^1$ | $\theta_{x_3^1 \mid x_2^1}$ |

Το σύνολο των παραμέτρων $\boldsymbol{\theta}$ του δικτύου θα είναι:

$$\boldsymbol{\theta} = (\boldsymbol{\theta}_{X_1}, \boldsymbol{\theta}_{X_2 \mid X_1}, \boldsymbol{\theta}_{X_3 \mid X_2})$$

με:

$\boldsymbol{\theta}_{X_1} = \left(\theta_{x_1^0}, \theta_{x_1^1}\right)$, $\boldsymbol{\theta}_{X_2 \mid X_1} = (\theta_{x_2^0 \mid x_1^0}, \theta_{x_2^1 \mid x_1^0}, \theta_{x_2^0 \mid x_1^1}, \theta_{x_2^1 \mid x_1^1})$ και

$\boldsymbol{\theta}_{X_3 \mid X_2} = (\theta_{x_3^0 \mid x_2^0}, \theta_{x_3^1 \mid x_2^0}, \theta_{x_3^0 \mid x_2^1}, \theta_{x_3^1 \mid x_2^1})$

Στην περίπτωση που θέλουμε να βρούμε την πιθανότητα $X_1 = x_1^0$, $X_2 = x_2^0$ και $X_3 = x_3^0$ θα είναι το εξής γινόμενο παραμέτρων:

$$P(x_1^0, x_2^0, x_3^0) = \theta_{x_1^0}\, \theta_{x_2^0 \mid x_1^0}\, \theta_{x_3^0 \mid x_2^0}$$

Έχοντας πλέον παραμετροποιήσει ένα μπεϋζιανό δίκτυο, το επόμενο βήμα είναι να βρεθούν εκείνες οι παράμετροι που ταιριάζουν καλύτερα στα δεδομένα, δηλαδή να βρεθεί το πιο ταιριαστό μοντέλο για το δίκτυο (fitting). Η διαδικασία αυτή λέγεται εύρεση παραμέτρων (parameter learning) και στη συνέχεια θα αναλυθούν δύο βασικές μέθοδοι εύρεσης παραμέτρων (Darwiche, 2009).

## 4.3 Εύρεση Παραμέτρων

Στο συγκεκριμένο υποκεφάλαιο, η ανάλυση που θα ακολουθήσει αφορά τις μεθόδους εύρεσης των παραμέτρων (parameter learning) ενός μπεϋζιανού δικτύου και γίνεται στηριζόμενη σε δύο βασικές υποθέσεις. Η πρώτη, είναι το γεγονός ότι αναφερόμαστε σε ένα διακριτό μπεϋζιανό δίκτυο, δηλαδή οι τυχαίες μεταβλητές του δικτύου είναι διακριτές. Η δεύτερη υπόθεση αφορά το γεγονός ότι έχουμε πλήρη δεδομένα, που σημάινει ότι γνωρίζουμε όλες τις παρατηρήσεις μας και δεν παρουσιάζονται κενά.

Οι δύο βασικές μέθοδοι για την εύρεση των παραμέτρων σε ένα μπεϋζιανό δίκτυο είναι η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας (ΕΜΠ) ή αλλιώς maximum likelihood estimation (MLE) και η μπεϋζιανή εκτίμηση παραμέτρων ή αλλιώς bayesian parameter estimation.

Η μέθοδος εκτίμησης μέγιστης πιθανοφάνειας βασίζεται στη κλασική στατιστική όπου θεωρεί ότι η θ είναι μια παράμετρος όπου μπορεί να πάρει διάφορες τιμές. Η μπεϋζιανή εκτίμηση, όμως, που βασίζεται στη μπεϋζιανή στατιστική, θεωρεί ότι η θ αποτελεί μια τυχαία μεταβλητή με δική της κατανομή.

Για την καλύτερη κατανόηση των μεθόδων είναι απαραίτητο να εξηγηθεί η έννοια του τυχαίου δείγματος. Ας υποτεθεί λοιπόν, οτι **X** είναι ένας πληθυσμός που ακολουθεί μια κατανομή $F$. Χαρακτηριστικά του πληθυσμού όπως για παράδειγμα ύψος, εισόδημα, αριθμός παιδιών και πολλά άλλα θεωρούνται τυχαίες μεταβλητές που είναι ανεξάρτητες και ισόνομες, όπου το ισόνομες δηλώνει ότι ακολουθούν όλες την ίδια κατανομή. Συνεπώς, τα χαρακτηριστικά $X_1, X_2, \ldots, X_n$ ενός πληθυσμού **X** αποτελούν ανεξάρτητες και ισόνομες τυχαίες μεταβλητές που ακολουθούν την ίδια κατανομή και μάλιστα την κατανομή που ακολουθεί ο **X**. Οι τιμές των τυχαίων μεταβλητών που παρατηρούνται καλούνται δεδομένα (Κοκολάκης και Φουσκάκης, 2009).

### *4.3.1 Μέθοδος Εκτίμησης Μέγιστης Πιθανοφάνειας (ΕΜΠ)*

Σύμφωνα με τη μέθοδο εκτίμησης μέγιστης πιθανοφάνειας, δοθέντος μιας συνάρτησης (μάζας ή πυκνότητας) πιθανότητας, δημιουργείται μια συνάρτηση που λέγεται συνάρτηση πιθανοφάνειας, η οποία εξαρτάται από μια παράμετρο θ ή ένα διάνυσμα πραμέτρων **θ**. Η παράμετρος θ ή το διάνυσμα παραμέρων **θ** είναι ποσότητες σταθερές αλλά άγνωστες. Η μεγιστοποίηση της πιθανοφάνειας έχει ως αποτέλεσμα την εύρεση

των παραμέτρων αυτών και άρα της κατανομής που ταιριάζει περισσότερο στα υπάρχοντα δεδομένα.

Έστω ένα τυχαίο δείγμα $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, ένα άγνωστο διάνυσμα των παραμέτρων $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ με $\boldsymbol{\theta} \in \Theta$, όπου $\Theta$ ο χώρος των παραμέτρων και p(**x**;**θ**) η από κοινού συνάρτηση (μάζας ή πυκνότητας) πιθανότητας των $X_1, X_2, \ldots, X_n$, με:

$$P(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} P(x_i; \boldsymbol{\theta}) = P(x_1; \boldsymbol{\theta}) P(x_2; \boldsymbol{\theta}) \ldots P(x_n; \boldsymbol{\theta})$$

Η παραγοντοποίηση της $p(\boldsymbol{x}; \boldsymbol{\theta})$ με το συγκεκριμένο τρόπο οφείλεται στο γεγονός ότι το $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ είναι ένα τυχαίο δείγμα και άρα οι $X_1, X_2, \ldots, X_n$ είναι ανεξάρτητες. Επίσης, ο συμβολισμός «**;**» στην $p(\boldsymbol{x}; \boldsymbol{\theta})$ υποδηλώνει ότι η κατανομή που θα ακολουθεί η $\boldsymbol{X}$ εξαρτάται από τις τιμές της παραμέτρου $\boldsymbol{\theta}$.

Όταν είναι γνωστές οι τιμές $x_i$ $(i = 1, \ldots, n)$ των τυχαίων μεταβλητών $X_i$ $(i = 1, \ldots, n)$, η συνάρτηση πιθανότητας p(**x**;**θ**) μπορεί να θεωρηθεί ως συνάρτηση της παραμέτρου **θ**, δηλαδή:

$$L(\boldsymbol{\theta}) = P(\boldsymbol{x}; \boldsymbol{\theta}) = \prod_{i=1}^{n} P(x_i; \boldsymbol{\theta})$$

όπου η συνάρτηση L(**θ**) ονομάζεται συνάρτηση πιθανοφάνειας (likelihood function) και εκφράζει το πόσο σύμφωνες με τις $X_1, X_2, \ldots, X_n$ είναι οι διάφορες τιμές της παραμέτρου **θ**, ή αλλιώς, πόσο πιθανό είναι να έχουν παραχθεί αυτά τα δεδομένα από την ορισμένη κατανομή με παράμετρο **θ**.

Σκοπός της μεθόδου είναι να βρεθεί ένα συγκεκριμένο διάνυσμα τιμών $\widehat{\boldsymbol{\theta}} = (\hat{\theta}_1, \hat{\theta}_2, \ldots, \hat{\theta}_m)$ των παραμέτρων $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ που να μεγιστοποιεί την συνάρτηση πιθανοφάνειας L(**θ**) ή ισοδύναμα τον νεπέριο λογάριθμο της L(**θ**) (loglikelihood). Δηλαδή αν Θ ο χωρος των παραμέτρων τότε

$$L(\widehat{\boldsymbol{\theta}}) = sup_{\theta \in \Theta} L(\boldsymbol{\theta})$$

ή ισοδύναμα

$$l(\widehat{\boldsymbol{\theta}}) = sup_{\theta \in \Theta} l(\boldsymbol{\theta})$$

με $l(\boldsymbol{\theta}) = \ln L(\boldsymbol{\theta})$, $\boldsymbol{\theta} \in \Theta$.

Αυτό συμβάινει διότι ο μετασχηματισμός της $L(\boldsymbol{\theta})$ μέσω του νεπέριου λογάριθμου, που είναι μια αύξουσα συνάρτηση, μας δίνει ως αποτέλεσμα μια αύξουσα συνάρτηση, την $l(\boldsymbol{\theta})$, η οποία μεγιστοποιείται για το ίδιο $\boldsymbol{\theta}$ που μεγιστοποιείται και η $L(\boldsymbol{\theta})$. Λύνοντας το σύστημα:

$$\frac{\partial l(\boldsymbol{\theta})}{\partial \theta_j} = 0 \ \ (j = 1, \dots, m)$$

θα βρεθεί το $\widehat{\boldsymbol{\theta}}$ για το οποίο η συνάρτηση πιθανοφάνειας L($\boldsymbol{\theta}$) παρουσιάζει αρκρότατο. Για να είναι πράγματι το $\widehat{\boldsymbol{\theta}}$ σημείο μεγίστου θα πρέπει ο Εσσιανός πίνακας:

$$\left[\frac{\partial^2 l(\boldsymbol{\theta})}{\partial \theta_i \partial \theta_j}\right]_{mxm} \ \ (i, j = 1, \dots, m)$$

να είναι αρνητικά ορισμένος για $\boldsymbol{\theta} = \widehat{\boldsymbol{\theta}}$ (Κοκολάκης και Φουσκάκης, 2009).

#### *4.3.1.1 Η ΕΜΠ σε Μπεϋζιανό Δίκτυο και η Ολική Αποσύνθεση Πιθανοφάνειας*

Ας υποτεθεί ότι έχουμε ένα μπεϋζιανό δίκτυο με δομή G, τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$ , D το σύνολο των δεδομένων και $P_G$ η από κοινού συνάρτηση (μάζας ή πυκνότητας) πιθανότητας του μπεϋζιανού δικτύου. Τότε σύμφωνα με τη γνωστή παραγοντοποίηση της από κοινού συνάρτησης πιθανότητας ενός μπεϋζιανού δικτύου έχουμε:

$$P_G(X_1, X_2, \dots, X_n) = \prod_{i=1}^{n} P(X_i \mid pa_{X_i})$$

Στον παρακάτω πίνακα (Πίνακας 7) φαίνεται η κατανομή του συνόλου των δεδομένων D και κάθε γραμμή του πίνακα ονομάζεται στιγμιότυπο των δεδομένων D.

**Πίνακας 7. Παρουσίαση της λογικής δεδομένων πολλών μεταβλητών και πολλών στιγμιοτύπων.**

| D\X | $X_1$ ..... $X_n$ |
|---|---|
| d[1] | $x_1[1]$ ..... $x_n[1]$ |
| . | . ..... . |
| . | . ..... . |
| d[m] | $x_1[m]$ ..... $x_n[m]$ |
| . | . ..... . |
| . | . ..... . |
| d[M] | $x_1[M]$ ..... $x_n[M]$ |

Όπου M είναι το σύνολο των στιγμιότυπων των δεδομένων, με d[m] να είναι το κάθε στιγμιότυπο για $1 \leq m \leq M$ και $x_i[m]$ υποδηλώνει την τιμή της τυχαίας μεταβλητής $X_i$ στο στιγμιότυπο m με $1 \leq i \leq n$.

Τότε, αν **θ** το διάνυσμα των παραμέτρων και υπό την υπόθεση ότι τα στιγμιότυπα d[m] είναι ανεξάρτητα, δηλαδή ότι κάθε στιγμιότυπο αποτελεί ένα ανεξάρτητο δείγμα, η συνάρτηση πιθανοφάνειας του μπεϋζιανου δικτύου θα είναι:

$$L(\boldsymbol{\theta}) = \prod_{m=1}^{M} P_G(d[m]; \boldsymbol{\theta})$$

$$= \prod_{m=1}^{M} P_G(x_1[m], \dots, x_n[m]; \boldsymbol{\theta})$$

$$= \prod_{m=1}^{M} P\left(x_1[m] \mid pa_{X_1}[m]; \boldsymbol{\theta}\right) * \dots * P\left(x_n[m] \mid pa_{X_n}[m]; \boldsymbol{\theta}\right) \quad (1)$$

$$= \prod_{m=1}^{M} \prod_{i=1}^{n} P\left(x_i[m] \mid pa_{X_i}[m]; \boldsymbol{\theta}\right)$$

$$= \prod_{i=1}^{n} \left[ \prod_{m=1}^{M} P\left(x_i[m] \mid pa_{X_i}[m]; \boldsymbol{\theta}\right) \right]$$

Κάθε όρος μέσα στις αγκύλες ονομάζεται δεσμευμένη πιθανοφάνεια (conditional likelihood) της τυχαίας μεταβλητής $X_i$ δεδομένου των γονιών της. Ωστόσο, η δεσμευμένη πιθανότητα $P(X_i \mid pa_{X_i}; \boldsymbol{\theta})$ δεν καθορίζεται από το σύνολο των παραμέτρων $\boldsymbol{\theta}$, αλλά από ένα υποσυνολό τους που συμβολίζουμε με $\boldsymbol{\theta}_{X_i \mid pa_{X_i}}$. Δηλαδή, η $P(X_i \mid pa_{X_i}; \boldsymbol{\theta})$ εξαρτάται από το υποσύνολο των παραμέτρων της τυχαίας μεταβλητής $X_i$ δεδομένου των γονιών της. Τότε, η συνάρτηση:

$$L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right) = P(D; \boldsymbol{\theta}_{X_i|pa_{X_i}}) = \prod_{m=1}^{M} P\left(x_i[m] \mid pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$$

ονομάζεται τοπική συνάρτηση πιθανοφάνειας (local likelihood function) της τυχαίας μεταβλητής $X_i$.

Επιπλέον, μπορεί να παρατηρήσει κανείς ότι το γινόμενο των πιθανοτήτων μέσα στις αγκύλες αφορά τη στήλη i του πίνακα των δεδομένων καθώς, το m διατρέχει τα στιγμιότυπα των δεδομένων της $X_i$, δηλαδή τις γραμμές. Αντιθέτως, το γινόμενο που βρίσκεται έξω από τις αγκύλες διατρέχει τις στήλες του πίνακα και άρα τις διάφορες τυχαίες μεταβλητές. Δηλαδή, το γινόμενο των δεσμευμένων πιθανοτήτων $P\left(x_i[m] \mid pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$ κάθε στήλης είναι η τοπική πιθανοφάνεια $L_i$ της κάθε τυχαίας μεταβλητής $X_i$ και έπειτα το γινόμενο των αποτελεσμάτων των στηλών είναι η πιθανοφάνεια $L(\theta)$. Η συνάρτηση πιθανοφάνειας μπορεί πλέον να γραφτεί ως εξής:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right) = \prod_{i=1}^{\boldsymbol{n}} P(D; \boldsymbol{\theta}_{X_i|pa_{X_i}})$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{m=1}^{M} P\left(x_i[m] \mid pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$$

Άρα, η συνάρτηση πιθανοφάνειας $L(\boldsymbol{\theta})$ είναι γινόμενο των τοπικών συναρτήσεων πιθανοφάνειας $L_i$ της κάθε τυχαίας μεταβλητής. Κάθε τοπική πιθανοφάνεια $L_i$ αντιστοιχεί στον πίνακα της δεσμευμένης συνάρτησης πιθανότητας (CPT) της $X_i$. Η διαδικάσία αυτή ονομάζεται ολική αποσύνθεση (global decomposition) της συνάρτησης πιθανοφάνειας. Επίσης, αν υποτεθεί ότι τα σύνολα των παραμέτρων $\boldsymbol{\theta}_{X_i|pa_{X_i}}$ των δεσμευμένων κατανομών $X_i|pa_{X_i}$ δεν έχουν κοινές παραμέτρους, δηλαδή $\boldsymbol{\theta}_{X_i|pa_{X_i}}$ και $\boldsymbol{\theta}_{X_j|pa_{X_j}}$ είναι ασύνδετα για κάθε $i \neq j$, με $i, j = 1, \dots, n$, τότε αν $\widehat{\boldsymbol{\theta}}_{X_i|pa_{X_i}}$ οι παράμετροι που μεγιστοποιούν την $L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$, το διάνυσμα παραμέτρων $\widehat{\boldsymbol{\theta}} = (\widehat{\boldsymbol{\theta}}_{X_1|pa_{X_1}}, \widehat{\boldsymbol{\theta}}_{X_2|pa_{X_2}}, \dots, \widehat{\boldsymbol{\theta}}_{X_n|pa_{X_n}})$ μεγιστοποιεί την πιθανοφάνεια $L(\boldsymbol{\theta})$.

Στο σημείο αυτό, θα πρέπει να τονιστεί το γεγονός πως η υπόθεση της ασυνδεσιμότητας των παραμέτρων είναι μια υπόθεση που δεν ισχύει πάντα, διότι υπάρχουν πολλές περιπτώσεις στην πραγματικότητα που διάφορες παράμετροι είναι κοινές σε ένα μπεϋζιανό δίκτυο. Συνεπώς, υπό την υπόθεση ότι οι παράμετροι δεν συνδέονται, το πρόβλημα της μεγιστοποίησης της πιθανοφάνειας $L(\boldsymbol{\theta})$ ανάγεται σε τοπικά προβήματα μεγιστοποίησης των τοπικών πιθανοφανειών $L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$ της κάθε τυχαίας μεταβλητής (Koller and Friedman, 2009).

#### *4.3.1.2 Η ΕΜΠ σε Μπεϋζιανό Δίκτυο και η Τοπική Αποσύνθεση Πιθανοφάνειας*

Στο σημείο αυτό, σύμφωνα με τη προϋπόθεση ότι ισχύει η προηγούμενη ανάλυση, δηλαδή ότι έχουμε ολική αποσύνθεση της πιθανοφάνειας και τις παραμέτρους των δεσμευμένων κατονομών ασύνδετες, γνωρίζουμε ότι η συνάρτηση πιθανοφάνειας είναι το γινόμενο των τοπικών πιθανοφανειών, κάθε τοπική πιθανοφάνεια αντιστοιχεί σε ένα CPT του μπεϋζιανού δικτύου και η μεγιστοποίηση της πιθανοφάνειας γίνεται μέσω της μεγιστοποίησης των τοπικών πιθανοφανειών. Έχοντας αυτή τη βάση η πιθανοφάνεια $L(\boldsymbol{\theta})$ μπορεί να αποσυντεθεί σε ακομά μικρότερα μέρη.

Για την καλύτερη κατανόηση και περισσότερο ευδιάκριτη καταγραφή των τύπων θα γίνει προσωρινή αλλαγή των συμβολισμών. Ας υποθέσουμε ότι έχουμε μια τυχαία μεταβλητή X, από το σύνολο των τυχαίων μεταβλητών του μπεϋζιανού δικτύου και **U** να είναι το σύνολο των γονέων της X. Τότε στη δεσμευμένη κατανομή X | **U** και στην τοπική πιθανοφάνειά της $L_X(\theta_{X\,|\,\boldsymbol{U}})$ θα αντιστοιχεί ένα CPT, με δεσμευμένη συνάρτηση μάζας πιθανότητας $P(X \mid \boldsymbol{U})$, που θα περιέχει τις παραμέτρους $\theta_{x\,|\,\boldsymbol{u}}$. Όπου $x[m] = x \in Val(X)$, με $Val(X)$ να είναι το σύνολο των τιμών που μπορεί να πάρει η

τυχαία μεταβλητή X και $pa_X[m] = \boldsymbol{u} \in Val(\boldsymbol{U})$ με $Val(\boldsymbol{U})$ να είναι το σύνολο των τιμών που μπορεί να πάρουν οι γονείς **U** της X. Το CPT που αντιστοιχεί στην $X \mid \boldsymbol{U}$ παρουσιάζεται στον πίνακα 8.

**Πίνακας 8. Εννοιολογική παρουσίαση CPT.**

| **U** | X | $\theta_{x \mid \boldsymbol{u}}$ |
|---|---|---|
| $Val(\boldsymbol{U})$ | $Val(X)$ | Οι παράμετροι δεδομένου των συνδυασμών των τιμών των X και **U** |

Σε αντιστοιχία με τους προηγούμενους συμβολισμούς η $L_X(\theta_{X \mid \boldsymbol{U}})$ θα είναι το αντίστοιχο της $L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$, η $\theta_{x[m] \mid \boldsymbol{u}[m]}$ είναι το αντίστοιχο της $P\left(x_i[m] \mid pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$ και η $\theta_{x \mid \boldsymbol{u}}$ είναι το αντίστοιχο της $P\left(x_i[m] = x \mid pa_{X_i}[m] = \boldsymbol{u}; \boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$, όπου έχουμε συγκεκριμένο συνδυασμό τιμών της μεταβλητής και των γονιών της στο m στιγμιότυπο.

Αν $M[\boldsymbol{u}, x]$ είναι ο αριθμός που δείχνει το πόσες φορές εμφανίζονται οι τιμές $x[m] = x$ και $\boldsymbol{u}[m] = \boldsymbol{u}$ για τα διάφορα στιγμιότυπα των δεδομένων D με $m = 1, \dots, M$, τότε η τοπική πιθανοφάνεια $L_X(\theta_{X \mid \boldsymbol{U}})$ της τυχαίας μεταβλητής X θα είναι:

$$L_X(\theta_{X \mid \boldsymbol{U}}) = P(D; \theta_{X \mid \boldsymbol{U}}) = \prod_{m=1}^{M} \theta_{x[m] \mid \boldsymbol{u}[m]}$$

$$= \theta_{x[1] \mid \boldsymbol{u}[1]}\, \theta_{x[2] \mid \boldsymbol{u}[2]} \dots\, \theta_{x[M] \mid \boldsymbol{u}[M]}$$

$$= \prod_{\boldsymbol{u} \in Val(\boldsymbol{U})} \left[ \prod_{x \in Val(X)} \theta_{x \mid \boldsymbol{u}}{}^{M[\boldsymbol{u},x]} \right]$$

Ο όρος που βρίσκεται μέσα στις αγκύλες αποτελεί μια απλή συνάρτηση πιθανοφάνειας (simple likelihood function) της X | **U** με δεδομένο ότι **U**=**u**. Άρα κάθε τοπική

πιθανοφάνεια που αντιστοιχεί σε ένα CPT μπορεί να αποσυντεθεί σε γινόμενο των απλών πιθανοφανειών. Δηλαδή:

$$L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right) = \prod_{\boldsymbol{u_i}\in Val(pa_{X_i})}\left[\prod_{x_i\in Val(X_i)} \theta_{x_i\,|\,\boldsymbol{u_i}}{}^{M[\,\boldsymbol{u_i},x_i]}\right]$$

όπου:

$Val(pa_{X_i})$: είναι το σύνολο των δυνατών συνδυασμών των τιμών των γονέων της $X_i$.

$Val(X_i)$: είναι το σύνολο των δυνατών συνδυασμών των τιμών της $X_i$.

$\boldsymbol{u_i}$: ένας συνδυασμός τιμών από το $Val(X_i)$ των γονιών της $X_i$.

$x_i$: μια τιμή από το $Val(X_i)$.

$M[\,\boldsymbol{u_i}, x_i]$: το πλήθος των εμφανίσεων των τιμών $\boldsymbol{u_i} \in Val(pa_{X_i})$ και $x_i \in Val(X_i)$ της $X_i$, στα διάφορα στιγμιότυπα των δεδομένων.

Αυτό σημαίνει ότι η τοπική συνάρτηση πιθανοφάνειας της $X_i|pa_{X_i}$ είναι ένα γινόμενο των απλών συναρτήσεων πιθανοφάνειας. Σαν αποτέλεσμα έχουμε ότι, κάθε τοπική πιθανοφάνεια $L_i$ της τυχάιας μεταβλητής $X_i$ μπορεί να αποσυντεθεί με τον συγκεκριμένο τρόπο και άρα η πιθανοφάνεια $L(\boldsymbol{\theta})$ που είναι το γινόμενο των τοπικών πιθανοφανειών αποσυντίθεται ακόμα περαιτέρω. Η συγκεκριμένη διαδικασία καλείται τοπική αποσύνθεση (local decomposition) της συνάρτησης πιθανοφάνειας.

Υποθέτουμε ότι οι παράμετροι $\boldsymbol{\theta}_{X_i|pa_{X_i}=\boldsymbol{u_i}}$ για κάθε διαφορετική τιμή $\boldsymbol{u_i}$ των γονιών $pa_{X_i}$ είναι ανεξάρτητες μεταξύ τους. Δηλαδή, αν $\boldsymbol{u_i}, \boldsymbol{u_i}' \in Val(pa_{X_i})$ τότε τα σύνολα των παραμέτρων $\boldsymbol{\theta}_{X_i|pa_{X_i}=\boldsymbol{u_i}}$ και $\boldsymbol{\theta}_{X_i|pa_{X_i}=\boldsymbol{u_i}'}$ είναι ανεξάρτητα μεταξύ τους και έτσι η μεγιστοποίηση της απλής πιθανοφάνειας που αντιστοιχεί στο σύνολο $\boldsymbol{\theta}_{X_i|pa_{X_i}=\boldsymbol{u_i}}$ δεν επηρεάζει τη μεγιστοποίηση της απλής πιθανοφάνειας που αντιστοιχεί στο σύνολο $\boldsymbol{\theta}_{X_i|pa_{X_i}=\boldsymbol{u_i}'}$ και αντίστροφα. Καταλήγουμε συνεπώς στο συμπέρασμα, ότι μπορεί να μεγιστοποιηθεί ξεχωριστά η κάθε συνάρτηση απλής πιθανοφάνειας $\prod_{x_i\in Val(X_i)} \theta_{x_i\,|\,\boldsymbol{u_i}}{}^{M[\,\boldsymbol{u_i},x_i]}$. Μάλιστα, στην περίπτωση που κάθε μεταβλητή $X_i$ είναι μια πολυωνυμική μεταβλητή, τότε οι παράμετροι που μεγιστοποιούν την κάθε συνάρτηση απλής πιθανοφάνειας είναι:

$$\hat{\theta}_{x_i \mid \boldsymbol{u_i}} = \frac{M[\,\boldsymbol{u_i}, x_i]}{M[\,\boldsymbol{u_i}]}$$

όπου $M[\,\boldsymbol{u_i}]$ είναι το πλήθος εμφάνισης του συνδυασμού των γονέων $\boldsymbol{u_i}$ για όλες τις δυνατές τιμές $x_i$ της μεταβλητής $X_i$:

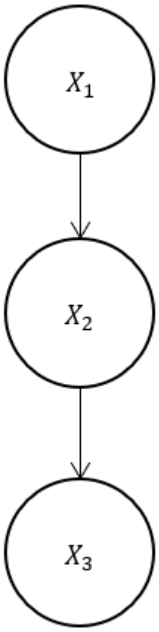
$$M[\,\boldsymbol{u_i}] = \sum_{x_i} M[\,\boldsymbol{u_i}, x_i]$$

δηλαδή είναι ο αριθμός εμφανίσεων του συνδυασμού $\boldsymbol{u_i}$ στα δεδομένα. Ο γενικός τύπος της συνάρτησης πιθανοφάνειας μπορεί πλέον μπορεί να γραφτεί ως εξής:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right) = \prod_{i=1}^{n}\prod_{m=1}^{M} P\left(x_i[m] \mid pa_{X_i}[m]; \boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$$

$$= \prod_{i=1}^{n} \prod_{\boldsymbol{u_i} \in Val(pa_{X_i})} \left[ \prod_{x_i \in Val(X_i)} \theta_{x_i \mid \boldsymbol{u_i}}{}^{M[\,\boldsymbol{u_i}, x_i]} \right]$$

Συνεπώς, η συνάρτηση πιθανοφάνειας $L(\boldsymbol{\theta})$ μεγιστοποιείται μέσω της μεγιστοποίησης της κάθε τοπικής πιθανοφάνειας $L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$**,** όπου κάθε τοπική πιθανοφάνεια μεγιστοποιείται μέσω της μεγιστοποίησης των απλών συναρτήσεων πιθανοφάνειας (Cowel *et al*. 1999; Koller and Friedman, 2009). Για παράδειγμα, ας υποθέσουμε ότι έχουμε ένα απλό μπεϋζιανό δίκτυο με μεταβλητές $X_1, X_2, X_3$ και τη δομή που παρουσιάζεται στην Εικόνα 18.




**Εικόνα 18. Παράδειγμα τριών κόμβων σε σύνδεση σε σειρά.**

τότε η από κοινού συνάρτηση μάζας πιθανότητας θα είναι:

$$P(X_1, X_2, X_3) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)$$

Οι τιμές που μπορούν να πάρουν οι μεταβλητές είναι $\{x_1^0, x_1^1\}$ για την $X_1$, $\{x_2^0, x_2^1\}$ για την $X_2$ και $\{x_3^0, x_3^1\}$ για την $X_3$. Επιπλέον έστω ότι έχουμε το σύνολο δεδομένων $D$ που παρουσιάζεται στον Πίνακα 9.

**Πίνακας 9. Πίνακας στιγμιοτύπων του παραδείγματος της Εικόνας 18.**

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| d[1] | $x_1^0$ | $x_2^0$ | $x_3^1$ |
| d[2] | $x_1^1$ | $x_2^0$ | $x_3^1$ |
| d[3] | $x_1^1$ | $x_2^1$ | $x_3^1$ |
| d[4] | $x_1^0$ | $x_2^0$ | $x_3^0$ |

Παραμετροποιώντας το δίκτυο, για την $X_1$ θα έχουμε ένα παραμετρικό πίνακα πιθανοτήτων(Πίνακας 10) και για τις $X_2$, $X_3$ τα παραμετρικά CPTs δίνονται στους Πίνακες 11 και 12.

**Πίνακας 10. Πίνακας CPT για τον κόμβο $X_1$του παραδείγματος της Εικόνας 18.**

| $X_1$ | $\boldsymbol{\theta}_{X_1}$ |
|---|---|
| $x_1^0$ | $\theta_{x_1^0}$ |
| $x_1^1$ | $\theta_{x_1^1}$ |

$+\rightarrow 1$

**Πίνακας 11. Πίνακας CPT για τον κόμβο $X_2$του παραδείγματος της Εικόνας 18.**

| $X_1$ | $X_2$ | $\boldsymbol{\theta}_{X_2 \mid X_1}$ |
|---|---|---|
| $x_1^0$ | $x_2^0$ | $\theta_{x_2^0 \mid x_1^0}$ |
| $x_1^0$ | $x_2^1$ | $\theta_{x_2^1 \mid x_1^0}$ |
| $x_1^1$ | $x_2^0$ | $\theta_{x_2^0 \mid x_1^1}$ |
| $x_1^1$ | $x_2^1$ | $\theta_{x_2^1 \mid x_1^1}$ |

$+\rightarrow 1$

$+\rightarrow 1$

**Πίνακας 12. Πίνακας CPT για τον κόμβο $X_3$του παραδείγματος της Εικόνας 18.**

| $X_2$ | $X_3$ | $\boldsymbol{\theta}_{X_3 \mid X_2}$ |
|---|---|---|
| $x_2^0$ | $x_3^0$ | $\theta_{x_3^0 \mid x_2^0}$ |
| $x_2^0$ | $x_3^1$ | $\theta_{x_3^1 \mid x_2^0}$ |
| $x_2^1$ | $x_3^0$ | $\theta_{x_3^0 \mid x_2^1}$ |
| $x_2^1$ | $x_3^1$ | $\theta_{x_3^1 \mid x_2^1}$ |

$+\rightarrow 1$

$+\rightarrow 1$

Άρα, για την πρώτη τοπική πιθανοφάνεια έχουμε:

$$L_1(\boldsymbol{\theta}_{X_1}) = P(D\,;\,\boldsymbol{\theta}_{X_1}) = \prod_{m=1}^{4} P(x_1[m]\,;\,\boldsymbol{\theta}_{X_1}) = \prod_{j=0}^{1} {\theta_{x_1^j}}^{M[x_1^j]}$$

όπου $M[x_1^j]$ το πλήθος εμφανίσεων της τιμής $x_1^j$ της $X_1$ στα δεδομένα $D$. Παρατηρούμε ότι στα δεδομένα η $x_1^0$ εμφανίζεται 2 φορές και η $x_1^1$ εμφανίζεται 2 φορές. Τότε:

$$L_1(\boldsymbol{\theta}_{X_1}) = \prod_{j=0}^{1} {\theta_{x_1^j}}^{M[x_1^j]} = {\theta_{x_1^0}}^2\,{\theta_{x_1^1}}^2$$

Για να βρούμε τις παραμέτρους που μεγιστοποιούν την $L_1$ μπορούμε να χρησιμοποιήσουμε τον τύπο της εύρεσης των παραμέτρων που μεγιστοποιούν τις απλές πιθανοφάνειες για πολυωνυμικές μεταβλητές:

$$\hat{\theta}_{x_i \,|\, \boldsymbol{u_i}} = \frac{M[\,\boldsymbol{u_i}, x_i]}{M[\,\boldsymbol{u_i}]}$$

Αυτό γιατί η περίπτωση των δύο δυνατών τιμών μιας μεταβλητής που έχουμε στο παράδειγμα μπορεί να θεωρηθεί ως μια πολυωνυμική μεταβλητή με δυο το πλήθος τιμές. Επίσης, στην περίπτωση που έχουμε μεταβλητή που δεν έχει γονείς ο τύπος γίνεται ως εξής:

$$\hat{\theta}_k = \frac{M[k]}{M}$$

όπου $M = 4$ το πλήθος των στιγμιοτύπων των δεδομένων και $M[k]$ ο αριθμός εμφάνισης της $k$ τιμής της μεταβλητής στα δεδομένα. Άρα, θα έχουμε:

$$\hat{\theta}_{x_1^0} = \frac{M[x_1^0]}{M} = \frac{2}{4} \quad \hat{\theta}_{x_1^1} = \frac{M[x_1^1]}{M} = \frac{2}{4}$$

ή

$$\hat{\theta}_{x_1^1} = 1 - \hat{\theta}_{x_1^0} = \frac{2}{4}$$

αφού

$$\hat{\theta}_{x_1^0} + \hat{\theta}_{x_1^1} = 1$$

Για τη δεύτερη τοπική πιθανοφάνεια που αντιστοιχεί στο CPT της $X_2$ έχουμε:

$$L_2\left(\boldsymbol{\theta}_{X_2 \mid X_1}\right) = P\left(D\,;\, \boldsymbol{\theta}_{X_2 \mid X_1}\right) = \prod_{m=1}^{4} P(x_2[m] \mid x_1[m]\,;\, \boldsymbol{\theta}_{X_2 \mid X_1}) =$$

$$= \prod_{j=0}^{1}\left[\prod_{k=0}^{1} \theta_{x_2^k \mid x_1^j}{}^{M[x_1^j,\, x_2^k]}\right] = \left[\theta_{x_2^0 \mid x_1^0}{}^2\, \theta_{x_2^1 \mid x_1^0}{}^0\right]\left[\,\theta_{x_2^0 \mid x_1^1}{}^1\, \theta_{x_2^1 \mid x_1^1}{}^1\right]$$

όπου στα δεδομένα ο συνδυασμός $x_2^0 \mid x_1^0$ εμφανίζεται δύο φορές, ο $x_2^1 \mid x_1^0$ δεν εμφανίζεται, ο $x_2^0 \mid x_1^1$ εμφανίζεται μια φορά και ο $x_2^1 \mid x_1^1$ επίσης μια φορά. Οι παράμετροι που μεγιστοποιούν την κάθε απλή πιθανοφάνεια που βρίσκεται σε αγκύλες για κάθε συνδυασμό των γονιών είναι:

$$\hat{\theta}_{x_2^0 \mid x_1^0} = \frac{M[x_1^0,\ x_2^0]}{M[x_1^0]} = \frac{2}{2} = 1 \quad \hat{\theta}_{x_2^1 \mid x_1^0} = \frac{M[x_1^0,\ x_2^1]}{M[x_1^0]} = \frac{0}{2} = 0$$

ή

$$\hat{\theta}_{x_2^1 \mid x_1^0} = 1 - \hat{\theta}_{x_2^0 \mid x_1^0} = 0$$

και

$$\hat{\theta}_{x_2^0 \mid x_1^1} = \frac{M[x_1^1,\ x_2^0]}{M[x_1^1]} = \frac{1}{2} \quad \hat{\theta}_{x_2^1 \mid x_1^1} = \frac{M[x_1^1,\ x_2^1]}{M[x_1^1]} = \frac{1}{2}$$

ή

$$\hat{\theta}_{x_2^1 \mid x_1^1} = 1 - \hat{\theta}_{x_2^0 \mid x_1^1} = \frac{1}{2}$$

με

$$\hat{\theta}_{x_2^0 \mid x_1^0} + \hat{\theta}_{x_2^1 \mid x_1^0} = 1 \quad \hat{\theta}_{x_2^0 \mid x_1^1} + \hat{\theta}_{x_2^1 \mid x_1^1} = 1$$

Τέλος, για την τρίτη τοπική πιθανοφάνεια που αντιστοιχεί στο CPT της $X_3$ έχουμε:

$$L_3\left(\boldsymbol{\theta}_{X_2 \mid X_1}\right) = P\left(D\ ;\ \boldsymbol{\theta}_{X_3 \mid X_2}\right) = \prod_{m=1}^{4} P(x_3[m] \mid x_2[m]\ ;\ \boldsymbol{\theta}_{X_3 \mid X_2}) =$$

$$= \prod_{j=0}^{1}\left[\prod_{k=0}^{1} \theta_{x_3^k \mid x_2^j}{}^{M[x_2^j,\, x_3^k]}\right] = \left[\theta_{x_3^0 \mid x_2^0}{}^{1}\ \theta_{x_3^1 \mid x_2^0}{}^{2}\right]\left[\ \theta_{x_3^0 \mid x_2^1}{}^{0}\ \theta_{x_3^1 \mid x_2^1}{}^{1}\right]$$

όπου στα δεδομένα ο συνδυασμός $x_3^0 \mid x_2^0$ εμφανίζεται μια φορά, ο $x_3^1 \mid x_2^0$ εμφανίζεται δύο φορές, ο $x_3^0 \mid x_2^1$ δεν εμφανίζεται και ο $x_3^1 \mid x_2^1$ εμφανίζεται μια φορά. Οι παράμετροι που μεγιστοποιούν την κάθε απλή πιθανοφάνεια που βρίσκεται σε αγκύλες για κάθε συνδυασμό των γονιών είναι:

$$\hat{\theta}_{x_3^0 \mid x_2^0} = \frac{M[x_2^0,\ x_3^0]}{M[x_2^0]} = \frac{1}{3} \qquad \hat{\theta}_{x_3^1 \mid x_2^0} = \frac{M[x_2^0,\ x_3^1]}{M[x_2^0]} = \frac{2}{3}$$

ή

$$\hat{\theta}_{x_3^1 \mid x_2^0} = 1 - \hat{\theta}_{x_3^0 \mid x_2^0} = \frac{2}{3}$$

και

$$\hat{\theta}_{x_3^0 \mid x_2^1} = \frac{M[x_2^1,\ x_3^0]}{M[x_2^1]} = \frac{0}{1} = 0 \qquad \hat{\theta}_{x_3^1 \mid x_2^1} = \frac{M[x_2^1,\ x_3^1]}{M[x_2^1]} = \frac{1}{1} = 1$$

ή

$$\hat{\theta}_{x_3^1 \mid x_2^1} = 1 - \hat{\theta}_{x_3^0 \mid x_2^1} = 1$$

με

$$\hat{\theta}_{x_3^0 \mid x_2^0} + \hat{\theta}_{x_3^1 \mid x_2^0} = 1 \qquad \hat{\theta}_{x_3^0 \mid x_2^1} + \hat{\theta}_{x_3^1 \mid x_2^1} = 1$$

Άρα με τη μεγιστοποίηση των απλών πιθανοφανειών και ως αποτέλεσμα των τοπικών πιθανοφανειών καταφέραμε να βρούμε τα $\hat{\theta}$ που μεγιστοποιούν την πιθανοφάνεια $L(\boldsymbol{\theta})$:

$$L(\boldsymbol{\theta}) = L_1\left(\boldsymbol{\theta}_{X_1}\right) L_2\left(\boldsymbol{\theta}_{X_2 \mid X_1}\right) L_3\left(\boldsymbol{\theta}_{X_2 \mid X_1}\right)$$

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{\boldsymbol{u_i} \in Val\left(pa_{X_i}\right)} \left[ \prod_{x_i \in Val(X_i)} \theta_{x_i \mid \boldsymbol{u_i}}{}^{M[\boldsymbol{u_i}, x_i]} \right] =$$

$$= \theta_{x_1^0}{}^1 \, \theta_{x_1^1}{}^3 \left[\theta_{x_2^0 \mid x_1^0}{}^2 \, \theta_{x_2^1 \mid x_1^0}{}^0\right] \left[\theta_{x_2^0 \mid x_1^1}{}^1 \, \theta_{x_2^1 \mid x_1^1}{}^1\right] \left[\theta_{x_3^0 \mid x_2^0}{}^1 \, \theta_{x_3^1 \mid x_2^0}{}^2\right] \left[\theta_{x_3^0 \mid x_2^1}{}^0 \, \theta_{x_3^1 \mid x_2^1}{}^1\right]$$

με:

$$L(\widehat{\boldsymbol{\theta}}) = sup_{\boldsymbol{\theta} \in \Theta} L(\boldsymbol{\theta})$$

όπου $\widehat{\boldsymbol{\theta}}$ είναι τα:

$$\widehat{\boldsymbol{\theta}} = (\hat{\theta}_{x_1^0}, \hat{\theta}_{x_1^1}, \hat{\theta}_{x_2^0 \mid x_1^0}, \hat{\theta}_{x_2^1 \mid x_1^0}, \hat{\theta}_{x_2^0 \mid x_1^1}, \hat{\theta}_{x_2^1 \mid x_1^1}, \hat{\theta}_{x_3^0 \mid x_2^0}, \hat{\theta}_{x_3^1 \mid x_2^0}, \hat{\theta}_{x_3^0 \mid x_2^1}, \hat{\theta}_{x_3^1 \mid x_2^1})$$

$$\widehat{\boldsymbol{\theta}} = (\frac{2}{4}, \frac{2}{4}, 1, 0, \frac{1}{2}, \frac{1}{2}, \frac{1}{3}, \frac{2}{3}, 0, 1)$$

Συνεπώς, τα CPTs του δικτύου μπορούν πλέον και διαμορφώνονται όπως φαίνεται στους Πίνακες 13 – 15.

**Πίνακας 13. Συμπληρωμένος πίνακας CPT για τον κόμβο $X_1$του παραδείγματος της Εικόνας 18.**

| $X_1$ | $P(X_1)$ |
|---|---|
| $x_1^0$ | $\frac{2}{4}$ |
| $x_1^1$ | $\frac{2}{4}$ |

**Πίνακας 14. Συμπληρωμένος πίνακας CPT για τον κόμβο $X_2$του παραδείγματος της Εικόνας 18.**

| $X_1$ | $X_2$ | $P(X_2 \mid X_1)$ |
|---|---|---|
| $x_1^0$ | $x_2^0$ | 1 |
| $x_1^0$ | $x_2^1$ | 0 |
| $x_1^1$ | $x_2^0$ | $\frac{1}{2}$ |
| $x_1^1$ | $x_2^1$ | $\frac{1}{2}$ |

**Πίνακας 15. Συμπληρωμένος πίνακας CPT για τον κόμβο $X_3$του παραδείγματος της Εικόνας 18.**

| $X_2$ | $X_3$ | $P(X_3 \mid X_2)$ |
|---|---|---|
| $x_2^0$ | $x_3^0$ | $\frac{1}{3}$ |
| $x_2^0$ | $x_3^1$ | $\frac{2}{3}$ |
| $x_2^1$ | $x_3^0$ | 0 |
| $x_2^1$ | $x_3^1$ | 1 |

Γνωρίζοντας τα CPTs μπορούμε να βρούμε την πιθανότητα κάθε συνδυασμού τιμών των $X_1, X_2, X_3$ μέσω της από κοινού συνάρτησης πιθανότητας:

$$P(X_1, X_2, X_3) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)$$

και επιπλέον μπορούμε να απαντήσουμε σε ερωτήσεις (queries) για οποιοδήποτε συνδυασμό μεταβλητών του δικτύου μέσω της μπεϋζιανής συμπερασματολογίας.

### *4.3.2 Μπεϋζιανή Εκτίμηση Παραμέτρων*

Στη μπεϋζιανή εκτίμηση παραμέτρων, η παράμετρος θ ή το διάνυσμα παραμέτρων $\boldsymbol{\theta}$ παύει να είναι ένα διάνυσμα παραμέτρων που οι τιμές τους είναι άγνωστες αλλά σταθερές και αποτελεί πλέον μια τυχαία μεταβλητή με δική της κατανομή $P(\boldsymbol{\theta})$.

Ας υποτεθεί πάλι ότι έχουμε ένα τυχαίο δείγμα $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$, ένα άγνωστο διάνυσμα παραμέτρων $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)$ που αποτελεί μια συνεχή m-διάστατη τυχαία μεταβλητή και $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ οι τιμές των $X_1, X_2, \ldots, X_n$, που είναι τα δεδομένα.

Η κατανομή $P(\boldsymbol{\theta})$ ονομάζεται εκ των προτέρων κατανομή (prior distribution) και αυτό συμβαίνει διότι η επιλογή της γίνεται χωρίς να ληφθούν υπόψιν τα δεδομένα. Η επιλογή της βασίζεται στην εμπειρία και στην πεποίθηση μας για τις διάφορες τιμές της $\boldsymbol{\theta}$. Δηλαδή, επιλέγοντας την κατανομή της $\boldsymbol{\theta}$, επιλέγουμε και τις πιθανότητες που μπορεί να έχει, κατά την κρίση μας, για τις διαφορετικές τιμές που μπορεί να πάρει. Με αυτό τον τρόπο, μετουσιώνεται η έννοια της υποκειμενικής πιθανότητας και σε αυτό το σημείο η μπεϋζιανή εκτίμηση παραμέτρων διαφέρει από την μέθοδο εκτίμησης μέγιστης πιθανοφάνειας.

Ο στόχος αρχικά, είναι να βρεθεί η δεσμευμένη κατανομή $P(\boldsymbol{\theta}|\boldsymbol{x})$ που ονομάζεται εκ των υστέρων κατανομή (posterior distribution). Η ονομασία της κατανομής είναι ενδεικτική της λειτουργίας της και αυτό διότι με την εκ των υστέρων κατανομή $P(\boldsymbol{\theta}|\boldsymbol{x})$ λαμβάνονται πλέον υπόψιν οι πληροφορίες που μεταφέρουν τα δεδομένα, επαναπροσδιορίζοντας την πεποίθηση που έχουμε για τις τιμές του διανύσματος των παραμέτρων $\boldsymbol{\theta}$. Δηλαδή, η $P(\boldsymbol{\theta}|\boldsymbol{x})$ εκφράζει την πιθανότητα να πάρει το $\boldsymbol{\theta}$ μια τιμή υπό τη συνθήκη πλέον των τιμών $\boldsymbol{x}$ των τυχαίων μεταβλητών $X_1, X_2, \ldots, X_n$.

Από την εφαρμογή του θεωρήματος Bayes έχουμε:

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{x})} \ , \boldsymbol{\theta} \in \Theta$$

με $P(\boldsymbol{x}|\boldsymbol{\theta}) = L(\boldsymbol{\theta})$ τη συνάρτηση πιθανοφάνειας, όπου ο συμβολισμός | χρησιμοποιείται αντί του **:** για να δηλώσει ότι δεν έχουμε απλά μια εξάρτηση από την παράμετρο, αλλά ότι η $\boldsymbol{\theta}$ είναι τυχαία μεταβλήτη. Επιπλέον, η $P(\boldsymbol{x})$ είναι η περιθώρια συνάρτηση πιθανότητας των $X_1, X_2, \ldots, X_n$ ή αλλιώς περιθώρια πιθανοφάνεια. Άρα:

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{L(\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(\boldsymbol{x})} \ , \boldsymbol{\theta} \in \Theta$$

επειδή όμως, $L(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})$ και $P(\boldsymbol{x}) = k^{-1}$ είναι μια σταθερά που δεν εξαρτάται από το **θ**, έχουμε:

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = kP(\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta}) \ , \boldsymbol{\theta} \in \Theta$$

Η σταθερά $k$ ονομάζεται σταθερά κανονικοποίησης (normalizing constant) και χρειάζεται ώστε να είναι η $P(\boldsymbol{\theta}|\boldsymbol{x})$ μια συνάρτηση πιθανότητας. Ωστόσο, επειδή δεν είναι αναγκαία για την εύρεση της κατανομής της $P(\boldsymbol{\theta}|\boldsymbol{x})$, πολλές φορές παραλείπεται κατά τη διαδικασία εύρεσης της κατανομής για λόγους ευκολίας. Άρα μπορούμε να γράψουμε:

$$P(\boldsymbol{\theta}|\boldsymbol{x}) \propto P(\boldsymbol{\theta})L(\boldsymbol{\theta})$$

δηλαδή, η εκ των υστέρων κατανομή $P(\boldsymbol{\theta}|\boldsymbol{x})$ είναι ανάλογη του γινομένου της εκ των προτέρων κατανομής $P(\boldsymbol{\theta})$ και της συνάρτησης πιθανοφάνειας $L(\boldsymbol{\theta})$.

Αν $P(\boldsymbol{x}, \boldsymbol{\theta})$ η από κοινού συνάρτηση πιθανότητας των $X_1, X_2, \ldots, X_n$ και **θ**, η σταθερά $k$ μπορεί να υπολογιστεί ως εξής:

$$k^{-1} = P(\boldsymbol{x}) = \int_{\Theta} P(\boldsymbol{x}, \boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

$$= \int_{\Theta} P(\boldsymbol{x}|\boldsymbol{\theta})P(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

$$= \int_{\Theta} L(\boldsymbol{\theta})P(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

$$= \int_{\Theta} L(\boldsymbol{\theta})P(\boldsymbol{\theta}) \, d\boldsymbol{\theta}$$

$$= \int_{\Theta} P(\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Ισοδύναμα, μπορεί να βρεθεί από το γεγονός ότι η $P(\boldsymbol{\theta}|\boldsymbol{x})$ θα είναι μια δεσμευμένη συνάρτηση πυκνότητας πιθανότητας (ή μάζας πιθανότητας) και άρα πρέπει:

$$\int_{\Theta} P(\boldsymbol{\theta}|\boldsymbol{x})\, d\boldsymbol{\theta} = 1$$

Στην περίπτωση που η $\boldsymbol{\theta}$ είναι μια διακριτή m-διάστατη τυχαία μεταβλητή με όμοιο τρόπο βρίσκουμε ότι:

$$k^{-1} = P(\boldsymbol{x}) = \sum_{\boldsymbol{\theta}\in\Theta} P(\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})$$

όπου πλέον και η εκ των υστέρων συνάρτηση πιθανότητας $P(\boldsymbol{\theta}|\boldsymbol{x})$ θα είναι μια δεσμεύμενη συνάρτηση μάζας πιθανότητας.

Οι συναρτήσεις $P(\boldsymbol{x})$ και $P(\boldsymbol{\theta})$ είναι διαφορετικές συναρτήσεις πιθανότητας σε αντίθεση με τις $P(x_i|\boldsymbol{\theta})$ με $i = 1, \dots, n$ που είναι ίδιες συναρτήσεις πιθανότητας διότι οι τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$ είναι ισόνομες (Κοκολάκης και Φουσκάκης, 2009). Συνεπώς, η ολοκληρωμένη έκφραση της εκ των υστέρων κατανομής λαμβάνοντας υπόψιν και τη σταθερά $k$ είναι:

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})}{\int_{\Theta} P(\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

Ενώ όταν η $\boldsymbol{\theta}$ είναι μια διακριτή m-διάστατη τυχαία μεταβλητή, η η ολοκληρωμένη έκφραση της εκ των υστέρων κατανομής είναι:

$$P(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{P(\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}\in\Theta} P(\boldsymbol{\theta}) \prod_{i=1}^{n} P(x_i|\boldsymbol{\theta})}$$

#### *4.3.2.1 Μπεϋζιανή Εκτίμηση Παραμέτρων σε Μπεϋζιανό Δίκτυο*

Στη γενική περίπτωση, όπου έχουμε ένα μπεϋζιανό δίκτυο με δομή G, τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ , το σύνολο των δεδομένων D και $\boldsymbol{\theta}$ το σύνολο των παραμέτρων, τότε η εκ των υστέρων κατανομή δίνεται ως εξής:

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)}$$

όπου η $P(\boldsymbol{\theta})$ είναι η εκ των προτέρων κατανομή που περιέχει την πληροφορία κατά την πεποιθησή μας για το σύνολο των παραμέτρων, χωρίς να ληφθεί υπόψιν η πληροφορία από τα δεδομένα, η $P(D|\boldsymbol{\theta})$ η συνάρτηση πιθανοφάνειας και $P(D)$ η περιθώρια συνάρτηση πιθανότητας των $X_1, X_2, \ldots, X_n$ και D.

Στη μέθοδο μέγιστης πιθανοφάνειας, υπό την υπόθεση ότι τα στιγμιότυπα είναι ανεξάρτητα μεταξύ, είδαμε τη ολική αποσύνθεση της πιθανοφάνειας, δηλαδή το γεγονός ότι η συνάρτηση πιθανοφάνειας παραγοντοποιείται σε ένα γινόμενο τοπικών πιθανοφανειών:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right) = \prod_{i=1}^{\boldsymbol{n}} P(D; \boldsymbol{\theta}_{X_i|pa_{X_i}})$$

Στην περίπτωση της μπεϋζιανής εκτίμησης όμως οι παράμετροι είναι τυχαίες μεταβλητές και άρα για να έχουμε την ίδια παραγοντοποίηση στη συνάρτηση της πιθανοφάνειας χρειζόμαστε μια νέα υπόθεση. Υποθέτουμε λοιπόν ότι, τα στιγμιότυπα των δεδομένων $D$ είναι ανεξάρτητα υπό τη συνθήκη των παραμέτρων. Η συγκεκριμένη υπόθεση θα αποδειχθεί στη συνέχεια του κεφαλαίου όπου θα εισαχθούν νέες έννοιες απαραίτητες για την απόδειξή της. Άρα και στην μπεϋζιανή εκτίμηση έχουμε:

$$L(\boldsymbol{\theta}) = P(D|\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i\left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$$

και αν $P(D, \boldsymbol{\theta})$ η από κοινού συνάρτηση πιθανότητας των δεδομένων D και του συνόλου των παραμέτρων $\boldsymbol{\theta}$, τότε η περθώρια συνάρτηση πιθανότητας $P(D)$ είναι:

$$P(D) = \int_{\Theta} P(D, \boldsymbol{\theta})\, d\boldsymbol{\theta}$$

$$P(D) = \int_{\Theta} P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Άρα, η εκ των υστέρων κατανομή μπορεί πλέον να γραφεί ως εξής:

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\int_{\Theta} P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})\, d\boldsymbol{\theta}}$$

Αντίστοιχα στην περίπτωση που η $\boldsymbol{\theta}$ είναι διακριτή τυχαία μεταβλητή:

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{\sum_{\boldsymbol{\theta}\in\Theta} P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}$$

Το επόμενο βήμα είναι να βρεί το $\boldsymbol{\theta}$ που να μεγιστοποιεί την $P(\boldsymbol{\theta}|D)$, μια διαδικασία που καλείται μεγιστοποίηση της εκ των υστέρων κατανομής ή maximum a posteriori probability estimation (MAP). Δηλαδή αν $\widehat{\boldsymbol{\theta}}$ είναι το διάνυσμα παραμέτρων που μεγιστοποιεί την εκ των υστέρων κατανομή τότε:

$$P\left(\widehat{\boldsymbol{\theta}}\middle|D\right) = \max_{\boldsymbol{\theta}\in\Theta} P(\boldsymbol{\theta}|D)$$

ή ισοδύναμα

$$\ln P\left(\widehat{\boldsymbol{\theta}}\middle|D\right) = \max_{\boldsymbol{\theta}\in\Theta} \ln P(\boldsymbol{\theta}|D)$$

Στο σημείο αυτό, αξίζει να δωθεί σημασία στο εξής: η μεγιστοποίηση γίνεται χρησιμοποιώντας τη μεταβλητή $\boldsymbol{\theta}$, η οποία δεν εμφανίζεται πουθενά στον παρονομαστή της posterior. Επομένως, για να μεγιστοποιηθεί η posterior, αρκεί να μεγιστοποιηθεί το γινόμενο του αριθμητή, δηλαδή η πιθανοφάνεια επί την prior κατανομή των παραμέτρων, και αυτό αποτελεί το βασικό πλεονέκτημα της μεθόδου MAP. Χαρακτηριστικά αξίζει να αναφερθεί ότι το ολοκλήρωμα του παρονομαστή είναι πολύ δύσκολο να υπολογιστεί σε προβλήματα υψηλών διαστάσεων, όχι μόνο με κλειστή μορφή, που μπορεί να μην υπάρχει, αλλά ακόμα και με υπολογιστική προσέγγιση. Καταληκτικά, η εύρεση του $\widehat{\boldsymbol{\theta}}$ που μεγιστοποιεί την εκ των υστέρων κατανομή, μπορεί να ερμηνευτεί ως η μεγιστοποίηση της πεποίθησης μας για επιλογή των παραμέτρων $\boldsymbol{\theta}$ υπό τη συνθήκη των δεδομένων $D$.

Στην αρχή, χωρίς να ληφθούν υπόψιν τα δεδομένα, η πεποίθηση μας για τις παραμέτρους και άρα για το μοντέλο των δεδομένων εκφράζεται μέσω της εκ των προτέρων κατανομής $P(\boldsymbol{\theta})$. Στη συνέχεια, αφού λάβουμε υπόψιν τις πληροφορίες που μεταφέρουν τα δεδομένα, καταλήγουμε σε μια δεσμευμένη κατανομή, την εκ των υστέρων κατανομή $P(\boldsymbol{\theta}|D)$. Η $P(\boldsymbol{\theta}|D)$ μας δίνει ένα πλήθος επιλογών, πεπερασμένων ή άπειρων σε αντιστοιχία με το αν το $\boldsymbol{\theta}$ είναι ένα διακριτό ή συνεχές διάνυσμα τυχαίων μεταβλητών, για τις οποίες μας εκφράζει το πόσο πιθανές είναι οι συγκεκριμένες επιλογές υπό τη συνθήκη πλέον της γνώσης των δεδομένων. Το $\widehat{\boldsymbol{\theta}}$ που μεγιστοποιεί την $P(\boldsymbol{\theta}|D)$ μέσω της μεθόδου MAP, εκφράζει την επιλογή των παραμέτρων $\boldsymbol{\theta}$ που μεγιστοποιούν την πεποίθησή μας ή αλλιώς μειώνουν όσο το δυνατόν περισσότερο την αβεβαιότητά μας για το μοντέλο που εκφράζει τα δεδομένα.

Επίσης, μπορεί να παρατηρηθεί μια σημαντική διαφορά σε σχέση με τη μέθοδο μέγιστης πιθανοφάνειας. Συγκεκριμένα, επειδή η $P(\boldsymbol{\theta}|D)$ είναι μια κατανομή, δεν είναι απαραίτητο να περιοριστεί κανείς στη εύρεση ενός συγκεριμένου $\widehat{\boldsymbol{\theta}}$ που τη μεγιστοποιεί αλλά θα μπορούσε να λάβει υπόψιν του τις παραμέτρους της κατανομής, για παράδειγμα είναι δυνατό να βρεθεί η μέση τιμή της κατανομής $E(\boldsymbol{\theta}|D)$ της $P(\boldsymbol{\theta}|D)$. Δηλαδή, η εκτίμηση $\widehat{\boldsymbol{\theta}}$ των παραμέτρων να είναι η μέση τιμή της posterior κατανομής $P(\boldsymbol{\theta}|D)$.

Αν υποθέσουμε ότι έχουμε μια παράμετρο $\theta \in \Theta$ με $\Theta$ τον παραμετρικό χώρο και την posterior $P(\theta|D)$ τότε:

$$\hat{\theta} = E(\theta|D) = \int_{\theta \in \Theta} \theta P(\theta|D)\, d\theta$$

όπου η εκτίμηση της παραμέτρου καταλήγει να είναι η μέση τιμή της posterior $P(\theta|D)$. Αντίστοιχα στη γενική περίπτωση που έχουμε ένα διάνυσμα παραμέτρων $\boldsymbol{\theta} = (\theta_1, \theta_2, \dots, \theta_m)$ και posterior κατανομή $P(\boldsymbol{\theta}|D)$, αν θέλουμε να βρούμε την εκτίμηση της $\hat{\theta}_i$ της παραμέτρου $\theta_i$ με $i = 1,2, \dots, m$, θα έχουμε:

$$P(\theta_i|D) = \int_{\boldsymbol{\theta} \setminus \theta_i} P(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$$

με $\boldsymbol{\theta} \setminus \theta_i$ να σημαίνει όλοτο $\boldsymbol{\theta}$ εκτός της $\theta_i$. Περισσότερο αναλυτικά έχουμε:

$$P(\theta_i|D) = \int_{\theta_m} \dots \int_{\theta_{i+1}} \int_{\theta_{i-1}} \dots \int_{\theta_1} P(\theta_1, \theta_2, \dots, \theta_m)|D)\, d\theta_1 \dots d\theta_{i-1}\, d\theta_{i+1} \dots d\theta_m$$

όπου $P(\theta_i|D)$ είναι η περιθώρια δεσμευμένη συνάρτηση πιθανότητας της από κοινού δεσμευμένης συνάρτησης πιθανότητας $P(\boldsymbol{\theta}|D)$.

Τότε:

$$\hat{\theta}_i = E(\theta_i|D) = \int_{\theta_i} \theta_i\, P(\theta_i|D)\, d\theta_i$$

όπου αντικαθιστώντας την $P(\theta_i|D)$ έχουμε:

$$\hat{\theta}_i = \int_{\theta_i} \theta_i \int_{\theta_m} \dots \int_{\theta_{i+1}} \int_{\theta_{i-1}} \dots \int_{\theta_1} P(\theta_1, \theta_2, \dots, \theta_m)|D)\, d\theta_1 \dots d\theta_{i-1}\, d\theta_{i+1} \dots d\theta_m\, d\theta_i$$

Συνεπώς, ,μπορούμε να βρούμε την εκτίμηση κάθε παραμέτρου μέσω της μέσης τιμής της posterior κατανομής.

Όπως έχει ήδη αναλυθεί, η πιθανοφάνεια στα μπεϋζιανά δίκτυα μπορεί να αποσυντεθεί σε περισσότερους όρους με τη διαδικασία της ολικής αποσύνθεσης και ύστερα με εκείνη της τοπικής αποσύνθεσης. Αυτό αποτελεί πλεονέκτημα στη περίπτωση εφαρμογής της μεθόδου μέγιστης πιθανοφάνειας, διότι μεγιστοποιείται η πιθανοφάνεια μέσω της μεγιστοποίησης αφενός περισσότερων όρων αλλά αφετέρου λιγότερο σύνθετων. Το ίδιο μπορεί να συμβεί και στην περίπτωση της μπεϋζιανής εκτίμησης παραμέτρων που σημαίνει ότι η εκ των υστέρων κατανομή υπόκειται επίσης στη διαδικασία της ολικής αποσύνθεσης και ύστερα σε εκείνη της τοπικής αποσύνθεσης. Για να αναλυθεί όμως αυτό θα πρέπει πρώτα να αναλυθεί η έννοια του μετα-δικτύου (meta-network) και οι υποθέσεις για τις παραμέτρους και τα δεδομένα που βασίζονται σε αυτό (Darwiche, 2009; Koller and Friedman, 2009).

#### *4.3.2.2 Μετα-Δίκτυο και Ολική Παραμετρική Ανεξαρτησία*

Το γράφημα G που είναι ένα DAG, όπως ήδη γνωρίζουμε, δηλώνει τη δομή του μπεϋζιανού δικτύου και τις γραφικές και κατ' επέκταση τις στοχαστικές εξαρτήσεις των μεταβλητών του δικτύου. Στην περίπτωση της μπεϋζιανής εκτίμησης, το σύνολο

των παραμέτρων που πλέον αποτελούν τυχαίες μεταβλητές, αλλά και τα διαφορετικά στιγμιότυπα των δεδομένων μπορούν να αποτελέσουν κόμβους του μπεϋζιανού δικτύου δημιουργώντας έτσι ένα δίκτυο που έχει βασιστεί στο αρχικό.

Ας υποτεθεί ότι έχουμε τη δομή G ενός μπεϋζιανού δικτύου και M το πλήθος των στιγμιοτύπων των δεδομένων. Τότε μπορεί να κατασκευαστεί ένα δίκτυο μεγέθους M για τη δομή G βασισμένο στα M το πλήθος στιγμιότυπα της G, όπου αν $X_i$ είναι μια τυχαία μεταβλητή στη G, τότε με $X_i[m]$ για $m = 1, \dots, M$ θα συμβολίζεται η τιμή της $X_i$ στο κάθε στιγμιότυπο της G. Για κάθε μεταβλητή $X_i$της G, στο δίκτυο θα περιέχεται το σύνολο των παραμέτρων $\boldsymbol{\theta}_{X_i|\boldsymbol{u}}$, όπου $\boldsymbol{u}$ είναι οι τιμές των γονιών **U** της $X_i$. Μάλιστα, για κάθε στιγμιότυπο, το σύνολο αυτό των παραμέτρων $\boldsymbol{\theta}_{X_i|\boldsymbol{u}}$ θα συνδέεται με τα $X_i[m]$ με $\boldsymbol{\theta}_{X_i|\boldsymbol{u}} \rightarrow X_i[1], \dots, X_i[M] \leftarrow \boldsymbol{\theta}_{X_i|\boldsymbol{u}}$. Το δίκτυο αυτό καλείται μετα-δίκτυο **(**Darwiche, 2009).

Οι Koller και Friedman (2009) παρουσιάζουν ένα παράδειγμα με βάση το οποίο μπορεί να γίνει περισσότερο ευδιάκριτη η έννοια του μετα-δικτύου και οι υποθέσεις για τις παραμέτρους που θα οδηγήσουν στην ολική και τοπική αποσύνθεση της εκ των υστέρων κατανομής.

Έστω ότι έχουμε δύο διακριτές τυχαίες μεταβλητές X, Y με τη κάθε μεταβλητή να μπορεί να πάρει μόνο δύο διαφορετικές τιμές και τη X να είναι ο γονέας της Y, δηλαδή $X \rightarrow Y$. Οι $x^0, x^1$ και $y^0, y^1$ είναι οι τιμές που μπορεί να πάρουν οι μεταβλητές X και Y αντίστοιχα. Κατά τη διδικασία παραμετροποίησης του δικτύου συμβολίζουμε με $\theta_{x^0}, \theta_{x^1}$ τις πιθανότητες των τιμών της X, με $\theta_{y^0 \mid x^0}, \theta_{y^1 \mid x^0}$ τις πιθανότητες της Y δεδομένου ότι $X = x^0$ και $\theta_{y^0 \mid x^1}, \theta_{y^1 \mid x^1}$ τις πιθανότητες της Y δεδομένου ότι $X = x^1$. Επιπλέον, για λόγους συντομίας ομαδοποιούμε τις πιθανότητες ως εξής:

$$\boldsymbol{\theta}_X = \{\theta_{x^0}, \theta_{x^1}\}$$

$$\boldsymbol{\theta}_{Y \mid x^0} = \{\theta_{y^0 \mid x^0}, \theta_{y^1 \mid x^0}\}$$

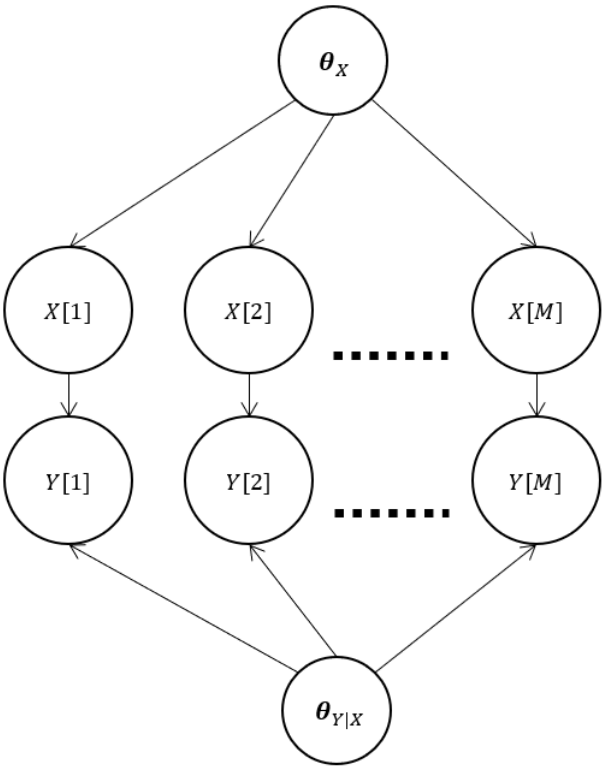
$$\boldsymbol{\theta}_{Y \mid x^1} = \{\theta_{y^0 \mid x^1}, \theta_{y^1 \mid x^1}\}$$

$$\boldsymbol{\theta}_{Y \mid X} = \boldsymbol{\theta}_{Y \mid x^0} \cup \boldsymbol{\theta}_{Y \mid x^1}.$$

όπου $\boldsymbol{\theta}$ το διάνυσμα των παραμέτρων του μπεϋζιανού δικτύου και επιπλεον:

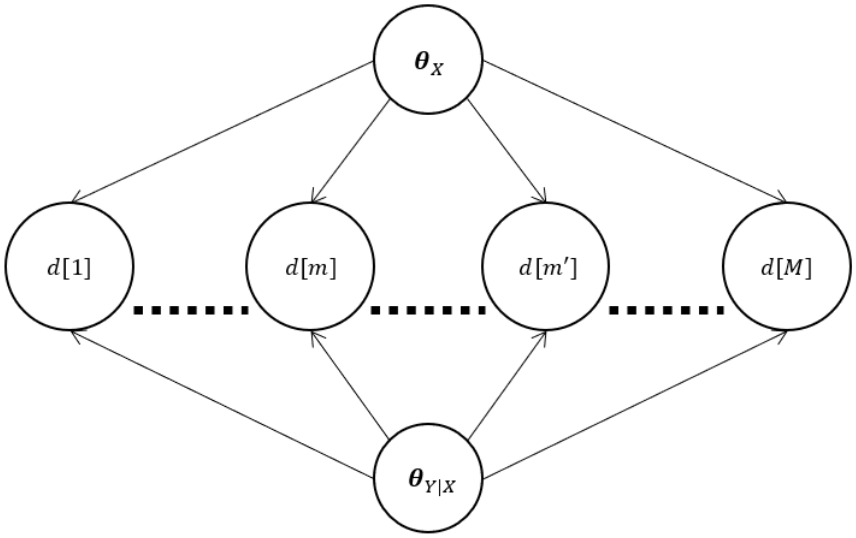
$$\boldsymbol{\theta} = \{\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y \mid X}\} = \{\theta_{x^0}, \theta_{x^1}, \theta_{y^0 \mid x^0}, \theta_{y^0 \mid x^0}, \theta_{y^1 \mid x^0}, \theta_{y^0 \mid x^1}, \theta_{y^1 \mid x^1}\}$$

Τότε το μετα-δίκτυο στην προκειμένη περίπτωση διαμορφώνεται όπως προκύπτει στην Εικόνα19.




**Εικόνα 19 Σχηματική αναπαράσταση μεταδικτύου (Koller and Friedman 2009).**

Το δίκτυο της Εικόνας 19, μπορεί να αναπαρασταθεί με διαφορετικό τρόπο (Εικόνα 20).




**Εικόνα 20. Σχηματική αναπαράσταση μεταδικτύου με χρήση στιγμιοτύπων (Koller and Friedman 2009).**

με $d[m] = \{X[m], Y[m]\}$ να είναι στιγμιότυπο των δεδομένων $\forall\, m = 1, \dots, M$. Αρχικά, παρατηρούμε ότι τα στιγμιότυπα είναι ανεξάρτητα δεδομένου των παραμέτρων και αυτό διότι αν πάρουμε δύο διαφορετικά στιγμιότυπα $d[m], d[m']$, όλες οι διαδρομές που τα ενώνουν είναι μπλοκαρισμένες δεδομένου των παραμέτρων. Συγκεκριμένα, αν πάρουμε την διαδρομή $d[m] \leftarrow \boldsymbol{\theta}_X \rightarrow d[m']$ περιέχει μια αποκλίνουσα σύνδεση και άρα υπό τη συνθήκη της $\boldsymbol{\theta}_X$ τα $d[m], d[m']$ είναι d-separated, δηλαδή η διαδρομή μπλοκάρεται. Το ίδιο ισχύει και για τη διαδρομή $d[m] \leftarrow \boldsymbol{\theta}_{Y \mid X} \rightarrow d[m']$ που επίσης μπλοκάρεται για τον ίδιο λόγο.

Αν τώρα, πάρουμε την διαδρομή $d[m] \leftarrow \boldsymbol{\theta}_X \rightarrow d[1] \leftarrow \boldsymbol{\theta}_{Y \mid X} \rightarrow d[m']$ παρατηρούμε ότι αποτελείται από τρείς τριάδες. Μόλις στην πρώτη από τις τρείς $d[m] \leftarrow \boldsymbol{\theta}_X \rightarrow d[1]$ η πληροφορία μπλοκάρεται δεδομένου της $\boldsymbol{\theta}_X$ και του γεγονότος ότι έχουμε αποκλίνουσα σύνδεση. Άρα μπλοκάρεται όλη η διαδρομή και συνεχίζοντας με την ίδια λογική όλες οι διαδρομές από $d[m]$ σε $d[m']$ είναι μπλοκαρισμένες δεδομένου της $\boldsymbol{\theta}_X$ ή της $\boldsymbol{\theta}_{Y \mid X}$. Συνεπώς, τα στιγμιότυπα είναι d-separated και άρα ανεξαρτητα υπό τη συνθήκη των παραμέτρων.

Στη συνέχεια, η βασική υπόθεση που μπορεί να γίνει είναι ότι οι παράμετροι $\boldsymbol{\theta}_X$ και $\boldsymbol{\theta}_{Y \mid X}$ είναι εκ των προτέρων ανεξάρτητες μεταξύ τους, γεγονός που δεν συμβαίνει πάντα. Δηλαδή, η γνώση της τιμής της μιας παραμέτρου δεν επηρεάζει την πεποίθησή μας για την τιμή της άλλης παραμέτρου. Μπορεί λοιπόν να γραφεί:

$$P(\boldsymbol{\theta}) = P(\boldsymbol{\theta}_X) P(\boldsymbol{\theta}_{Y \mid X})$$

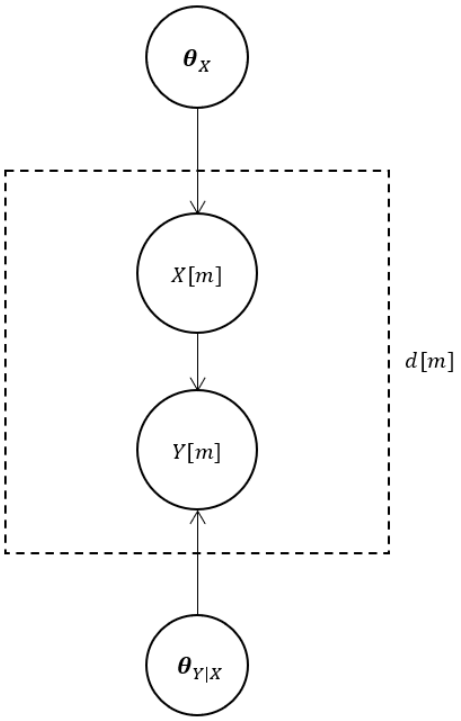
Η συγκεκριμένη υπόθεση ονομάζεται καθολική παραμετρική ανεξαρτησία (global parameter independence) και στη γενική περίπτωση ενός μπεϋζιανού δικτύου γράφεται ως εξής:

$$P(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(\boldsymbol{\theta}_{X_i \mid pa_{X_i}})$$

όπου η prior κατανομή πιθανότητας παραγοντοποιείται στο γινόμενο των prior του κάθε διανύσματος παραμέτρων $\boldsymbol{\theta}_{X_i \mid pa_{X_i}}$ που αντιστοιχεί στο CPT της $X_i \mid pa_{X_i}$ (Koller and Friedman, 2009).

#### *4.3.2.3 Ολική Αποσύνθεση της εκ των Υστέρων Κατανομής*

Μπορεί να παρατηρήσει κανείς, αναλύοντας ένα τυχαίο στιγμιότυπο από το μεταδίκτυο, ότι η διαδρομή από $\boldsymbol{\theta}_X$ σε $\boldsymbol{\theta}_{Y \mid X}$ περιέχει δύο τριάδες κόμβων, όπως φαίνεται στην Εικόνα 21.




**Εικόνα 21. Απομόμωση συγκεκριμένου στιγμιοτύπου από το μεταδίκτυο.**

Η πρώτη τριάδα είναι η $\boldsymbol{\theta}_X \rightarrow X[m] \rightarrow Y[m]$, που είναι μια σύνδεση σε σειρά και η δεύτερη είναι η $X[m] \rightarrow Y[m] \leftarrow \boldsymbol{\theta}_{Y \mid X}$, που είναι μια συγκλίνουσα σύνδεση (v-δομή). Επειδή τα δεδομένα είναι γνωστά, υπό τη συνθήκη του $X[m]$, έχουμε ότι $\boldsymbol{\theta}_X$ και $Y[m]$ είναι d-separated, που σημάινει ότι η μετάδοση της πληροφορίας μπλοκάρεται. Στη δεύτερη τριάδα λόγω της v-δομής και του γεγονότος ότι το $Y[m]$ είναι γνωστό συμπεραίνουμε ότι η τα $X[m]$ και $\boldsymbol{\theta}_{Y \mid X}$ έχουν εξάρτηση ή αλλίως η πληροφορία δεν μπλοκάρεται. Ωστόσο, η πρώτη τριάδα μπλοκάρει την μετάδοση της πληροφορίας και αυτό είναι αρκετό ώστε να καταλήξουμε ότι όλη η διαδρομή είναι μπλοκαρισμένη. Άρα, $\boldsymbol{\theta}_X$ και $\boldsymbol{\theta}_{Y \mid X}$ είναι d-separated δεδομένου του $d[m]$, που σημαίνει ότι οι $\boldsymbol{\theta}_X$ και $\boldsymbol{\theta}_{Y \mid X}$ είναι ανεξάρτητες υπό τη συνθήκη του $d[m]$. Συνεπώς:

$$P\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y \mid X} \mid d[m]\right) = P(\boldsymbol{\theta}_X \mid d[m]) P(\boldsymbol{\theta}_{Y \mid X} \mid d[m])$$

Επειδή το στιγμιότυπο ήταν τυχάιο μπορούμε να γενικεύσουμε το αποτέλεσμα ως εξής:

$$P\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y \mid X} \mid D\right) = P(\boldsymbol{\theta}_X \mid D) P(\boldsymbol{\theta}_{Y \mid X} \mid D)$$

Η παραπάνω σχέση δηλώνει ότι οι παράμετροι είναι ανεξάρτητες μεταξύ τους υπό τη συνθήκη των δεδομένων D. Επειδή όμως οι συναρτήσεις $P\left(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y \mid X} \mid D\right)$, $P(\boldsymbol{\theta}_X \mid D)$ και $P(\boldsymbol{\theta}_{Y \mid X} \mid D)$ είναι εκ των υστέρων συναρτήσεις πιθανότητας, σημαίνει ότι η εκ των υστέρων συνάρτηση πιθανότητας αποσυντίθεται σε μικρότερους όρους όπως ακριβώς είχε γίνει και με την πιθανοφάνεια.

Γενικά, σε ένα μπεϋζιανό δίκτυο με διάνυσμα παραμέτρων $\boldsymbol{\theta}$, τυχαίες μεταβλητές $X_1, X_2, \ldots, X_n$ και σύνολο δεδομένων $D$, η εκ των υστέρων κατανομή πιθανότητας που είναι:

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta}) P(\boldsymbol{\theta})}{P(D)}$$

με την πιθανοφάνεια να είναι $P(D|\boldsymbol{\theta}) = \prod_{i=1}^{n} L_i \left(\boldsymbol{\theta}_{X_i|pa_{X_i}}\right)$. Όμως, υπό την υπόθεση ότι έχουμε καθολική παραμετρική ανεξαρτησία:

$$P(\boldsymbol{\theta}) = \prod_{i=1}^{n} P(\boldsymbol{\theta}_{X_i \mid pa_{X_i}})$$

οπότε η posterior κατανομή γράφεται:

$$P(\boldsymbol{\theta}|D) = \frac{\prod_{i=1}^{n} L_i(\boldsymbol{\theta}_{X_i|pa_{X_i}}) \prod_{i=1}^{n} P(\boldsymbol{\theta}_{X_i \mid pa_{X_i}})}{P(D)}$$

$$= \frac{\prod_{i=1}^{n} L_i(\boldsymbol{\theta}_{X_i|pa_{X_i}}) P(\boldsymbol{\theta}_{X_i \mid pa_{X_i}})}{P(D)}$$

με $P(\boldsymbol{\theta}_{X_i|pa_{X_i}} | D) = \prod_{i=1}^{n} L_i(\boldsymbol{\theta}_{X_i|pa_{X_i}}) P(\boldsymbol{\theta}_{X_i \mid pa_{X_i}})$ να είναι η τοπική posterior κατανομή της $\boldsymbol{\theta}_{X_i|pa_{X_i}}$.

Άρα:

$$P(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} P(\boldsymbol{\theta}_{X_i|pa_{X_i}} | D)$$

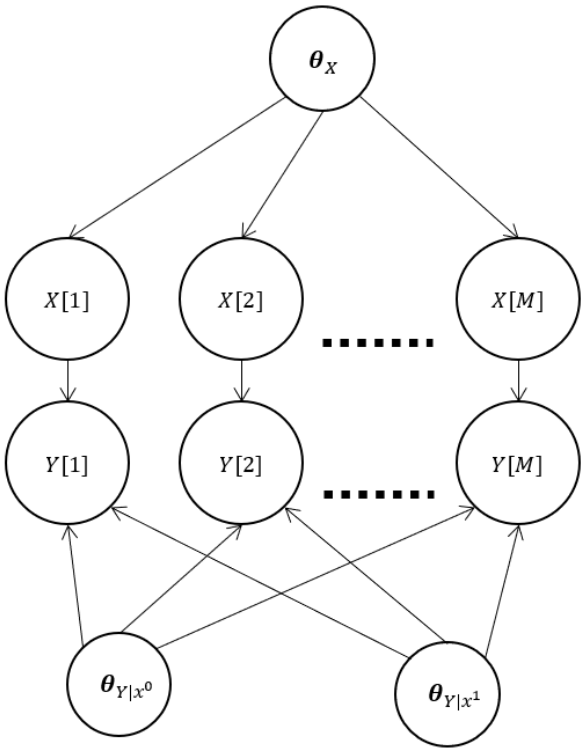
Συνεπώς, η posterior κατανομή πιθανότητας, εαν υποτεθεί ότι έχουμε καθολική παραμετρική ανεξαρτησία, είναι το γινόμενο των τοπικών posterior κατανομών πιθανότητας, που είναι αποτέλεσμα της ίδιας διαδικασίας που εφαρμόστηκε και για

την πιθανοφάνεια δηλαδή, της ολικής αποσύνθεσης της εκ των υστέρων κατανομής (Koller and Friedman, 2009).

#### *4.3.2.4 Τοπική Αποσύνθεση της εκ των Υστέρων Κατανομής*

Η εκ των υστέρων κατανομή μπορεί να αποσυντεθεί σε ακόμα περισσότερους όρους, όπως έγινε και με την πιθανοφάνεια, μια διαδικασία που καλείται τοπική αποσύνθεση της εκ των υστέρων κατανομής.

Η διαδικασία της τοπικής αποσύνθεσης της posterior κατανομής, βασίζεται στην υπόθεση ότι τα σύνολα των παραμέτρων της ίδιας μεταβλητής για διαφορετικές τιμές των γονιών της, είναι ανεξάρτητα μεταξύ τους. Η υπόθεση αυτή ονομάζεται τοπική παραμετρική ανεξαρτησία (local parameter independence). Αναλύοντας περαιτέρω το μετα-δίκτυο του προηγούμενου παραδείγματος θα έχουμε το μετα-δίκτυο της εικόνας 22.




**Εικόνα 22. Περαιτέρω ανάλυση του μεταδικτύου της Εικόνας 19 (Σχηματική αναπαράσταση μεταδικτύου (Koller and Friedman 2009).**

Με την υπόθεση της τοπικής παραμετρικής ανεξαρτησίας, βάση της οποίας για διαφορετικές τιμές των γονιών οι παράμετροι είναι ανεξάρτητες, έχουμε:

$$P(\boldsymbol{\theta}_{Y \mid X}) = P(\boldsymbol{\theta}_{Y \mid x^0}, \boldsymbol{\theta}_{Y \mid x^1}) = P(\boldsymbol{\theta}_{Y \mid x^0}) P(\boldsymbol{\theta}_{Y \mid x^1})$$

Δηλαδή, η εκ των προτέρων συνάρτηση της $P(\boldsymbol{\theta}_{Y \mid X})$ σπάει σε δύο εκ των προτέρων συναρτήσεις, μια για το διάνυσμα παραμέτρων $\boldsymbol{\theta}_{Y \mid x^0}$ και άλλη μια για το $\boldsymbol{\theta}_{Y \mid x^1}$.

Στη συνέχεια, υπό την υπόθεση ότι οι παράμετροι για κάθε διαφορετική τιμή του γονέα της $Y$ είναι ανεξάρτητες υπό τη συνθήκη των δεδομένων, η τοπική posterior κατανομή $P(\boldsymbol{\theta}_{Y \mid X} | D)$ γίνεται:

$$P(\boldsymbol{\theta}_{Y \mid X} | D) = P(\boldsymbol{\theta}_{Y \mid x^0}, \boldsymbol{\theta}_{Y \mid x^1} | D) = P(\boldsymbol{\theta}_{Y \mid x^0} | D) P(\boldsymbol{\theta}_{Y \mid x^1} | D)$$

Άρα η posterior κατανομή θα είναι:

$$P(\boldsymbol{\theta} | D) = P(\boldsymbol{\theta}_X, \boldsymbol{\theta}_{Y \mid X} \mid D) = P(\boldsymbol{\theta}_X \mid D) P(\boldsymbol{\theta}_{Y \mid X} \mid D)$$

όπου

$$P(\boldsymbol{\theta} | D) = P(\boldsymbol{\theta}_X \mid D) P(\boldsymbol{\theta}_{Y \mid x^0} | D) P(\boldsymbol{\theta}_{Y \mid x^1} | D)$$

Άρα, στη περίπτωση ενός μπεϋζιανού δικτύου, κάθε τοπική posterior κατανομή $P(\boldsymbol{\theta}_{X_i | pa_{X_i}} | D)$ μπορεί να αποσυντεθεί ως εξής:

$$P(\boldsymbol{\theta}_{X_i | pa_{X_i}} | D) = \prod_{pa_{X_i} \in Val(pa_{X_i})} P(\boldsymbol{\theta}_{X_i | pa_{X_i}} | D)$$

όπου $Val(pa_{X_i})$ είναι το σύνολο των διαφορετικών συνδυασμών τιμών που μπορούν να πάρουν οι γονείς της $X_i$. Δηλαδή, αποσυντίθεται σε ένα γινόμενο απλών posterior κατανομών που για κάθε διαφορετική τιμή των γονέων της $X_i$ θα αντιστοιχεί μια τέτοια κατανομή.

Γενικά, έστω ένα μπεϋζιανό δίκτυο με δομή $G$, με $X_1, X_2, \ldots, X_n$ τυχαίες μεταβλητές και σύνολο δεδομένων $D$. Αν η εκ των προτέρων κατανομή $P(\boldsymbol{\theta})$ ικανοποιεί τις υποθέσεις της καθολικής και της τοπικής παραμετρικής ανεξαρτησίας, τότε η εκ των υστέρων κατανομή στην μπεϋζιανή εκτίμηση γίνεται:

$$P(\boldsymbol{\theta} | D) = \prod_{i=1}^{n} P(\boldsymbol{\theta}_{X_i | pa_{X_i}} | D) = \prod_{i=1}^{n} \prod_{pa_{X_i} \in Val(pa_{X_i})} P(\boldsymbol{\theta}_{X_i | pa_{X_i}} | D)$$

Αυτό γίνεται διότι, η καθολική παραμετρική ανεξαρτησία και η τοπική παραμετρική ανεξαρτησία της prior κατανομής, οδηγεί αντίστοιχα στην ολική και τοπική αποσύνθεση της posterior κατανομής.

Συνεπώς, η τοπική αποσύνθεση της posterior κατανομής έχει ως σημαντικό αποτέλεσμα το γεγονός ότι αν χρησιμοποιηθεί η μέθοδος MAP για την εύρεση του

διανύσματος $\boldsymbol{\theta}$ τότε η μεγιστοποίηση της posterior ανάγεται στη μεγιστοποίηση των απλών posterior κατανομών $P(\boldsymbol{\theta}_{X_i|pa_{X_i}}|\, D)$ για κάθε διαφορετική τιμή των γονιών $pa_{X_i}$ της $X_i$ (Darwiche, 2009; Koller and Friedman, 2009).

#### *4.3.2.5 Εκ των Προτέρων Κατανομές*

Η επιλογή της prior κατανομής, όπως έχει ήδη αναφερθεί, είναι σημαντική διαδικασία διότι μέσω της κατανομής αυτής μεταφέρεται η πεποίθησή μας και η προηγούμενη γνώση μας για το ποιές θα μπορούσε να είναι οι συγκεκριμένες παράμετροι. Ένα επιπλέον κριτήριο επιλογής είναι η δυσκολία υπολογισμού της περιθώριας συνάρτησης πιθανότητας $P(D)$, που είναι η σταθερά κανονικοποίησης της posterior κατανομής $P(\boldsymbol{\theta}|D)$. Αυτό συμβαίνει διότι για τον υπολογισμό της $P(D)$ χρειάζεται να βρεθεί το ολοκήρωμα πάνω στο γινόμενο της prior κατανομής με την συνάρτηση πιθανοφάνειας:

$$P(D) = \int_{\Theta} P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})\, d\boldsymbol{\theta}$$

Με την posterior να έχει τη μορφή:

$$P(\boldsymbol{\theta}|D) = \frac{P(D|\boldsymbol{\theta})P(\boldsymbol{\theta})}{P(D)}$$

Είναι επίσης σημαντικό, η posterior κατανομή που θα έχουμε ως αποτέλεσμα, να αποτελεί μια εύκολα διαχειρίσιμη και συμπαγή μαθηματική περιγραφή. Γενικά, ανάλογα με τη συνάρτηση πιθανοφάνειας $P(D|\boldsymbol{\theta})$ που έχουμε, μπορούμε να επιλέξουμε μια prior κατανομή η οποία θα μας δίνει posterior κατανομή που θα ανήκει στην ίδια οικογένεια με τη prior.

Μια prior κατανομή $P(\boldsymbol{\theta})$ καλείται συζυγής prior (conjugate prior distribution) της συνάρτησης πιθανοφάνειας $P(D|\boldsymbol{\theta})$, αν η posterior κατανομή $P(\boldsymbol{\theta}|D)$ ανήκει στην ίδια οικογένεια κατανομών που ανήκει και η prior κατανομή. Υπάρχουν έτοιμοι πίνακες που ανάλογα με τη συνάρτηση πιθανοφάνειας αποτυπώνουν το ποιά θα πρέπει να είναι η prior κατανομή και με ποιές παραμέτρους, έτσι ώστε να είναι συζυγής στην συνάρτηση πιθανοφάνειας, δηλαδή η posterior να ανήκει στην ίδια οικογένεια κατανομών με τη prior κατανομή. Οι παράμετροι της prior κατανομής, για να μπορούν να διαφοροποιηθούν από τις παραμέτρους του μοντέλου, καλούνται υπερπαράμετροι (hyperparameters).

Μια από τις συνήθεις περιπτώσεις είναι, για παράδειγμα, μια διακριτή τυχάια μεταβλητή X με δυνατότητα να πάρει μόνο δύο τιμές, δηλαδή μία διωνυμική μεταβλητή, και ένα σύνολο δεδομένων $D = \{X[1], \dots, X[M]\}$. Τότε αν επιλέξουμε ως prior μια κατανομή Βήτα, θα πάρουμε ως posterior επίσης μια κατανομή Βήτα.

Για παράδειγμα, έστω ότι έχουμε ένα πείραμα ριψής μιας πινέζας, με $\Omega = \{\kappa, o\}$, όπου «κ» είναι το ενδεχόμενο να έρθει κεφαλή και «ο» να έρθει η ουρά της πινέζας. Τότε:

$$X(\omega) = \begin{cases} x^1, & \alpha\nu\ \omega = \kappa \\ x^0, & \alpha\nu\ \omega = o \end{cases}$$

και η διακριτή τυχαία μεταβλητή X ακολουθεί κατανομή Bernoulli με συνάρτηση μάζας πιθανότητας:

$$P(X[m]) = \begin{cases} \theta, & \alpha\nu\ X[m] = x^1 \\ 1-\theta, & \alpha\nu\ X[m] = x^0 \end{cases}$$

όπου το $m$, με $1 \le m \le M$, συμβολίζει τον αριθμό της ρίψης ή αλλιώς το m-οστό στιγμιότυπο των δεδομένων $D$ και $0 \le \theta \le 1$. Η συνάρτηση πιθανοφάνειας θα είναι:

$$P(D|\theta) = \prod_{m=1}^{M} P(X[m]|\theta) = P(X[1]|\theta)P(X[2]|\theta)\dots P(X[M]|\theta)$$
$$= \theta^{M[1]}(1-\theta)^{M[0]}$$

όπου $M[1]$ και $M[0]$ μας δίνουν το πλήθος των εμφανίσεων της κεφαλής και της ουράς της πινέζας στο σύνολο των δεδομένων αντίστοιχα. Άρα, αν διαλέξουμε για την prior $P(\theta)$ μια κατανομή Βήτα με παραμέτρους $a_0, a_1$, όπου στην προκειμένη περίπτωση τις λέμε υπερπαραμέτρους, τότε:

$$P(\theta) = \frac{1}{B(a_1, a_0)}\theta^{a_1-1}(1-\theta)^{a_0-1}$$

Συνεπώς, η posterior κατανομή θα είναι:

$$P(\theta|D) \propto P(D|\theta)P(\theta) = \theta^{M[1]}(1-\theta)^{M[0]}\frac{1}{B(a_1, a_0)}\theta^{a_1-1}(1-\theta)^{a_0-1}$$

Δηλαδή:

$$P(\theta|D) \propto \theta^{a_1+M[1]-1}(1-\theta)^{a_0+M[0]-1}$$

Άρα η posterior κατανομή είναι μια κατανομή Βήτα με παραμέτρους τις $a_1 + M[1]$ και $a_0 + M[0]$.

Ένα ακόμα παράδειγμα, αποτελεί η περίπτωση που έχουμε μια τυχαία μεταβλητή διάστασης K που ακολουθεί την πολυωνυμική κατανομή. Συγκεκριμένα, έστω μια

τυχαία μεταβλητή $X$ με $x^1, x^2, \ldots, x^K$ να είναι οι πιθανές διαφορετικές τιμές της. Συμβολίζουμε ως $D = \{d_1, d_2, \ldots, d_M\}$ το -Μ το πλήθος- σύνολο των δεδομένων, με $d_m = X[m]$ $(m = 1, \ldots, M)$ να είναι το m-οστό στιγμιότυπο των δεδομένων. Στην ουσία κάθε στιγμιότυπο m αποτελεί μια ανεξάρτητη δοκιμή Bernoulli διαβαθμισμένης επιτυχίας με πιθανές τιμές τις $x^1, x^2, \ldots, x^K$. Επιπλέον, έστω $\theta_i$ η πιθανότητα να εμφανιστεί η $x^\kappa$ τιμή, δηλαδή:

$$P(X = x^\kappa) = \theta_\kappa$$

Τότε αν οι τυχαίες μεταβλητές $X_1, X_2, \ldots, X_K$ συμβολίζουν το πλήθος των εμφανίσεων των $x^1, x^2, \ldots, x^K$ διαφορετικών τιμών στα Μ το πλήθος στιγμιότυπα, έχουμε ότι η:

$$\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$$

είναι μια Κ-διάστατη τυχαία μεταβλητή που ακολουθεί πολυωνυμική κατανομή, με συνάρτηση μάζας πιθανότητας:

$$f(M[1], M[2], \ldots, M[K]) = P(X_1 = M[1], X_2 = M[2], \ldots, X_K = M[K])$$
$$= \binom{M}{M[1], M[2], \ldots, M[K]} {\theta_1}^{M[1]} {\theta_2}^{M[2]} \ldots {\theta_K}^{M[K]}$$

όπου $M[\kappa]$ είναι το πλήθος εμφανίσεων της $x^\kappa$ τιμής στο σύνολο των δεδομένων $D$.Τα στιγμιότυπα είναι ανεξάρτητα μεταξύ τους, οπότε κάθε στιγμιότυπο μπορεί να θεωρηθεί ότι ακολουθεί διαβασμισμένη Bernoulli, και άρα η συνάρτηση πιθανοφάνειας θα είναι:

$$P(D|\boldsymbol{\theta}) = \prod_{m=1}^{M} P(d_m|\theta) = P(d_1|\boldsymbol{\theta}) P(d_2|\boldsymbol{\theta}) \ldots P(d_M|\boldsymbol{\theta}) = \prod_{\kappa=1}^{K} {\theta_\kappa}^{M[\kappa]}$$

με $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ το διάνυσμα των παραμέτρων. Βασιζόμενοι στην πιθανοφάνεια, αν διαλέξουμε την prior $P(\boldsymbol{\theta})$ να είναι μια κατανομή Dirichlet με υπερπαραμέτρους $\alpha_1, \alpha_2, \ldots, \alpha_K$ και $\boldsymbol{\alpha} = \alpha_1 + \alpha_2 + \cdots + \alpha_K$, τότε:

$$P(\boldsymbol{\theta}) = f(\theta_1, \theta_2, \ldots, \theta_K \,;\, \alpha_1, \alpha_2, \ldots, \alpha_K) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{\kappa=1}^{K} {\theta_\kappa}^{\alpha_\kappa - 1}$$

Άρα αν πάρουμε την posterior κατανομή έχουμε:

$$P(\boldsymbol{\theta}|D) \propto P(D|\boldsymbol{\theta}) P(\boldsymbol{\theta})$$

$$P(\boldsymbol{\theta}|D) \propto \prod_{\kappa=1}^{K} {\theta_\kappa}^{M[\kappa]} \prod_{\kappa=1}^{K} {\theta_\kappa}^{\alpha_\kappa - 1} = \prod_{\kappa=1}^{K} {\theta_\kappa}^{\alpha_\kappa + M[\kappa] - 1}$$

Συνεπώς η posterior κατανομή θα είναι μια κατανομή μια Dirichlet κατανομή με παραμέτρους $\alpha_1 + M[1],\ \alpha_2 + M[2], \dots,\ \alpha_K + M[K]$.

Άρα, παρατηρούμε ότι αν έχουμε μια μεταβλητή που ακολουθεί την κατανομή Bernoulli, η κατανομή Βήτα είναι μια συζυγής prior της συνάρτησης πιθανοφάνειας, ενώ αν έχουμε μεταβλητή που ακολουθεί την πολυωνυμική κατανομή, η κατανομή Dirichlet είναι μια συζυγής prior της συνάρτησης πιθανοφάνειας (Carlin et al. 2013; Koller and Friedman, 2009; Neapolitan, 2003).

Με βάση την ανάλυση που προηγήθηκε, σε ένα μπεϋζιανό δίκτυο, μπορούμε να διαλέξουμε μια prior κατανομή που να είναι συζυγής της συνάρτησης πιθανοφάνειας με αποτέλεσμα να έχουμε μια posterior κατανομη της ίδιας οικογένειας με την prior. Αν λάβουμε υπόψιν την ολική και τοπική αποσύνθεση της posterior κατανομής:

$$P(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} P(\boldsymbol{\theta}_{X_i|pa_{X_i}}|\ D) = \prod_{i=1}^{n} \prod_{pa_{X_i} \in Val(pa_{X_i})} P(\boldsymbol{\theta}_{X_i|pa_{X_i}}|\ D)$$

τότε, χρειάζεται να επιλέξουμε μια prior για το διάνυσμα παραμέτρων $\boldsymbol{\theta}_{X_i|pa_{X_i}}$ για κάθε διαφορετικό συνδυασμό τιμών των γονέων $pa_{X_i}$ της $X_i$. Έστω ότι έχουμε ένα μπεϋζιανό δίκτυο και έναν κόμβο $X_i$, όπου η $X_i|pa_{X_i}$ είναι μια πολυωνυμική κατανομή για κάθε διαφορετικό συνδυασμό των γονιών $pa_{X_i}$ της $X_i$. Τότε σε κάθε μια $X_i|pa_{X_i}$ με συγκεκριμένο συνδυασμό τιμών $pa_{X_i}$θα αντιστοιχεί και μια Dirichlet prior κατανομή με υπερπαραμέτρους:

$$\boldsymbol{a}_{X_i|pa_{X_i}} = (a_{X_i|pa_{X_i}}{}^{1}, a_{X_i|pa_{X_i}}{}^{2}, \dots, a_{X_i|pa_{X_i}}{}^{K_i})$$

όπου $K_i$ είναι το πληθος των τιμών που μπορεί να πάρει η $X_i$. Άρα, η posterior κατανομή θα είναι μια Dirchlet κατανομή με παραμέτρους

$$a_{X_i|pa_{X_i}}{}^{1} + M_i[1],\ a_{X_i|pa_{X_i}}{}^{2} + M_i[2], \dots, a_{X_i|pa_{X_i}}{}^{K_i} + M_i[K_i]$$

όπου $M_i[1], M_i[2], \dots, M_i[K_i]$ είναι το πλήθος εμφάνισης της $1^{\eta\varsigma}$ , $2^{\eta\varsigma}$ και $K_i$ τιμής της $X_i$ αντίστοιχα, στο σύνολο των δεδομένων. Άρα, $\forall\ X_i|pa_{X_i} \in Val(pa_{X_i})$ έχουμε ότι:

$P(\boldsymbol{\theta}_{X_i|pa_{X_i}}|\ D) = f(\theta_{X_i|pa_{X_i}}{}^{1}, \theta_{X_i|pa_{X_i}}{}^{2}, \dots, \theta_{X_i|pa_{X_i}}{}^{K_i};\ a_{X_i|pa_{X_i}}{}^{1} + M_i[1],\ a_{X_i|pa_{X_i}}{}^{2} +$
$+M_i[2], \dots, a_{X_i|pa_{X_i}}{}^{K_i} + M_i[K_i]$

Συνεπώς, για κάθε συνδυασμό των γονιών $pa_{X_i}$ της $X_i$ θα αντιστοιχεί και μια Dirichlet prior κατανομή και άρα στο CPT της $X_i|pa_{X_i}$ θα αντιστοιχούν πολλαπλές Dirichlet prior και άρα πολλαπλές Dirichlet posterior, μια για κάθε συνδυασμό των γονιών $pa_{X_i}$

της $X_i$. Κάτι που με τη σειρά του θα συμβαίνει για κάθε μεταβλητή $X_i$ του μπεϋζιανού δικτύου.

Το αποτέλεσμα αυτό μπορεί να γενιξευτεί και άρα σε κάθε περίπτωση διακριτού μπεϋζιανού δικτύου που έχουμε ολική και τοπική αποσύνθεση της posterior κατανομής μπορούμε να επιλέξουμε συζυγείς prior κατανομές στις απλές συναρτήσεις πιθανοφάνειας ώστε να έχουμε απλές posterior κατανομές που να ανήκουν στην ίδια οικογένεια κατανομών με τις prior κατανομές (Koller and Friedman, 2009).

### *4.3.3 Εύρεση Παραμέτρων Χωρίς Πλήρη Δεδομένα με τον Αλγόριθμο ΕΜ*

Εύλογα δημιουργείται το ερώτημα του πως θα βρεθεί το στατιστικό μοντέλο που ταιριάζει περισσότερο στα δεδομένα όταν δεν έχουμε πλήρη δεδομένα, δηλαδή υπάρχουν στιγμιότυπα των οποίων οι μεταβλητές δεν έχουν τιμή, είτε γιατί δεν υπήρχε παρατήρηση εξ αρχής, γιατί χάθηκε πληροφορία σε κάποια μεταφορά κτλ. Στην περίπτωση αυτή η απάντηση δίνεται από τον αλγόριθμο ΕΜ (Expectation-Maximization algorithm). Στόχος του αλγόριθμου είναι να βρεθεί το σύνολο των παραμέτρων $\widehat{\boldsymbol{\theta}}$ που δίνουν την μεγαλύτερη δυνατή πιθανοφάνεια. Ο αλγόριθμος χωρίζεται σε δύο βασικά βήματα με πρώτο το βήμα της αναμονής (expectation step), στο οποίο χρησιμοποιείται μια εκτίμηση $\widehat{\boldsymbol{\theta}}$ των παραμέτρων ώστε να συμπληρωθούν τα δεδομένα. Το δεύτερο βήμα καλείται «βήμα μεγιστοποίησης» (maximization step), κατά το οποίο χρησιμοποιούνται τα δεδομένα που συμπληρώθηκαν ώστε να βρεθεί η επόμενη εκτίμηση του συνόλου των δεδομένων $\widehat{\boldsymbol{\theta}}'$, βάσει του οποίου θα έχουμε πιθανοφάνεια μεγαλύτερη από τη προηγούμενη. Η διαδικασία αυτή επαναλαμβάνεται μέχρι η απόσταση των πιθανοφανειών να είναι ικανοποιητική ή επαναλαμβάνεται για έναν προκαθορισμένο αριθμό επαναλήψεων. Για να αναλύσουμε καλύτερα τον ΕΜ θα λάβουμε υπόψιν το ίδιο παράδειγμα που χρησιμοποιήθηκε στην μέθοδο μέγιστης πιθανοφάνειας. Έστω ότι έχουμε μπεϋζιανό δίκτυο μεταβλητών $X_1, X_2, X_3$ με δομή $X_1 \rightarrow X_2 \rightarrow X_3$. Η από κοινού συνάρτηση πιθανότητας θα είναι:

$$P(X_1, X_2, X_3) = P(X_1)P(X_2 \mid X_1)P(X_3 \mid X_2)$$

Αυτή τη φορά όμως έστω ότι τα δεδομένα δεν είναι πλήρη (Πίνακας 16).

**Πίνακας 16. Παράδειγμα μη πλήρους πίνακα δεδομένων.**

| | $X_1$ | $X_2$ | $X_3$ |
|---|---|---|---|
| d[1] | ? | $x_2^0$ | $x_3^1$ |
| d[2] | $x_1^1$ | $x_2^0$ | $x_3^1$ |
| d[3] | $x_1^1$ | $x_2^1$ | $x_3^1$ |
| d[4] | $x_1^0$ | $x_2^0$ | ? |

Παρατηρούμε ότι λείπει η τιμή της $X_1$ και της $X_3$ στα στιγμιότυπα d[1] και d[4] αντίστοιχα. Αυτό έχει ως αποτέλεσμα να μην γνωρίζουμε τα CPTs αφού δεν γνωρίζουμε το σύνολο των παραμέτρων. Για να ξεκινήσει η διαδικασία, υποθέτουμε ότι το σύνολο των παραμέτρων θα είναι:

$$\widehat{\boldsymbol{\theta}} = (0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5, 0.5)$$

και για ευκολία θεωρούμε ότι όλες είναι ίδιες και έχουν τιμή ίση με $0.5$, όπως φαίνεται στους Πίνακες 17-19.

**Πίνακας 17. Αρχικοποίηση πίνακα CPT για τον κόμβο $X_1$.**

| $X_1$ | $\boldsymbol{\theta}^0{}_{X_1}$ |
|---|---|
| $x_1^0$ | 0.5 |
| $x_1^1$ | 0.5 |

$+\rightarrow 1$

**Πίνακας 18. Αρχικοποίηση πίνακα CPT για τον κόμβο $X_2$**

| $X_1$ | $X_2$ | $\boldsymbol{\theta}^0{}_{X_2 \mid X_1}$ | |
|---|---|---|---|
| $x_1^0$ | $x_2^0$ | 0.5 | |
| $x_1^0$ | $x_2^1$ | 0.5 | +→1 |
| $x_1^1$ | $x_2^0$ | 0.5 | |
| $x_1^1$ | $x_2^1$ | 0.5 | +→1 |

**Πίνακας 19. Αρχικοποίηση πίνακα CPT για τον κόμβο $X_3$**

| $X_2$ | $X_3$ | $\boldsymbol{\theta}^0{}_{X_3 \mid X_2}$ | |
|---|---|---|---|
| $x_2^0$ | $x_3^0$ | 0.5 | |
| $x_2^0$ | $x_3^1$ | 0.5 | +→1 |
| $x_2^1$ | $x_3^0$ | 0.5 | |
| $x_2^1$ | $x_3^1$ | 0.5 | +→1 |

Η πιθανοφάνεια θα είναι:

$$L(\boldsymbol{\theta}^0) = P(D \mid \boldsymbol{\theta}^0) = \prod_{m=1}^{4} P_{\boldsymbol{\theta}^0}(\, d[m])$$

όπου για χάριν ευκολίας αντί της $P(d[m] \mid \boldsymbol{\theta}^0)$ γράφουμε $P_{\boldsymbol{\theta}^0}(d[m])$. Επειδή δεν έχουμε πλήρη όλα τα στιγμιότυπα η πιθανοφάνεια γράφεται:

$L(\boldsymbol{\theta}^0) = P_{\boldsymbol{\theta}^0}(d[1])P_{\boldsymbol{\theta}^0}(d[2])P_{\boldsymbol{\theta}^0}(d[3])P_{\boldsymbol{\theta}^0}(d[4])$

$L(\boldsymbol{\theta}^0) = P_{\boldsymbol{\theta}^0}(x_2^0, x_3^1)P_{\boldsymbol{\theta}^0}(x_1^1, x_2^0, x_3^1)P_{\boldsymbol{\theta}^0}(x_1^1, x_2^1, x_3^1)P_{\boldsymbol{\theta}^0}(x_1^0, x_2^0,)$

Για να βρεθεί η πιθανοφάνεια χρειάζεται να εφαρμόσουμε μπεϋζιανή συμπερασματολογία με βάση τα CPTs του συνόλου των παραμέτρων $\boldsymbol{\theta}^0$:

$$P_{\boldsymbol{\theta}^0}(x_2^0, x_3^1) = \sum_{x_1} P_{\boldsymbol{\theta}^0}(x_1, x_2^0, x_3^1) = P_{\boldsymbol{\theta}^0}(x_1^0, x_2^0, x_3^1) + P_{\boldsymbol{\theta}^0}(x_1^1, x_2^0, x_3^1)$$

$$= P_{\boldsymbol{\theta}^0}(x_1^0) P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^0) P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0) + P_{\boldsymbol{\theta}^0}(x_1^1) P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^1) P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0)$$

$$= 0.5 * 0.5 * 0.5 + 0.5 * 0.5 * 0.5 = \frac{1}{4}$$

ομοίως

$$P_{\boldsymbol{\theta}^0}(x_1^1, x_2^0, x_3^1) = P_{\boldsymbol{\theta}^0}(x_1^1) P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^1) P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0) = \frac{1}{8}$$

$$P_{\boldsymbol{\theta}^0}(x_1^1, x_2^1, x_3^1) = P_{\boldsymbol{\theta}^0}(x_1^1) P_{\boldsymbol{\theta}^0}(x_2^1 \mid x_1^1) P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^1) = \frac{1}{8}$$

$$P_{\boldsymbol{\theta}^0}(x_1^0, x_2^0, ) = \sum_{x_3} P_{\boldsymbol{\theta}^0}\,(x_1^0, x_2^0, x_3)$$

$$= P_{\boldsymbol{\theta}^0}(x_1^0) P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^0) P_{\boldsymbol{\theta}^0}(x_3^0 \mid x_2^0) + P_{\boldsymbol{\theta}^0}(x_1^0) P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^0) P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0) = \frac{1}{4}$$

Άρα:

$$L(\boldsymbol{\theta}^0) = \frac{1}{4} * \frac{1}{8} * \frac{1}{8} * \frac{1}{4} = 9.77 * 10^{-4}$$

Ως στόχο έχουμε να βρούμε το $\boldsymbol{\theta}^1$ που δίνει μεγαλύτερη πιθανοφάνεια από τη $L(\boldsymbol{\theta}^0)$. Κάτι που δεν μπορεί να γίνει άμεσα όπως θα δούμε παρακάτω. Αν τα δεδομένα ήταν πλήρη, για να βρούμε για παράδειγμα τις εκτιμήσεις των παραμέτρων ${\theta^1}_{x_1^0}$ και ${\theta^1}_{x_3^0 \mid x_2^0}$ που θα ανήκαν στο $\boldsymbol{\theta}^1$ θα είχαμε:

$${\theta^1}_{x_1^0} = P_{\boldsymbol{\theta}^0}(X_1 = x_1^0) = \frac{N(X_1 = x_1^0)}{N}$$

$${\theta^1}_{x_3^0 \mid x_2^0} = P_{\boldsymbol{\theta}^0}(X_3 = x_3^0 \mid X_2 = x_2^0) = \frac{P_{\boldsymbol{\theta}^0}(X_3 = x_3^0,\ X_2 = x_2^0)}{P_{\boldsymbol{\theta}^0}(X_2 = x_2^0)}$$

$$= \frac{\frac{N(X_3 = x_3^0,\ X_2 = x_2^0)}{N}}{\frac{N(X_2 = x_2^0)}{N}} = \frac{N(X_3 = x_3^0,\ X_2 = x_2^0)}{N(X_2 = x_2^0)}$$

όπου $N$ το πλήθος των στιγμιοτύπων των δεδομένων, $N(X_1 = x_1^0)$ ο αριθμός εμφάνισης της τιμής $x_1^0$ στα δεδομένα, $N(X_2 = x_2^0)$ ο αριθμός εμφάνισης της τιμής $x_2^0$ στα δεδομένα και $N(X_3 = x_3^0,\ X_2 = x_2^0)$ ο αριθμός εμφάνισης των τιμών $x_3^0, x_2^0$ στα δεδομένα. Όμως δεν έχουμε πλήρη δεδομένα και συγκεκριμένα δεν ξέρουμε στο στιγμιότυπο d[1] τι τιμή παίρνει η $X_1$. Με αποτέλεσμα να μην ξέρουμε τον αριθμό

$N(X_1 = x_1^0)$. Τότε παίρνουμε την αναμενόμενη τιμή του αριθμού εμφάνισης της $x_1^0$ στα δεδομένα:

$$E[N(X_1 = x_1^0)] = P_{\boldsymbol{\theta}^0}(x_1^0 \mid x_2^0, x_3^1) + 1 + 0 + 0$$

δηλαδή αθροίζουμε όσες φορές ξέρουμε ότι εμφανίστηκε στο κάθε στιγμιότυπο και εκεί που δεν ξέρουμε αθροίζουμε την πιθανότητα του να παρθεί η τιμή $x_1^0$ δεδομένου του υπόλοιπου στιγμιοτύπου. Για να βρεθεί η $P_{D,\boldsymbol{\theta}^0}(X_1 = x_1^0 \mid x_2^0, x_3^1)$ χρειάζεται μπεϋζιανή συμπερασματολογία με βάση τα CPTs από το $\boldsymbol{\theta}^0$:

$$P_{\boldsymbol{\theta}^0}(X_1 = x_1^0 \mid x_2^0, x_3^1) = \frac{P_{\boldsymbol{\theta}^0}(x_1^0, x_2^0, x_3^1)}{P_{\boldsymbol{\theta}^0}(x_2^0, x_3^1)} = \frac{P_{\boldsymbol{\theta}^0}(x_1^0)P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^0)P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0)}{\sum_{x_1} P_{\boldsymbol{\theta}^0}(x_1)P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1)P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0)}$$

$$= \frac{P_{\boldsymbol{\theta}^0}(x_1^0)P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^0)P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0)}{P_{\boldsymbol{\theta}^0}(x_1^0)P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^0)P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0) + P_{\boldsymbol{\theta}^0}(x_1^1)P_{\boldsymbol{\theta}^0}(x_2^0 \mid x_1^1)P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_2^0)} = \frac{1}{2}$$

Ομοίως η αναμενόμενη τιμή του αριθμού εμφάνισης των τιμών $x_2^0, x_3^0$ θα είναι:

$$E[N(X_3 = x_3^0, X_2 = x_2^0)] = 0 + 0 + 0 + P_{\boldsymbol{\theta}^0}(x_3^0 \mid x_1^0, x_2^0)$$

όπου βρίσκουμε την $P_{\boldsymbol{\theta}^0}(x_3^0 \mid x_1^0, x_2^0)$ με βάση τα CPTs από το $\boldsymbol{\theta}^0$ όπως και προηγουμένως, ενώ η αναμενόμενη τιμή του αριθμού εμφάνισης της τιμής $x_2^0$ θα είναι:

$$E[N(X_2 = x_2^0)] = 1 + 1 + 0 + 1$$

όπου δεν χρειάζεται να βρεθεί κάποια πιθανότητα αφού σε όλα τα στιγμιότυπα παρατηρείται τιμή της $X_2$. Στο σημείο αυτό συμπληρώνουμε στιγμιότυπα στα δεδομένα λαμβάνοντας υπόψιν κάθε πιθανή εκδοχή των τιμών της $X_1$. Το ίδιο κάνουμε και για κάθε άλλη μεταβλητή που τις λείπουν τιμές στα δεδομένα, όπως την $X_3$. Τελικά, για κάθε στιγμιότυπο με μη πλήρη δεδομένα, προκύπτουν πιθανοτικώς σταθμισμένα ψευδοστιγμιότυπα όπως φαίνεται στον Πίνακα 20.

**Πίνακας 20. Πίνακας ψευδοστιγμιοτύπων πρώτης επανάληψης.**

| | $X_1$ | $X_2$ | $X_3$ | Πιθανότητα κάθε στιγμιοτύπου βάση το $\boldsymbol{\theta}^0$ |
|---|---|---|---|---|
| d[1] | ? | $x_2^0$ | $x_3^1$ | |
| | $x_1^0$<br>$x_1^1$ | $x_2^0$<br>$x_2^0$ | $x_3^1$<br>$x_3^1$ | $P_{\boldsymbol{\theta}^0}(x_1^0 \mid x_2^0, x_3^1)$<br>$P_{\boldsymbol{\theta}^0}(x_1^1 \mid x_2^0, x_3^1)$ |
| d[2] | $x_1^1$ | $x_2^0$ | $x_3^1$ | |
| d[3] | $x_1^1$ | $x_2^1$ | $x_3^1$ | |
| d[4] | $x_1^0$ | $x_2^0$ | ? | |
| | $x_1^0$<br>$x_1^0$ | $x_2^0$<br>$x_2^0$ | $x_3^0$<br>$x_3^1$ | $P_{\boldsymbol{\theta}^0}(x_3^0 \mid x_1^0, x_2^0)$<br>$P_{\boldsymbol{\theta}^0}(x_3^1 \mid x_1^0, x_2^0)$ |

Ύστερα, βρίσκουμε και τις υπόλοιπες αναμενόμενες τιμές του αριθμού εμφάνισης των τιμών με τον εξής τύπο που διαφαίνεται και από το παράδειγμα:

$$E_{\boldsymbol{\theta}^0}[N(X_i, pa_{X_i} \mid D)] = \sum_{m=1}^{4} P_{\boldsymbol{\theta}^0}(X_i, pa_{X_i} \mid d[m]) = \sum_{m=1}^{4} P(X_i, pa_{X_i} \mid d[m],\ \boldsymbol{\theta}^0)$$

Δηλαδή, για κάθε μεταβλητή μπορεί να δημιουργηθεί ένας πίνακας αναμενόμενων τιμών για κάθε συνδυασμό τιμών της μεταβλητής και των γονιών της. Η συμπλήρωση των δεδομένωνμε τα στιγμιότυπα, που γίνεται με μπεϋζιανή συμπερασματολογία και η εύρεση των αναμενόμενων τιμών εμφάνισης των διάφορων τιμών για την κάθε μεταβλητή είναι το βήμα του expectation.

Στο βήμα του maximization, που μπορεί πλέον να πραγματοποιηθεί, βρίσκουμε το $\boldsymbol{\theta}^1$ που δίνει πιθανοφάνεια μεγαλύτερη της $L(\boldsymbol{\theta}^0)$. Στο συγκεκριμένο παράδειγμα, το ${\theta^1}_{x_3^0 \mid x_2^0}$ πλέον γράφεται ως εξής:

$${\theta^1}_{x_3^0 \mid x_2^0} = \frac{E_{\boldsymbol{\theta}^0}[N(X_3 = x_3^0, pa_{X_3} = x_2^0 \mid D)]}{E[N(X_2 = x_2^0)]}$$

$$\theta^1{}_{x_3^0 \mid x_2^0} = \frac{E_{\boldsymbol{\theta}^0}[N(X_3 = x_3^0, pa_{X_3} = x_2^0 \mid D)]}{\sum_{x_3} E_{\boldsymbol{\theta}^0}[N(x_3, X_2 = x_2^0) \mid D]}$$

όπου $\sum_{x_3} E_{\boldsymbol{\theta}^0}[N(x_3, X_2 = x_2^0) \mid D] = E[N(X_2 = x_2^0)]$ και με τον ίδιο τρόπο βρίσκουμε τις υπόλοιπες παραμέτρους. Αφού βρεθεί το $\boldsymbol{\theta}^1$ βρίσκουμε την $L(\boldsymbol{\theta}^1)$ με μπεϋζιανή συμπερασματολογία χρησιμοποιώντας πλέον τα νέα CPTs με γνωστές τις τιμές των παραμέτρων $\theta^1{}_{X_i \mid pa_{X_i}}$. Με τη διαδικασία αυτή, βάση θεωρητικώς αποδεδειγμένων αποτελεσμάτων, ο αλγόριθμος έχει εξασφαλίσει ότι:

$$L(\boldsymbol{\theta}^1) \geq L(\boldsymbol{\theta}^0)$$

και βέβαια:

$$\log L(\boldsymbol{\theta}^1) \geq \log L(\boldsymbol{\theta}^0)$$

Η ίδια διαδικασία επαναμβάνεται και σε κάθε επανάληψη $t$ θα έχουμε:

$$\log L(\boldsymbol{\theta}^{t+1}) \geq \log L(\boldsymbol{\theta}^t)$$

Γενικά, σύμφωνα με τους Jensen και Nielsen (2007), ο αλγόριθμος EM για μπεϋζιανά δίκτυα είναι ο εξής:

1. Επιλέγουμε ένα $\varepsilon > 0$ που θα είναι η απόσταση κάτω από την οποία θα πρέπει να σταματήσουν οι επαναλήψεις
2. $\boldsymbol{\theta}^0 = \{\theta_{ijk}\}$ είναι οι αρχική εκτίμηση των παραμέτρων που επιλέγεται αυθαίρετα με $i$ να είναι ο δείκτης για τη μεταβλητή $X_i$, $j$ είναι για τον συνδυσμό των γονιών της $X_i$ και $k$ είναι για τις τιμές της $X_i$.
3. Θέτουμε $t = 0$
4. Επαναλαμβάνουμε:

   Expectation βήμα: Για κάθε $i$ υπολογίζουμε τον πίνακα των αναμενόμενων τιμών:

   $$E_{\boldsymbol{\theta}^t}[N(X_i,\ pa_{X_i} \mid D)] = \sum_{d \in D} P(X_i, pa_{X_i} \mid d[m],\ \boldsymbol{\theta}^t)$$

   Maximization βήμα: Χρησιμοποιούμε τις αναμενόμενες τιμές σαν ήταν πραγματικές ώστε να βρούμε την νέα εκτίμηση $\widehat{\boldsymbol{\theta}}$ των παραμέτρων $\hat{\theta}_{ijk}$:

   $$\hat{\theta}_{ijk} = \frac{E_{\boldsymbol{\theta}^t}[N(X_i = k,\ pa_{X_i} = j \mid D)]}{\sum_h E_{\boldsymbol{\theta}^t}[N(X_i = h,\ pa_{X_i} = j \mid D)]}$$

   Θέτουμε $\boldsymbol{\theta}^{t+1} = \widehat{\boldsymbol{\theta}}$ και $t = t + 1$

   (Επαναλαμβάνουμε) Μέχρι $|\log(\boldsymbol{\theta}^{t+1}) - \log(\boldsymbol{\theta}^t)| \leq \varepsilon$

Όπως έχει αναφερθεί και προηγουμένως, αντί μιας δεδομένης απόστασης που θα πρέπει να φτάσουν οι πιθανοφάνειες θα μπορούσε να χρησιμοποιηθεί και ένα προκαθορισμένος αριθμός επαναλήψεων. Θα πρέπει να σημειωθεί ότι ο αλγόριθμος ΕΜ δεν εξασφαλίζει την εύρεση ολικού μεγίστου για την πιθανοφάνεια. Εφόσον κάθε νέα εκτίμηση παραμέτρων θα δίνει μεγαλύτερη πιθανοφάνεια, αύτο που εξασφαλίζεται είναι ότι στο τέλος των επαναλήψεων ο αλγόριθμος θα έχει βρεί ένα τοπικό μέγιστο για την πιθανοφάνεια. Τέλος, ο ΕΜ αλγόριθμος μπορεί επιπλέον να χρησιμοποιηθεί και για την εύρεση των παραμέτρων που μεγιστοποιούν την posterior συνάρτηση πιθανότητας όταν δεν έχουμε πλήρη δεδομένα. Δηλαδή εκεί που είχαμε την λογαριθμική πιθανοφάνεια πλέον έχουμε να προστίθεται και ο λογάριθμος της prior κατανομής:

$$\log L(\boldsymbol{\theta})\, P(\boldsymbol{\theta}) = \log L(\boldsymbol{\theta}) + \log P(\boldsymbol{\theta})$$

Με αυτό το τρόπο ο ΕΜ λείτουργεί έτσι ώστε σε κάθε επανάληψη να βρίσκονται οι παράμετροι $\boldsymbol{\theta}$ που δίνουν μεγαλύτερο άθροισμα $\log L(\boldsymbol{\theta}) + \log P(\boldsymbol{\theta})$ σε σχέση με προηγουμένως. Η περιθώρια πιθανότητα $P(D)$ δεν επηρεάζει τη διαδικασία γιατί δεν μεταβάλλεται σε κάθε επανάληψη (Darwiche, 2009; Jensen and Nielsen, 2007; Koller and Friedman, 2009).

## 4.4 Εύρεση Δομής Μπεϋζιανού Δικτύου

Η δομή του δικτύου, όπως εξετάστηκε μέχρι στιγμής, είναι έτοιμη ως το αποτέλεσμα της άποψης ειδικών για το πως θα έπρεπε να διαμορφωθεί το δίκτυο. Η χρήση των δεδομένων αφορούσε την εύρεση των παραμέτρων που δίνουν το μοντέλο που ακολουθούν τα δεδομένα, μια διαδικασία που, όπως έχει ήδη αναφερθεί, καλείται εκπαίδευση του αλγόριθμου στα δεδομένα ή αλλιώς fitting. Υπάρχει ωστόσο και η δυνατότητα εύρεσης της δομής ενός μπεϋζιανού δικτύου από τα ίδια τα δεδομένα (structure learning). Στο πλαίσιο αυτό υπάρχουν δύο βασικές κατηγορίες μεθόδων με πρώτη την εύρεση της δομής βασισμένη σε περιορισμούς (constrained based) και δεύτερη την εύρεση της δομής βασισμένη σε κάποια βαθμολογία (score based).

Η πρώτη κατηγορία (constrained based μεθόδοι) βασίζεται στο να βρίσκει τις υπό συνθήκη ανεξαρτησίες των μεταβλητών που μπορούν να παραχθούν από τα δεδομένα, μέσα από τεστ που εξετάζουν την ύπαρξη υπό συνθήκη ανεξαρτησίας (conditional independence tests). Οι ανεξαρτησίες αντανακλούν την γραφική ιδιότητα του d-separation που είναι αναγκαία για την εύρεση της δομής του μπεϋζιανού δικτύου. Η δεύτερη κατηγορία (score based μεθόδοι) βασίζεται στην παραγωγή υποψήφιων μπεϋζιανών δικτύων με χρήση κατάλληλων συναρτήσεων βαθμολόγησης (score functions), βάση των οποίων γίνεται η επιλογή του δικτύου με το μεγαλύτερο βαθμό. Επομένως, επιλέγεται η δομή του δικτύου που είναι πιθανότερο να έχει παραξεί τα δεδομένα ή αλλιώς επιλέγεται η δομή που ταιριάζει περισσότερο στα δεδομένα.

Κατά τη διαδικασία εύρεσης της δομής εμφανίζονται δύο βασικά προβλήματα τα οποία καλούνται οι αλγόριθμοι να αποφύγουν. Το πρώτο πρόβλημα είναι ότι ο αριθμός των υποψήφιων μπεϋζιανών δικτύων είναι πολύ μεγάλος και μεγαλώνει όσο περισσότερο αυξάνεται το πλήθος των κόμβων. Συγκεκριμένα, η συνάρτηση $f(n)$ δίνει το πλήθος των διαφορετικών δομών σε σχέση με τους κόμβους $n$:

$$f(n) = \sum_{i=1}^{n} (-1)^{i+1} \frac{n!}{(n-i)!\, n!} 2^{i(n-i)} f(n-1)$$

Το δεύτερο πρόβλημα αφορά το γεγονός ότι μπορεί να υπάρχει παραπάνω από μια υποψήφια δομή δικτύου που να ταιριάζει στα δεδομένα.

### *4.4.1 Constraint based μέθοδοι*

Στην περίπτωση των constrained based μεθόδων, οι δύο δημοφιλέστεροι αλγόριθμοι, που έχουν χρησιμοποιηθεί και στη μελέτη περίπτωσης, είναι ο GS (Grow-Shrink) (Margaritis, 2003) και ο IAMB (Incremental Association Markov Blanket) (Tsamardinos *et al.*, 2003). Στόχος και των δύο αλγόριθμων είναι η εύρεση του Markov blanket (MB) κάθε κόμβου.

Ο αλγόριθμος GS χωρίζεται σε δύο φάσεις, με πρώτη τη φάση της ανάπτυξης (growing phase) και δεύτερη τη φάση της συρρίκνωσης (shrinking phase). Οι φάσεις αυτές αφορούν την ανάπτυξη και ύστερα συρρίκνωση ενός συνόλου S, που στο τέλος θα καταλήξει να είναι το Markov blanket ενός κόμβου. Συγκεκριμένα, έστω ότι έχουμε έναν κόμβο $X \in U$, με $U$ να είναι το σύνολο των κόμβων του δικτύου. Τότε ο αλγόριθμος GS είναι ο εξής:

1. $S \leftarrow \emptyset$.
2. Ενώ $\exists\, Y \in U - \{X\}$ τέτοιο ώστε $!(Y \perp\!\!\!\perp X \mid S)$, ανάθεσε $S \leftarrow S \cup \{Y\}$. (growing phase)
3. Ενώ $\exists\, Y \in S$, τέτοιο ώστε $Y \perp\!\!\!\perp X \mid S - \{Y\}$, ανάθεσε $S \leftarrow S - \{Y\}$. (shrinking phase)
4. $B(X) \leftarrow S$.

όπου $B(X)$ συμβολίζει το Markov blanket του κόμβου $X$.

Όμοια με τον GS, ο αλγόριθμος IAMB αποτελείται από δύο φάσεις και έχει ως στόχο την εύρεση του Markov blanket($K$), δηλαδή του Markov blanket ($MB$) κάθε κόμβου $K$. Στη πρώτη φάση δημιουργείται ένα σύνολο $CMB(K)$ που αντιστοιχεί στο υποψήφιο Markov blanket (candidate Markov blanket) για τον υπό εξέταση κόμβο $T$. Στη πρώτη φάση το $CMB(T)$ μπορεί να περιέχει και κόμβους που δεν ανήκουν στο $MB(T)$. Στη δεύτερη φάση αφαιρούνται από το σύνολο $CMB(T)$ οι κόμβοι που δεν ανήκουν στο $MB(T)$, με αποτέλεσμα στο τέλος να ταυτίζονται τα σύνολα $CMB(T) = MB(T)$ και να έχει βρεθεί το Markov blanket του κόμβου $T$ (Margaritis, 2003).

### *4.4.2 Score based μέθοδοι*

Στη συνέχεια, στις score based μεθόδους, οι συναρτήσεις βαθμολόγησης έχουν το βασικό ρόλο να αποδίδουν μια βαθμολογία σε κάθε υποψήφια δομή ενός μπεϋζιανού δικτύου, με στόχο να βρεθεί εκείνο που είναι το πιθανότερο να έχει παραχθεί από τα

δεδομένα. Μια βασική ιδιότητα που θα πρέπει να ικανοποιεί μια τέτοια συνάρτηση είναι να ισορροπεί την ακρίβεια της υποψήφιας δομής με την πολυπλοκότητα της δομής. Δηλαδή η αύξηση του μεγέθους της δομής θα επιφέρει μείωση της πιθανότητας η δομή που έχει βρεθεί να έχει παράξει τα δεδομένα. Επιπλέον, μια συνάρτηση βαθμολόγησης θα πρέπει να προσφέρει τη δυνατότητα διαχειρίσιμων υπολογισμών, ώστε να καταλήξει σε εύλογο διάστημα σε μια βαθμολόγηση. Μια συνάρτηση βαθμολόγησης, που θα χρησιμοποιηθεί στη μελέτη περίπτωσης, και η οποία ικανοποιεί τις παραπάνω ιδιότητες είναι η BIC (Bayesian Information Criterion) (Schwarz, 1978). Αν υποτεθεί ότι $S$ είναι η υποψήφια δομή και $D$ τα δεδομένα, τότε η βαθμολογία της BIC δίνεται ως εξής:

$$BIC(S \mid D) = \ln P(D \mid \widehat{\boldsymbol{\theta}}_S) - \frac{size(S)}{2} \ln N$$

όπου $\widehat{\boldsymbol{\theta}}_S$ είναι η εκτίμηση των παραμέτρων που μεγιστοποιούν την πιθανοφάνεια της δομής $S$ και $P(D \mid \widehat{\boldsymbol{\theta}}_S)$ είναι η μέγιστη τιμή της συνάρτησης πιθανοφάνειας. Επιπλέον, το $size(S)$ συμβολίζει το πλήθος των ανεξάρτητων παραμέτρων στη δομή $S$ και το $N$ είναι το πλήθος των στιγμιοτύπων των δεδομένων. Η πολυπλοκότητα της δομής $S$ αναπαρίσταται από τον όρο:

$$\frac{size(S)}{2} \ln N$$

και αυξάνει καθώς αυξάνονται το πλήθος των ανεξάρτητων παραμέτρων και το πλήθος των στιγμιοτύπων των δεδομένων. Η ακρίβεια της δομής αναπαρίσταται από τον όρο:

$$\ln P(D \mid \widehat{\boldsymbol{\theta}}_S)$$

και μεγιστοποιείται όταν μεγιστοποιείται η πιθανοφάνεια.Συνεπώς, μπορεί κανείς να παρατηρήσει και πρακτικά την ανταλλαγή που γίνεται μεταξύ της ακρίβειας και της πολυπλοκότητας της δομής. Άρα, υπό τη συνήθη υπόθεση ότι τα στιγμιότυπα είναι ανεξάρτητα των παραμέτρων η συνάρτηση BIC γίνεται:

$$BIC(S \mid D) = \ln \prod_{i=1}^{N} P(d_i \mid \widehat{\boldsymbol{\theta}}_S) - \frac{size(S)}{2} \ln N$$

$$BIC(S \mid D) = \sum_{i=1}^{N} \ln P(d_i \mid \widehat{\boldsymbol{\theta}}_S) - \frac{size(S)}{2} \ln N$$

Μάλιστα, γνωρίζουμε ότι η πιθανοφάνεια είναι:

$$L(\boldsymbol{\theta}) = \prod_{i=1}^{n} \prod_{\boldsymbol{u_i} \in Val(pa_{X_i})} \left[ \prod_{x_i \in Val(X_i)} \theta_{x_i \mid \boldsymbol{u_i}}^{M[\boldsymbol{u_i}, x_i]} \right]$$

με

$$\hat{\theta}_{x_i \mid \boldsymbol{u_i}} = \frac{M[\boldsymbol{u_i}, x_i]}{M[\boldsymbol{u_i}]}$$

Τότε, κατά αντιστοιχία μπορούμε να γράψουμε:

$$BIC(S \mid D) = \ln L(\hat{\boldsymbol{\theta}}_S) - \frac{size(S)}{2} \ln N$$

$$BIC(S \mid D) = \ln \prod_{i=1}^{n} \prod_{j=1}^{q_i} \prod_{k=1}^{r_i} \left( \frac{N_{ijk}}{N_{ij}} \right)^{N_{ijk}} - \frac{size(S)}{2} \ln N$$

$$BIC(S \mid D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \ln \left( \frac{N_{ijk}}{N_{ij}} \right) - \frac{\ln N}{2} \sum_{i=1}^{n} q_i (r_i - 1)$$

όπου:

- ο $N_{ijk}$ είναι ο αριθμός εμφάνισης στα δεδομένα, της $k$ περίπτωσης τιμής της $X_i$ μεταβλητής (κόμβου) με τον $j$ συνδυασμό τιμών των γονιών της $X_i$.
- ο $N_{ij}$ είναι ο αριθμός εμφάνισης στα δεδομένα του $j$ συνδυασμού τιμών των γονιών της $X_i$.
- $n$ το πλήθος των μεταβλητών (κόμβων)
- και $q_i, r_i$ το πλήθος των συνδυασμών των τιμών που αφορούν τους γονείς της $X_i$ και το πλήθος των τιμών της $X_i$ αντίστοιχα.

Μια ακόμα συνάρτηση βαθμολόγησης, η οποία θα χρησιμοποιηθεί στη μελέτη περίπτωσης, είναι η AIC (Akaike Information Criterion) (Bozdogan, 1987)που δίνεται ως εξής:

$$AIC = 2k - 2 \ln L(\hat{\boldsymbol{\theta}}_S)$$

όπου $k$ είναι το πλήθος των ανεξάρτητων παραμέτρων και $L(\hat{\boldsymbol{\theta}}_S)$ είναι η μέγιστη τιμή της πιθανοφάνειας.

Αν συγκρίνουμε τους δύο όρους πολυπλοκότητας στις συναρτήσεις βαθμολόγησης BIC και AIC μπορούμε να παρατηρήσουμε ότι η BIC προσδίδει μεγαλύτερη ποινή για την πολυπλοκότητα σε σχέση με την AIC, όταν το πλήθος των δεδομένων είναι

μεγάλο. Πράγματι, αν $size(S) = k$ είναι το πλήθος των ανεξάρτητων παραμέτρων της δομής τότε συγκρίνοντας τους δύο όρους πολυπλοκότητας έχουμε:

$$k \ln N > 2k \Rightarrow \ln N > 2 \Rightarrow e^{\ln N} > e^2$$

$$\Rightarrow N > e^2$$

Άρα, όταν έχουμε δεδομένα με πλήθος στιγμιοτύπων $N$ μεγαλύτερο του $e^2$, δηλαδή για $N \geq 8$, η BIC τιμωρεί την πολυπλοκότητα σε μεγαλύτερο βαθμό από την AIC.

Μια ιδιότητα που είναι επιθυμητό να έχει μια συνάρτηση βαθμολόγησης, όπως θα φανεί παρακάτω, είναι η ικανότητά της να αποσυντίθεται σε μικρότερα μέρη (decomposition property). Γενικά, μια συνάρτηση βαθμολόγησης καλείται decomposable αν μπορεί να γραφτεί ως το άθροισμα τοπικών βαθμολογήσεων με κάθεμια από αυτές να αντιστοιχεί σε μια οικογένεια κόμβων της δομής. Δηλαδή, αν $X_i, pa_{X_i}$ είναι η οικογένεια του κόμβου $X_i$ τότε μια decomposable συνάρτηση βαθμολόγησης θα είναι το άθροισμα των βαθμολογιών των οικογενειών κάθε κόμβου:

$$score(D, S) = \sum_{i=1}^{n} score(X_i, pa_{X_i}, D)$$

όπου $score(X_i, pa_{X_i}, D)$ είναι η τοπική βαθμολόγηση της οικογένειας $X_i, pa_{X_i}$. Η συνάρτηση BIC είναι μια decomposable συνάρτηση βαθμολόγησης (decomposable score function). Πράγματι:

$$BIC(S \mid D) = \sum_{i=1}^{n} \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \ln\left(\frac{N_{ijk}}{N_{ij}}\right) - \frac{\ln N}{2} \sum_{i=1}^{n} q_i (r_i - 1)$$

$$BIC(S \mid D) = \sum_{i=1}^{n} \left[ \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \ln\left(\frac{N_{ijk}}{N_{ij}}\right) - \frac{1}{2} q_i (r_i - 1) \ln N \right]$$

όπου η τοπική βαθμολόγηση της κάθε οικογένειας είναι:

$$score\left(X_i, pa_{X_i}, D\right) = \left[ \sum_{j=1}^{q_i} \sum_{k=1}^{r_i} N_{ijk} \ln\left(\frac{N_{ijk}}{N_{ij}}\right) - \frac{1}{2} q_i (r_i - 1) \ln N \right]$$

Δεδομένου πλέον των συναρτήσεων βαθμολόγησης, ο τρόπος εύρεσης των υποψήφιων μπεϋζιανών δικτύων βασίζεται σε heuristic διαδικασίες. Γενικά, κατά τις διαδικασίες αυτές, γίνονται σε βήματα μικρές αλλαγές σε ένα DAG, κάθεμια από τις οποίες για να

είναι νόμιμη θα πρέπει να οδηγεί πάλι σε DAG. Οι αλλαγές αυτές έχουν τοπικό χαρακτήρα και αφορούν την πρόσθεση ακμής μεταξύ δύο μη γειτονικών κόμβων, την αφαίρεση της ακμής μεταξύ δύο κόμβων και την αντιστροφή της κατεύθυνσης μιας ακμής. Κάθε μια από αυτές τις αλλαγές καλείται χειριστής (operator) και κάθε operator δουλεύει υπό τη συνθήκη ότι αυτό που θα προκύψει θα πρέπει να είναι πάλι ένα DAG.

Μια τετοια οικογένεια από heuristic αλγορίθμους είναι οι αλγόριθμοι greedy search, κατά τους οποίους μπορεί κανείς να ξεκινήσει με μια άδεια δομή χωρίς ακμές (empty structure) ή με μια τυχαία δομή (random structure) ή ακόμα και με μια εκ των προτέρων διαμορφωμένη δομή DAG (prior structure) από είδικούς. Δύο χαρακτηριστικοί αλγόριθμοι greedy search είναι ο hc (hill-climbing) και ο tabu search, που ξεκινάνε από ένα δίκτυο και προσθέτουν, αφαιρούν και αντιστρέφουν ακμές μέσω των operators.

Γενικά η λογική των greedy search αλγορίθμων περιγράφεται από τους Jensen και Nielsen (2007) ως εξής:

1. Έστω $S$ μια αρχική δομή που μπορεί να είναι άδεια, τυχαία ή εκ των προτέρων διαμορφωμένη.
2. Επαναλαμβάνουμε τη παρακάτω διαδικασία.
    - Υπολογίζουμε $\Delta(A)$ τη βαθμολογία για κάθε νόμιμο χειρισμό $A$ με $\Delta^* = max_A \Delta(A)$ για $A = A^*$ να είναι ο χειρισμός που μεγιστοποιεί τη βαθμολογία.
    - Εάν $\Delta^* > 0$, τότε θέτουμε τη νέα δομή $S$ να είναι εκείνη που παράχθηκε από τον χειρισμό $(op)$ με τη μεγαλύτερη βαθμολογία $op(S, A^*)$, δηλαδή $S = op(S, A^*)$.
3. Η επανάληψη γίνεται μέχρι $\Delta^* \leq 0$, δηλαδή μέχρι να μην μπορεί να βαθμολογηθεί θετικά ο καλύτερα βαθμολογούμενος χειρισμός $A^*$.

Αν επιλεγεί ο hc αλγόριθμος, η διαδικασία ξεκινάει με μία δομή $S$ και χρησιμοποιώντας για παράδειγμα την BIC υπολογίζει τη βαθμολογία του δικτύου. Στο επόμενο βήμα δημιουργείται ένα σύνολο πιθανών δομών με τη πρόσθεση, αφαίρεση και αντιστροφή πάντα μιας ακμής στην υπάρχουσα δομή $S$. Η νέα υποψήφια δομή DAG είναι εκείνη που θα έχει τη μεγαλύτερη βαθμολογία από τη BIC. Η διαδικασία επαναλαμβάνεται μέχρι να βρεθεί το DAG με τη μεγαλύτερη βαθμολογία.

Σε αυτό το σημείο μπορεί κανείς να παρατηρήσει τη χρησιμότητα των decomposable συναρτήσεων βαθμολόγησης όπως η BIC. Για παράδειγμα αν προστεθεί μια ακμή με κατεύθυνση ανάμεσα στους κόμβους $X_i, X_j$ τέτοια ώστε $X_i \rightarrow X_j$ έχει ως αποτέλεσμα την αλλαγή της οικογένειας του $X_j$. Προηγουμένως η οικογένεια του ήταν $X_j, pa_{X_j}$ ενώ πλέον είναι $X_j, pa_{X_j} \cup \{X_i\}$. Χρησιμοποιώντας μια decomposable συνάρτηση βαθμολόγησης μπορούν να βαθμολογηθούν οι οικογένειες και να παρθεί η διαφορά τους ώστε να αξιολογηθεί αυτή η αλλαγή στη δομή:

$$\Delta(X_i \rightarrow X_j) = score\left(X_j, pa_{X_j} \cup \{X_i\}, D\right) - score(X_i, pa_{X_i}, D)$$

$$\Delta(X_i \rightarrow X_j) = BIC\left(X_j, pa_{X_j} \cup \{X_i\}, D\right) - BIC(X_i, pa_{X_i}, D)$$

Συνεπώς, σε δύο κόμβους $X_i, X_j$ η πρόσθεση και η αφαίρεση μιας ακμής ανάμεσα τους με $X_i \rightarrow X_j$ αλλάζει την οικογένεια του $X_j$ και στην περίπτωση που ήδη υπήρχε ακμή η αντιστροφή της κατεύθυνσης $X_i \leftarrow X_j$ αλλάζει τις οικογένειες και των δύο κόμβων. Γενικά, μέσω των decomposable συναρτήσεων βαθμολόγησης που βαθμολογούν την οικογένεια κάθε κόμβου, κάθε τέτοια αλλαγή μπορεί πλέον να αξιολογηθεί.

Είναι σημαντικό να σημειωθεί πως οι greedy search αλγόριθμοι δεν εξασφαλίζουν άμεσα την εύρεση της δομής με τη μεγαλύτερη δυνατή βαθμολογία αλλά καταλήγουν σε ένα τοπικό μέγιστο. Για την αποφυγή της παραμονής σε ένα τοπικό μέγιστο και την πληρέστερη διερεύνυση του πεδίου των λύσεων, χρησιμοποιούνται greedy search αλγόριθμοι με πολλαπλές επανεκκινήσεις (greedy search with multiple restarts). Σύμφωνα με αυτή την προσέγγιση, αφού βρεθεί μια δομή που αποτελεί τοπικό μέγιστο, ο αλγόριθμος επανεκκινεί με μια τυχάια δομή καταλήγοντας σε άλλη δομή. Αυτό γίνεται για έναν προκαθορισμένο αριθμό επαναλήψεων όπου στο τέλος επιλέγεται η δομή με τη μεγαλύτερη βαθμολογία (Gelman *et al*. 2013; Jensen and Nielsen, 2007; Kjaerulff and Madsen, 2013; Scutari and Denis, 2014).

### *4.4.3 Δομικός Αλγόριθμος EM*

Εύλογα δημιουργείται το ερώτημα τι γίνεται στην περίπτωση της εύρεσης της δομής ενός μπεϋζιανού δικτύου όταν δεν έχουμε πλήρη δεδομένα. Στην περίπτωση που δεν έχουμε πλήρη δεδομένα και θέλουμε βρούμε τη δομή του δικτύου, μπορούμε να χρησιμοποιήσουμε τον δομικό αλγόριθμο EM (structural EM) όπως χρησιμοποιήσαμε αντίστοιχα τον αλγόριθμο EM για την εύρεση παραμέτρων από μη πλήρη δεδομένα.

Ας υποθέσουμε ότι έχουμε μια υποψήφια δομή $G$ και τα μη πλήρη δεδομένα $D$. Η γενική προσέγγιση προτείνει την συμπλήρωση των δεδομένων, η οποία θα γίνει με βάση το σύνολο των παραμέτρων $\overline{\boldsymbol{\theta}}_G$, που θα είναι οι παράμετροι που μεγιστοποιούν την posterior συνάρτηση πιθανότητας των παραμέτρων (MAP παράμετροι) για την υποψήφια δομή $G$. Στη συνέχεια με βάση τα συμπληρωμένα δεδομένα $D^*_{G,\overline{\boldsymbol{\theta}}_G}$ μπορούμε με κάποια συνάρτηση βαθμολόγησης να αξιολογήσουμε κάθε αλλαγή που επιβάλλει ένας operator ($o$) στην υποψήφια δομή $G$. Δηλαδή:

$$\Delta_{D^*_{G,\overline{\boldsymbol{\theta}}_G}}(G : o) = score\left(o(G) : D^*_{G,\overline{\boldsymbol{\theta}}_G}\right) - score(G : D^*_{G,\overline{\boldsymbol{\theta}}_G})$$

όπου:

- $score\left(o(G) : D^*_{G,\overline{\boldsymbol{\theta}}_G}\right)$ είναι η βαθμολογία της νεας δομής $o(G)$ ύστερα από την λειτουργία του operator $o$ με βάση τα συμπληρωμένα δεδομένα $D^*_{G,\overline{\boldsymbol{\theta}}_G}$.
- $score(G : D^*_{G,\overline{\boldsymbol{\theta}}_G})$ είναι η βαθμολογία της δομής $G$ με βάση τα συμπληρωμένα δεδομένα $D^*_{G,\overline{\boldsymbol{\theta}}_G}$.

Κεντρικό ρόλο στον δομικό ΕΜ διαδραματίζει το ακόλουθο θεώρημα των Koller και Friedman (2009):

Έστω $G_0$ μια γραφική δομή και $\overline{\boldsymbol{\theta}}_0$ οι MAP παράμετροι για την $G_0$, δεδομένου ενός συνόλου δεδομένων $D$. Τότε για κάθε γραφική δομή $G$:

$$BIC\left(G : D^*_{G_0,\overline{\boldsymbol{\theta}}_0}\right) - BIC\left(G_0 : D^*_{G_0,\overline{\boldsymbol{\theta}}_0}\right) \leq BIC(G : D) - BIC(G_0 : D)$$

όπου:

- η διαφορά $BIC(G : D) - BIC(G_0 : D)$ είναι η πραγματική βελτίωση της βαθμολογίας με τη συνάρτηση BIC για την δομή $G$ σε σχέση με τη δομή $G_0$ γιατί χρησιμοποιεί τα πραγματικά δεδομένα $D$.
- η διαφορά $BIC\left(G : D^*_{G_0,\overline{\boldsymbol{\theta}}_0}\right) - BIC\left(G_0 : D^*_{G_0,\overline{\boldsymbol{\theta}}_0}\right)$ είναι η εκτιμώμενη βελτίωση της βαθμολογίας, αφού χρησιμοποιούμε το συμπληρωμένο σύνολο δεδομένων $D^*_{G_0,\overline{\boldsymbol{\theta}}_0}$ το οποίο συμπληρώθηκε χρησιμοποιώντας της δομή $G_0$.

Αυτό σημαίνει ότι η πραγματική βελτίωση της βαθμολογίας με τη συνάρτηση BIC για την δομή $G$ σε σχέση με τη δομή $G_0$ είναι τουλάχιστον ίση είναι με την εκτιμώμενη βελτίωση της βαθμολογίας όταν χρησιμοποιούμε το συμπληρωμένο σύνολο

δεδομένων $D^*_{G,\bar{\boldsymbol{\theta}}_0}$. Άρα αν $\boldsymbol{\theta}_0$ είναι οι αρχικές παράμετροι της δομής $G_0$ μπορούμε να γράψουμε συνοπτικά την παραπάνω σχέση ως εξής:

$$\Delta_{D^*_{G_0,\bar{\boldsymbol{\theta}}_0}} \leq \Delta_{D_{G_0,\boldsymbol{\theta}_0}}$$

Συνεπώς, από το θεώρημα εξασφαλίζεται ένα κάτω φράγμα για την πραγματική μεταβολή $\Delta_{D_{G_0,\boldsymbol{\theta}_0}}$ της βαθμολογίας BIC. Το κάτω φράγμα είναι η εκτιμώμενη μεταβολή $\Delta_{D^*_{G_0,\bar{\boldsymbol{\theta}}_0}}$ της βαθμολογίας BIC που σημαίνει ότι η βελτίωση της θα βελτιώσει αναγκαστικά και την πραγματική βαθμολογία. Το συγκεκριμένο αποτέλεσμα δεν εξασφαλίζεται μόνο μια αλλαγή της δομής $G_0$, δηλαδή την εφαρμογή ενός μόνο operator $o$, αλλά ισχύει και για πολλές αλλαγές δηλαδή για αυθαίρετα μεγάλες ακολουθίες operators για τη δομή $G_0$.

Ο δομικός αλγόριθμος EM όπως παρουσιάζεται από τους Koller και Friedman (2009). Εστω $G^0$ η αρχική δομή ενός μπευζιανού δικτύου, $\boldsymbol{\theta}_0$ οι αρχικές παράμετροι για τη δομή $G^0$ και $D$ τα μη πλήρη δεδομένα. Τότε:

1. για κάθε $t = 0,1,\ldots$ επανέλαβε μέχρι να επιτευχθεί σύγκλιση
2. // προαιρετικό βήμα εύρεσης παραμέτρων
3. $\boldsymbol{\theta}^{t'} \leftarrow$ Expectation-Maximization$(G^t, \boldsymbol{\theta}^t, D)$ για την posterior
4. // τρέχοντας τον EM παράγονται τα αναμενόμενα επαρκή στατιστικά για το σύνολο των δεδομένων $D^*_{G^t,\boldsymbol{\theta}^{t'}}$
5. $G^{t+1} \leftarrow$ Εύρεση δομής $(D^*_{G^t,\boldsymbol{\theta}^{t'}})$
6. $\boldsymbol{\theta}^{t+1} \leftarrow$ Εύρεση παραμέτρων $(D^*_{G^t,\boldsymbol{\theta}^{t'}}, G^{t+1})$
7. επέστρεψε $G^t$, $\boldsymbol{\theta}^t$ (αφού έχει γίνει σύγκλιση)

Δηλαδή ξεκινάμε με αρχική δομή $G^0$, μη πλήρη δεδομένα $D$ και αρχικό σύνολο παραμέτρων $\boldsymbol{\theta}_0$. Χρησιμοποιούμε τον αλγόριθμο EM για την posterior κατανομή με βάση τη δομή $G_0$, τις παραμέτρους $\boldsymbol{\theta}_0$ και τα μη πλήρη δεδομένα $D$ και βρίσκουμε το σύνολο των παραμέτρων $\boldsymbol{\theta}^{0'}$ που μεγιστοποιούν την posterior συνάρτηση πιθανότητας. Συμπληρώνουμε το σύνολο των δεδομένων με βάση τις νέες παραμέτρους $\boldsymbol{\theta}^{0'}$ (αφού θα έχουν αλλάξει και τα CPTs) και έχουμε το σύνολο δεδομένων $D^*_{G^0,\boldsymbol{\theta}^{0'}}$. Με βάση το νέο σύνολο δεδομένων $D^*_{G^0,\boldsymbol{\theta}^{0'}}$, κάνουμε εύρεση δομής χρησιμοποιώντας τη συνάρτηση βαθμολόγησης BIC όπως περιγράφηκε προηγουμένως. Για μια ακολουθία από εφαργογές operators στη δομή $G_0$ βελτιώνουμε

τη διαφορά $BIC\left(G^1 : D^*_{G^0,\boldsymbol{\theta}^{0\prime}}\right) - BIC\left(G^0 : D^*_{G^0,\boldsymbol{\theta}^{0\prime}}\right)$, όπου $G^1$ η δομή μετά την εφαρμογή των operators. Η διαφορά με τη σειρά της βελτιώνει την πραγματική διαφορά $BIC(G^1 : D) - BIC(G_0 : D)$ ως αποτέλεσμα του θεωρήματος. Λαμβάνουμε τότε την $G^1$ ως νέα δομή. Στη συνέχεια γίνεται εύρεση των παραμέτρων $\boldsymbol{\theta}^1$ βασισμένη στα συμπληρωμένα δεδομένα $D^*_{G^0,\boldsymbol{\theta}^{0\prime}}$ για την δομή $G^1$. Πλεόν οι παράμετροι $\boldsymbol{\theta}^1$, η δομή $G^1$ και τα μη πλήρη δεδομένα $D$ θα είναι η βάση για την επόμενη επανάληψη όπως μόλις περιγράφτηκε. Όταν γίνει σύγκλιση της εκτιμώμενης διαφοράς

$BIC\left(G^t : D^*_{G^{t-1},\boldsymbol{\theta}^{t-1\prime}}\right) - BIC\left(G^{t-1} : D^*_{G^{t-1},\boldsymbol{\theta}^{t-1\prime}}\right)$ θα γίνει σύγκλιση και της πραγματικής διαφοράς στη βαθμολόγηση των δομών. Τότε, γίνεται επιστροφή της δομής $G^t$ στην οποία κατέληξε ο αλγόριθμος όπως και το αντίστοιχο σύνολο παρμέτρων της δομής $\boldsymbol{\theta}^t$.

Τέλος, θα πρέπει να τονιστεί ότι σε αντίθεση με τον αλγόριθμο ΕΜ που έβρισκε ένα τοπικό μέγιστο της πιθανοφάνειας, ο δομικός αλγόριθμος ΕΜ δεν εξασφαλίζει ότι η δομή στην οποία θα καταλήξει θα είναι τοπικό μέγιστο για το σύνολο των υποψήφιων δομών. Ωστόσο, η δομή που θα καταλήξει εξασφαλίζεται ότι θα έχει καλυτερη βαθμολογία σε σχέση με την αρχική δομή (Koller and Friedman, 2009).

## 4.5 Διαδικασία Πρόβλεψης

Μέχρι στιγμής, έχει γίνει αναφορά σε περιπτώσεις που αφορούν την εύρεση των παραμέτρων με τις μεθόδους που έχουν αναλυθεί έχοντας σταθερά δεδομένα $D$. Εύλογα γεννάται το ερώτημα, τι γίνεται στην περίπτωση που έχουμε εισαγωγή νέων δεδομένων, δηλαδή ποιά θα είναι η πιθανότητα των νέων δεδομένων υπό τη συνθήκη εκείνων που ήδη ξέρουμε.

Για να απαντηθεί το ερώτημα, ας υποθέσουμε ότι έχουμε πάλι τη γενική περίπτωση ενός μπεϋζιανού δικτύου με τυχαίες μεταβλητές $X_1, X_2, \dots, X_n$ και σύνολο δεδομένων $D = \{d[1], \dots, d[M]\}$, όπου $d[m]$, με $m = 1, \dots, M$, αντιστοιχεί στο m στιγμιότυπο του συνόλου των M στιγμιοτύπων των δεδομένων. Επιπλέον, έστω $\boldsymbol{\theta}$ το διάνυσμα των παραμέτρων, με $\boldsymbol{\theta} \in \Theta$ όπου $\Theta$ ο παραμετρικός χώρος. Αν $d[M+1]$ είναι ένα νέο στιγμιότυπο δεδομένων με

$$d[M+1] = \{X_1[M+1], X_2[M+1], \dots, X_n[M+1]\}$$

και $X_i[M+1], i = 1, \dots, n$ να είναι οι τιμές των $X_1, X_2, \dots, X_n$ στο $M+1$ στιγμιότυπο, τότε το ζητούμενο είναι να βρεθεί η πιθανότητα $P(d[M+1] \mid D)$. Άρα:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} P(d[M+1], \boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

όπου $P(d[M+1] \mid D)$ η περιθώρια δεσμευμένη συνάρτηση πιθανότητας της από κοινού δεσμευμένης συνάρτησης πιθανότητας $P(d[M+1], \boldsymbol{\theta} \mid D)$. Με την εφαρμογή του θεωρηματος της δεσμευμένης πιθανότητας έχουμε:

$$P(d[M+1] \mid \boldsymbol{\theta}, D) = \frac{P(d[M+1], \boldsymbol{\theta} \mid D)}{P(\boldsymbol{\theta} \mid D)} \Leftrightarrow$$

$$\Leftrightarrow P(d[M+1], \boldsymbol{\theta} \mid D) = P(d[M+1] \mid \boldsymbol{\theta}, D) P(\boldsymbol{\theta} \mid D)$$

Άρα η $P(d[M+1] \mid D)$ γίνεται:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} P(d[M+1] \mid \boldsymbol{\theta}, D) P(\boldsymbol{\theta} \mid D)\, d\boldsymbol{\theta}$$

Όμως, όπως ήδη έχουμε αναλύσει, τα στιγμιότυπα είναι ανεξάρτητα υπό τη συνθήκη των παραμέτρων και άρα $d[M+1]$ και $D$ θα είναι ανεξάρτητα υπό τη συνθήκη του $\boldsymbol{\theta}$. Τελικά, η πρόβλεψη των νέων δεδομένων διαμορφώνεται ως εξής:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} P(d[M+1] \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$$

και αν $\boldsymbol{\theta} = (\theta_{X_1|pa_{X_1}}, \theta_{X_2|pa_{X_2}}, \dots, \theta_{X_n|pa_{X_n}})$ τότε:

$P(d[M+1] \mid D) =$

$$= \int_{\theta_{X_n|pa_{X_n}}} \dots \int_{\theta_{X_1|pa_{X_1}}} P(d[M+1] \mid \theta_1, \theta_2, \dots, \theta_m) P(\theta_1, \theta_2, \dots, \theta_m|D)\, d\theta_{X_1|pa_{X_1}} \dots d\theta_{X_n|pa_{X_n}}$$

Παρατηρούμε ότι η $P(\boldsymbol{\theta}|D)$ είναι η posterior κατανομή των δεδομένων και αποτελεί μια συνάρτηση πυκνότητας πιθανότητας της $\boldsymbol{\theta}$. Επιπλέον, το στιγμιότυπο $d[M+1]$ είναι ένα δείγμα τιμών των $X_1, X_2, \dots, X_n$ και άρα η $P(d[M+1] \mid \boldsymbol{\theta})$ είναι η συνάρτηση πιθανοφάνειας του στιγμιότυπου $d[M+1]$ δηλαδή είναι επίσης μια συνάρτηση της $\boldsymbol{\theta}$.

Μπορούμε λοιπόν να γράψουμε:

$P(\boldsymbol{\theta}|D) = f(\boldsymbol{\theta})$

$P(d[M+1] \mid \boldsymbol{\theta}) = g(\boldsymbol{\theta})$

άρα σύμφωνα με τη μέση τιμή συνάρτησης τυχαίων μεταβλητών έχουμε:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} g(\boldsymbol{\theta}) f(\boldsymbol{\theta})\, d\boldsymbol{\theta} = E[g(\boldsymbol{\theta})]$$

αν αντικαταστήσουμε ξανά έχουμε:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} P(d[M+1] \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} = E[P(d[M+1] \mid \boldsymbol{\theta})]$$

Δηλαδή, βρίσκουμε το πόσο πιθανό είναι να συμβεί το στιγμιότυπο $d[M+1]$ για όλες τις δυνατές τιμές της $\boldsymbol{\theta}$ και ύστερα υπολογίζουμε τη μέση τιμή αυτών που βρήκαμε. Αυτή η μέση τιμή, θα είναι η εκτίμηση του να συμβεί ο συγκεκριμένος συνδυασμός νέων δεδομένων υπό τη συνθήκη των παλιών δεδομένων. Συνεπώς, μπορούμε να καταλήξουμε στο συμπέρασμα ότι η εκτίμηση ενός νέου στιγμιότυπου υπό τη συνθήκη των δεδομένων είναι η μέση τιμή της πιθανοφάνειας του.

Φυσικά, τα δεδομένα $D$ αποτελούν τα δεδομένα εκπαίδευσης του μοντέλου, ενώ το νέο στιγμιότυπο μπορεί να αποτελεί μέρος των δεδομένων επιβεβαίωσης που χρησιμοποιούνται για να εξετάσουμε τη συμπεριφορά του αλγόριθμου σε δεδομένα

στα οποία δεν έχει εκπαιδευτεί (Bernardo and Smith, 2000; Darwiche, 2009; Koller and Friedman, 2009).

Η πρόβλεψη του νέου στιγμιότυπου των δεδομένων μπορεί να αναλυθεί ακόμα περαιτέρω αν ληφθούν υπόψιν η δομή του μπεϋζιανού δικτύου που καθορίζει τις υπό συνθήκη ανεξαρτησίες των μεταβλητών και η ολική αποσύνθεση της posterior κατανομής. Πράγματι έχουμε ότι:

$$P(d[M+1] \mid \boldsymbol{\theta}) = P(X_1[M+1], X_2[M+1], \ldots, X_n[M+1] \mid \boldsymbol{\theta})$$

$$= \prod_{i=1}^{n} P\left(X_i \mid pa_{X_i}[M+1], \boldsymbol{\theta}\right)$$

$$= \prod_{i=1}^{n} P\left(X_i \mid pa_{X_i}[M+1], \theta_{X_i|pa_{X_i}}\right)$$

αυτό γίνεται διότι, όπως ήδη γνωρίζουμε, οι τυχαίες μεταβλητές είναι ανεξάρτητες μεταξύ τους υπό συνθήκη των γονιών τους και τις αφορούν μόνο οι παράμετροι που είναι οι αντίστοιχες προς αυτές και όχι όλο το σύνολο των παραμέτρων. Επιπλέον, αν λάβουμε υπόψιν την ολική αποσύνθεση της posterior κατανομής έχουμε:

$$P(\boldsymbol{\theta}|D) = \prod_{i=1}^{n} P(\theta_{X_i|pa_{X_i}} \mid D)$$

Άρα το γινόμενο $P(d[M+1] \mid \boldsymbol{\theta})P(\boldsymbol{\theta}|D)$ γράφεται:

$$P(d[M+1] \mid \boldsymbol{\theta})P(\boldsymbol{\theta}|D) =$$

$$= \prod_{i=1}^{n} P\left(X_i[M+1] \mid pa_{X_i}[M+1], \theta_{X_i|pa_{X_i}}\right) \prod_{i=1}^{n} P(\theta_{X_i|pa_{X_i}} \mid D)$$

$$= \prod_{i=1}^{n} P\left(X_i[M+1] \mid pa_{X_i}[M+1], \theta_{X_i|pa_{X_i}}\right) P(\theta_{X_i|pa_{X_i}} \mid D)$$

Συνεπώς για την εκτίμηση των νέων δεδομένων έχουμε:

$$P(d[M+1] \mid D) =$$

$$= \int_{\theta_{X_n|pa_{X_n}}} \ldots \int_{\theta_{X_1|pa_{X_1}}} \prod_{i=1}^{n} P\left(X_i[M+1] \mid pa_{X_i}[M+1], \theta_{X_i|pa_{X_i}}\right) P(\theta_{X_i|pa_{X_i}} \mid D)\, d\theta_{X_1|pa_{X_1}} \ldots d\theta_{X_n|pa_{X_n}}$$

$$= \int_{\theta_{X_1|pa_{X_1}}} P\left(X_1[M+1] \,\middle|\, pa_{X_1}[M+1], \theta_{X_1|pa_{X_1}}\right) P(\theta_{X_1|pa_{X_1}} | D) d\theta_{X_1|pa_{X_1}} \dots$$

$$\dots \int_{\theta_{X_n|pa_{X_n}}} P\left(X_n[M+1] \,\middle|\, pa_{X_n}[M+1], \theta_{X_n|pa_{X_n}}\right) P(\theta_{X_n|pa_{X_n}} | D) d\theta_{X_n|pa_{X_n}}$$

$$= \prod_{i=1}^{n} \int_{X_i|pa_{X_i}} P\left(X_i[M+1] \,\middle|\, pa_{X_i}[M+1], \theta_{X_i|pa_{X_i}}\right) P(\theta_{X_i|pa_{X_i}} | D)\, d\theta_{X_i|pa_{X_i}}$$

Στο σημείο αυτό, έχει αξία να αναφερθούν κάποια απλά παραδείγματα πρόβλεψης νέων δεδομένων λαμβάνοντας υπόψιν και το είδος της κατανομής που ακολουθεί η posterior κατανομή. Ας υποθέσουμε ότι έχουμε μια διωνυμική μεταβλητή $X$ με πιθανές τιμές της $x^0, x^1$ και με $P(X = x^1) = \theta$. Επιπλέον, έχουμε ένα σύνολο δεδομένων $D$ όπου $M$ το πλήθος των στιγμιότυπων, με $m$ να συμβολίζει το $m$ στιγμιότυπο των δεδομένων και $1 \leq m \leq M$. Από προηγούμενη ανάλυση γνωρίζουμε ότι αν επιλέξουμε για prior της $\theta$ μια κατανομή Βήτα με υπερπαραμέτρους $a_0, a_1$ τότε η posterior κατανομή θα επίσης μια κατανομή Βήτα με παραμέτρους $a_1 + M[1]$ και $a_0 + M[0]$, όπου $M[1]$ και $M[0]$ μας δίνουν το πλήθος των εμφανίσεων της τιμής $x^1$ και $x^0$ στο σύνολο των δεδομένων αντίστοιχα. Αν λοιπόν θέλουμε να κάνουμε πρόβλεψη για ένα νέο στιγμιότυπο $d[M+1]$ όπου $d[M+1] = X[M+1] = x^1$ θα έχουμε ότι:

$$P(d[M+1] \mid D) = \int_{\theta \in \Theta} P(d[M+1] \mid \theta) P(\theta|D)\, d\theta = E[P(d[M+1] \mid \theta)]$$

όμως υπό την υπόθεση ότι η πιθανότητα η $X$ να πάρει την τιμή $x^1$ παραμένει η ίδια δεδομένου της παραμέτρου, με $P(d[M+1] \mid \theta) = \theta$ τότε:

$$P(d[M+1] \mid D) = \int_{\theta \in \Theta} \theta P(\theta|D)\, d\theta$$

Παρατηρούμε ότι το ολοκλήρωμα είναι η μέση τιμή της της posterior κατανομής της $\theta$:

$$E[\theta] = \int_{\theta \in \Theta} \theta P(\theta|D)\, d\theta$$

άρα:

$$P(d[M+1] \mid D) = E[\theta]$$

Η κατανομή όμως που ακολουθεί η $\theta$, αφού έχουν ληφθεί υπόψιν τα δεδομένα, είναι η posterior κατανομή, που είναι μια κατανομή Βήτα με παραμέτρους $a_1 + M[1]$ και $a_0 + M[0]$. Άρα η μέση τιμή θα είναι:

$$P(d[M+1] \mid D) = E[\theta] = \frac{a_1 + M[1]}{a_1 + M[1] + a_0 + M[0]}$$

που μπορεί να γραφτεί και ως εξής:

$$P(d[M+1] \mid D) = E[\theta] = \frac{a_1 + M[1]}{a + M}$$

με $M[0] + M[1] = M$ να είναι το άθροισμα των εμφανίσεων των τιμών $x^0, x^1$ στο σύνολο των δεδομένων, που όπως είναι λογικό στην περίπτωσή μας θα κάνει το πλήθος των στιγμιότυπων $M$ και $a_0 + a_1 = a$ είναι το άθροσμα των παραμέτρων της prior κατανομής.

Άρα καταλήγουμε στο συμπέρασμα ότι η πρόβλεψη του νέου στιγμιότυπου, σε μια διωνυμική μεταβλητή με prior μια κατανομή βήτα, είναι η μέση τιμή της κατανομής της τυχαίας μεταβλητής της παραμέτρου $\theta$, δηλαδή η μέση τιμή της posterior κατανομής, που είναι μια Βήτα κατανομή.

Μια ακόμα περίπτωση, είναι εκείνη που έχουμε μια πολυωνυμική μεταβλητή $X$ με $x^1, x^2, \ldots, x^K$ διαφορετικές τιμές. Ξέρουμε τότε ότι η K διαστάσεων τυχαία μεταβλητή $\boldsymbol{X} = (X_1, X_2, \ldots, X_K)$, με τις τυχαίες μεταβλητές $X_1, X_2, \ldots, X_K$ να είναι το πλήθος εμφάνισης των διαφορετικών τιμών $x^1, x^2, \ldots, x^K$ κατά αντιστοιχία στο σύνολο των δεδομένων, ακολουθεί πολυωνυμική κατανομή. Έστω $D = \{d[1], d[2], \ldots, d[M]\}$ το σύνολο δεδομένων όπου:

$$d[m] = X[m], (m = 1, \ldots, M)$$

να είναι το υπ'αριθμόν m στιγμιότυπο των δεδομένων με $d[m] = X[m]$.

Επιπλέον, έστω $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_K)$ να είναι η K-διάστατη τυχαία μεταβλητή των παραμέτρων, με $P(X = x^\kappa) = \theta_\kappa$ για κάθε $\kappa = 1, \ldots, K$. Από προηγούμενη ανάλυση γνωρίζουμε ότι αν επιλέξουμε για prior της $\boldsymbol{\theta}$ μια κατανομή Dirichlet με

υπερπαραμέτρους $\alpha_1, \alpha_2, \ldots, \alpha_{K,}$ τότε και η posterior κατανομή θα είναι μια κατανομή Dirichlet με παραμέτρους $\alpha_1 + M[1],\ \alpha_2 + M[2], \ldots,\ \alpha_K + M[K]$ με $M[\kappa]$ να είναι το πληθος εμφάνισης της τιμής $x^\kappa$ στα δεδομένα.

Τότε, αν $d[M+1] = X[M+1] = x^\kappa$ είναι ένα νέο στιγμιότυπο των δεδομένων, η πιθανότητα να συμβεί υπό τη συνθήκη των υπόλοιπων δεδομένων είναι η μέση τιμή της πιθανοφάνειας του στιγμιότυπου:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} P(d[M+1] \mid \boldsymbol{\theta}) P(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} = E[P(d[M+1] \mid \boldsymbol{\theta})]$$

όμως:

$$P(d[M+1] \mid \boldsymbol{\theta}) = P(X[M+1] = x^\kappa \mid \boldsymbol{\theta}) = \theta_\kappa$$

άρα:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} \theta_\kappa P(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta}$$

όπου η $P(\boldsymbol{\theta}|D)$ είναι η συνάρτηση πυκνότητας πιθανότητας της Dirichlet posterior κατανομής. Αυτό όμως σημαίνει ότι η $P(d[M+1] \mid D)$ είναι η μέση τιμή της $\theta_\kappa$:

$$P(d[M+1] \mid D) = \int_{\boldsymbol{\theta} \in \Theta} \theta_\kappa P(\boldsymbol{\theta}|D)\, d\boldsymbol{\theta} = E[\theta_\kappa]$$

Η μέση τιμή της Dirichlet κατανομής για την $\theta_\kappa$ θα είναι:

$$E[\theta_\kappa] = \frac{\alpha_\kappa + M[\kappa]}{\alpha + M}$$

με $\alpha = \alpha_1 + \alpha_2 + \cdots + \alpha_K$ και $M = M[1] + M[2] + \cdots + M[K]$.

Άρα καταλήγουμε στο συμπέρασμα ότι η πρόβλεψη του νέου στιγμιότυπου με τιμή $x^\kappa$, σε μια πολυωνυμική μεταβλητή όπου η prior είναι μια κατανομή Dirichlet, είναι η μέση τιμή της αντίστοιχης παραμέτρου $\theta_\kappa$ της πολυδιάστατης μεταβλητής $\boldsymbol{\theta}$ που η κατανομή της είναι η Dirichlet κατανομή της posterior (Koller and Friedman, 2009; Neapolitan, 2003).

### *4.5.1 Μέθοδος Διαγραφής Μεταβλητών*

Η μπεϋζιανή συμπερασματολογία (Bayesian inference) είναι κομμάτι της διαδικασίας πρόβλεψης και αφορά την εξαγωγή συμπερασμάτων που επικαιροποιούν την

πεποίθησή μας (belief updating) για την κατατάσταση των μεταβλητών υπό την επήρεια κάποιων δεδομένων (evidence). Τα συμπεράσματα αυτά αποτελούν απαντήσεις σε ερωτήσεις (queries) που γίνονται για τις μεταβλητές και βασίζονται στη χρήση του θεωρήματος δεσμευμένης πιθανότητας και στην ανάλυση που έχει ήδη γίνει με τα CPTs.

Γενικά, έστω ότι έχουμε ένα μπεϋζιανό δίκτυο με $X_1, X_2, \ldots, X_n$ τυχαίες μεταβλητές και με $\boldsymbol{e} = (x_{m+1}, x_{m+2}, \ldots, x_n)$ να είναι οι τιμές των $X_{m+1}, X_{m+2}, \ldots, X_n$ $(m \leq n)$ που γνωρίζουμε ως δεδομένες (evidence). Τότε η απάντηση στην ερώτηση ποια είναι η πιθανότητα της $X_1 = x_1$ δεδομένου του $\boldsymbol{e}$ είναι:

$$P(x_1 \mid \boldsymbol{e}) = \frac{P(x_1, \boldsymbol{e})}{P(\boldsymbol{e})}$$

όπου η $P(x_1, \boldsymbol{e})$ και $P(\boldsymbol{e})$ είναι περιθώριες συναρτήσεις πιθανότητας της από κοινού συνάρτησης πιθανότητας $P(X_1, X_2, \ldots, X_n)$ του δικτύου με:

$$P(x_1, \boldsymbol{e}) = P(x_1, x_{m+1}, x_{m+2}, \ldots, x_n) = \sum_{x_m} \ldots \sum_{x_2} P(x_1, X_2, \ldots, X_m, \ldots, x_n)$$

$$P(\boldsymbol{e}) = P(x_{m+1}, x_{m+2}, \ldots, x_n) = \sum_{x_m} \ldots \sum_{x_1} P(X_1, X_2, \ldots, X_m, x_{m+1}, x_{m+2}, \ldots, x_n)$$

άρα:

$$P(x_1 \mid \boldsymbol{e}) = \frac{\sum_{x_m} \ldots \sum_{x_2} P(x_1, X_2, \ldots, X_m, \ldots, x_n)}{\sum_{x_m} \ldots \sum_{x_1} P(X_1, X_2, \ldots, X_m, x_{m+1}, x_{m+2}, \ldots, x_n)}$$

Θα μπορούσε επίσης να εξαχθούν συμπεράσματα για την πιθανότητα να παίρνει συγκεκριμένη τιμή μια μεταβλητή ή ένα υποσύνολο των μεταβλητών, χρησιμοποιώντας τις περιθώριες συναρτήσεις πιθανότητας της από κοινού συνάρτησης πιθανότητας του δικτύου. Δηλαδή, η απάντηση στην ερώτηση του ποια είναι η πιθανότητα η μεταβλητή $X_i$ να παίρνει την τιμή $x_i$ είναι:

$$P(x_i) = \sum_{x_n} \ldots \sum_{x_{i-1}} \sum_{x_{i+1}} \ldots \sum_{x_1} P(X_1, \ldots, X_{i-1}, x_i, X_{i+1} \ldots, X_n)$$

Μπορεί κανείς να παρατηρήσει ότι όσο αυξάνεται το πλήθος των μεταβλητών τόσο αυξάνεται και το πλήθος των υπολογισμών που πρέπει να γίνουν. Το γεγονός αυτό καθιστά την εξαγωγή συμπερασμάτων όλο και πιο δύσκολη. Η παραγοντοποίηση της από κοινού συνάρτησης πιθανότητας με βάση τις υπό συνθήκη ανεξαρτησίες που αντανακλά η δομή του δικτύου μειώνει τους υπολογισμούς αλλά δεν είναι πάντα αρκετή. Η λύση στο συγκεκριμένο πρόβλημα βασίζεται στη μέθοδο διαγραφής των μεταβλητών (variable elimination).

Με βάση το παράδειγμα των Koller και Friedman (2009), ας υποτεθεί ότι έχουμε ένα μπεϋζιανό δίκτυο με μεταβλητές $A, B, C, D$, όπου κάθεμια μπορεί να πάρει δύο τιμές και η δομή του δικτύου είναι $A \to B \to C \to D$. Τότε η από κοινού συνάρτηση πιθανότητας του δικτύου είναι:

$$P(A, B, C, D) = P(A)P(B \mid A)P(C \mid B)P(D \mid C)$$

Αν συμβολίσουμε με $a^1, a^2, b^1, b^2, c^1, c^2, d^1, d^2$ τιμές των $A, B, C, D$ αντίστοιχα, τότε η πιθανότητα η $D$ να πάρει τις τιμές $d^1$ ή $d^2$ γράφεται ως εξής:

$$P(d^1) = \sum_C \sum_B \sum_A P(A, B, C, d^1) = \sum_C \sum_B \sum_A P(A)P(B \mid A)P(C \mid B)P(d^1 \mid C)$$

$$P(d^2) = \sum_C \sum_B \sum_A P(A, B, C, d^2) = \sum_C \sum_B \sum_A P(A)P(B \mid A)P(C \mid B)P(d^2 \mid C)$$

παρατηρούμε όμως ότι το γινόμενο $P(C \mid B)P(d^1 \mid C)$ δεν εξαρτάται από τις τιμές της $A$ και άρα μπορεί να βγεί εκτός του αθροίσματος:

$$P(d^1) = \sum_C \sum_B P(C \mid B)P(d^1 \mid C) \sum_A P(A)P(B \mid A)$$

επειδή όμως ούτε ο υπολογισμός της πιθανότητας $P(d^1 \mid C)$ εξαρτάται από τις τιμές της $B$, μπορεί και αυτός ο όρος να βγεί εκτός του αθροίσματος:

$$P(d^1) = \sum_C P(d^1 \mid C) \sum_B P(C \mid B) \sum_A P(A)P(B \mid A)$$

Αν θεωρήσουμε την συνάρτηση $\varphi_1 : Val(B) \to \mathbb{R}$, όπου $Val(B)$ το σύνολο τιμών της $B$, τότε έχουμε $\varphi_1(B) = \sum_A P(A)P(B \mid A)$ όπου:

$$\varphi_1(b^1) = P(a^1)P(b^1 \mid a^1) + P(a^2)P(b^1 \mid a^2)$$

$$\varphi_1(b^2) = P(a^1)P(b^2 \mid a^1) + P(a^2)P(b^2 \mid a^2)$$

άρα:

$$P(d^1) = \sum_C P(d^1 \mid C) \sum_B P(C \mid B)\, \varphi_1(B)$$

όπου έχουμε διαγράψει την μεταβλητή $A$ αφού όλη η επιρροή της καλύπτεται από την $\varphi_1$.Ομοίως θεωρούμε την $\varphi_2: Val(C) \to \mathbb{R}$, με $\varphi_2(C) = \sum_B P(C \mid B)\, \varphi_1(B)$ όπου:

$$\varphi_2(c^1) = P(c^1 \mid b^1)\varphi_1(b^1) + P(c^1 \mid b^2)\varphi_1(b^2)$$

$$\varphi_2(c^2) = P(c^2 \mid b^1)\varphi_1(b^1) + P(c^2 \mid b^2)\varphi_1(b^2)$$

και παρατηρούμε ότι τις $\varphi_1(b^1), \varphi_1(b^2)$ δεν τις υπολογίζουμε ξανά διότι τις ξέρουμε από τη προηγούμενη ανάλυση. Άρα:

$$P(d^1) = \sum_C P(d^1 \mid C)\, \varphi_2(C)$$

$$P(d^1) = P(d^1 \mid c^1)\varphi_2(c^1) + P(d^1 \mid c^2)\varphi_2(c^2)$$

όπου πλέον έχουμε διαγράψει την μεταβλητή $B$ αφού όλη η επιρροή της καλύπτεται από την $\varphi_2$. Με την ίδια λογική μπορούμε να βρούμε την $P(d^2)$, δηλαδή χρησμοποιώντας πάλι τις $\varphi_1, \varphi_2$ έχουμε:

$$P(d^2) = \sum_C P(d^2 \mid C) \sum_B P(C \mid B) \sum_A P(A)P(B \mid A)$$

$$P(d^2) = \sum_C P(d^2 \mid C) \sum_B P(C \mid B)\, \varphi_1(B)$$

$$P(d^2) = \sum_C P(d^2 \mid C)\, \varphi_2(C)$$

$$P(d^2) = P(d^2 \mid c^1)\varphi_2(c^1) + P(d^2 \mid c^2)\varphi_2(c^2)$$

Το αποτέλεσμα είναι ότι καταλήξαμε σε έναν απλούστερο τύπο για τον υπολογισμό των $P(d^1),\ P(d^2)$ έχοντας διαγράψει τις μεταβλητές $A, B$. Όσον αφορά το πλήθος των υπολογισμών για να βρεθούν οι $P(d^1), P(d^2)$ χρειαζόμαστε 4 γινόμενα και 2 αθροίσματα για τις $\varphi_1(b^1), \varphi_1(b^2), 4$ γινόμενα και 2 αθροίσματα για τις $\varphi_2(c^1), \varphi_2(c^2)$ και αφού ξέρουμε τις τιμές των $\varphi_1, \varphi_2$ για τις $P(d^1),\ P(d^2)$, χρειαζόμαστε επιπλέον 4 γινόμενα και 2 αθροίσματα. Άρα στο σύνολο χρειάζεται να κάνουμε 18 το πλήθος υπολογισμούς, αριθμός κατά πολύ μικρότερος εκείνου που θα χρειαζόταν αν υπολογίζαμε τις $P(d^1), P(d^2)$ μέσω της από κοινού συνάρτησης πιθανότητας (48 γινόμενα και 14 αθροίσματα σύνολο). Συνεπώς, με τη μέθοδο διαγραφής μεταβλητών καταλήξαμε σε μια απλούστερη μορφή για την περιθώρια συνάρτηση πιθανότητας $P(D)$, χρησιμοποιήθηκαν λιγότεροι υπολογισμοί για την εύρεση των πιθανοτήτων και στην ουσία διαγράφονται όλες οι μεταβλητές $A, B, C$ που δεν συμμετείχαν στην ερώτηση που έγινε. Το τελευταίο μπορεί εύκολα να γίνει αντιληπτό αν κάνουμε ένα παραπάνω βήμα και θέσουμε:

$$\varphi_3(D) = \sum_C P(D \mid C)\, \varphi_2(C)$$

Πράγματι, η εύρεση της περιθώριας $P(D)$ ανάγεται στον υπολογισμό της $\varphi_3$ που αφορά μόνο την μεταβλητή $D$ για την οποία γίνεται η ερώτηση:

$$P(D) = \varphi_3(D)$$

Οι συναρτήσεις $\varphi_1, \varphi_2, \varphi_3$που χρησιμοποιήθηκαν στη μέθοδο διαγραφής μεταβλητών καλούνται παράγοντες (factors) και γενικά διαθέτουν κάποιες σημαντικές ιδιότητες. Έστω ότι έχουμε ένα σύνολο μεταβλητών $\boldsymbol{X}$ και μια μεταβλητή $Y$ που δεν ανήκει στο σύνολο $\boldsymbol{X}$. Επιπλέον, έστω $\varphi_{\boldsymbol{X},Y}(\boldsymbol{X}, Y)$ να είναι ο παράγοντας των $\boldsymbol{X}, Y$ τότε η συνάρτηση:

$$\varphi_{\boldsymbol{X}}(\boldsymbol{X}) = \sum_Y \varphi_{\boldsymbol{X},Y}(\boldsymbol{X}, Y)$$

είναι ο παράγοντας των μεταβλητών $\boldsymbol{X}$. Δηλαδή, αθροίζοντας τις τιμές της μιας μεταβλητής παράγεται ο παράγοντας των υπόλοιπων μεταβλητών.

Πράγματι, στο παράδειγμα που προηγήθηκε θα μπορούσαμε να θεωρήσουμε ως παράγοντα $\varphi_{A,B}$ των $A, B$ την $\varphi_{A,B}(A,B) = P(A,B) = P(A)P(B \mid A)$. Τότε αθροίζοντας όλες τις τιμές της $A$ θα πάρουμε τον παράγοντα $\varphi_B$ της $B$:

$$\varphi_B(B) = \sum_A \varphi_{A,B}(A,B) = \sum_A P(A)P(B \mid A)$$

ο οποίος ταυτίζεται με την $\varphi_1(B)$. Επίσης, αν έχουμε δύο παράγοντες $\varphi_1, \varphi_2$ και η μεταβλητή $X$ δεν εμφανίζεται στον $\varphi_1$ τότε:

$$\sum_X \varphi_1 \varphi_2 = \varphi_1 \sum_X \varphi_2$$

αθροίζοντας τις τιμές της $X$ ο $\varphi_1$ μπορεί να βγεί εκτός αθροίσματος. Αυτό είναι ένα γεγονός που μπορούμε να παρατηρήσαμε και στο παράδειγμα αν θεωρήσουμε ότι αρχικά οι παράγοντες είναι:

$\varphi_A = P(A)$, $\varphi_{B|A} = P(B \mid A)$, $\varphi_{C|B} = P(C \mid B)$ και $\varphi_{D|C} = P(D \mid C)$

Τότε η περιθώρια συνάρτηση πιθανότητας $P(D)$ γράφεται πλέον ώς εξής:

$$P(D) = \sum_C \sum_B \sum_A P(A)P(B \mid A)P(C \mid B)P(D \mid C)$$

$$P(D) = \sum_C \sum_B \sum_A \varphi_A \ \varphi_{B|A} \ \varphi_{C|B} \ \varphi_{D|C}$$

και σύμφωνα με την ιδιότητα των παραγόντων έχουμε:

$$P(D) = \sum_C \sum_B \varphi_{C|B} \ \varphi_{D|C} \sum_A \varphi_A \varphi_{B|A}$$

$$P(D) = \sum_C \varphi_{D|C} \sum_B \varphi_{C|B} \sum_A \varphi_A \varphi_{B|A}$$

όπου μπορούμε τώρα να εφαρμόσουμε ακριβώς την ίδια διαδικασία διαγραφής μεταβλητών μόνο που πλέον:

$$\varphi_1(B) = \sum_A \varphi_A \varphi_{B|A}\,, \qquad \varphi_2(C) = \sum_B \varphi_{C|B} \ \varphi_1(B), \qquad \varphi_3(D) = \sum_C \varphi_{D|C} \ \varphi_2(C)$$

και όπως είδαμε, η εύρεση της περιθώριας συνάρτησης πιθανότητας να ανάγεται στον υπλογισμό των τιμών ενός παράγοντα:

$$P(D) = \varphi_3(D)$$

Γενικά, έστω ένα μπεϋζιανό δίκτυο με σύνολο μεταβλητών $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$ και ένα σύνολο μεταβλητών $\boldsymbol{Y}$, υποσύνολο του $\boldsymbol{X}$, για το οποίο γίνεται ένα ερώτημα με σκοπό να εξαχθεί κάποιο συμπέρασμα. Τότε με $\boldsymbol{Z} = \boldsymbol{X} - \boldsymbol{Y}$ συμβολίζουμε το σύνολο των μεταβλητών του δικτύου που δεν εχουν σχέση με το ερώτημα. Επιπλέον, θεωρούμε ως $\Phi$ το σύνολο των παραγόντων του δικτύου όπου για κάθε CPT θα αντιστοιχεί και ένα παράγοντας δηλαδή:

$$\Phi = \{\varphi_{X_1}, \ldots, \varphi_{X_n}\}$$

με $\varphi_{X_i} = P(X_i \mid pa_{X_i})$ να είναι η συνάρτηση που θα μου δίνει τις πιθανότητες τις $X_i$ δεδομένου των γονιών της $pa_{X_i}$ για κάθε $i = 1, \ldots, n$. Τότε, μπορούμε να ορίσουμε τον παράγοντα των μεταβλητών $\boldsymbol{Y}$ ως εξής:

$$\varphi_Y(\boldsymbol{Y}) = \sum_{\boldsymbol{Z}} \prod_{i=1}^{n} \varphi_{X_i}$$

όπου στη συνέχεια ακολουθείτε η διαδικασία διαγραφής των μεταβλητών $\boldsymbol{Z}$ όπως περιγράφηκε στο παράδειγμα για εκείνες τις μεταβλητές που δεν σχετίζονταν με το ερώτημα.

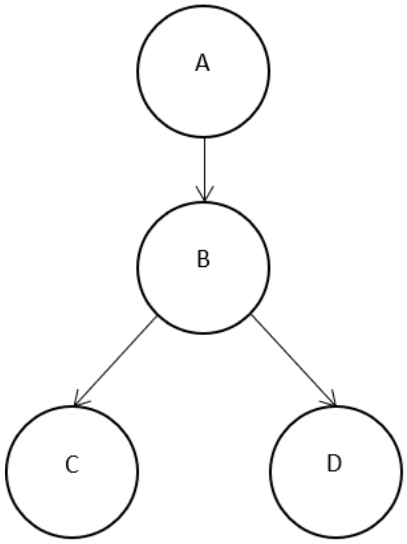
Ωστόσο, για μπορέσουν να εξαχθούν συμπεράσματα με βάση τα ερωτήματα που τίθονται για τις διάφορες μεταβλητές, θα πρέπει να είναι γνωστά τα CPTs. Αυτό σημάινει πως θα πρέπει να είναι γνωστό το μοντέλο του μπεϋζιανού δικτύου και άρα οι κατανομές που ακολουθούν οι μεταβλητές του δικτύου. Συνεπώς, θα πρέπει από τα δεδομένα που δίνονται να εξαχθεί το μοντέλο του δικτύου ώστε να μπορεί να εφαρμοστεί η μπεϋζιανή συμπερασματολογία (Darwiche, 2009; Koller and Friedman, 2009; Korb and Nicholson 2011; Neapolitan, 2003).

### *4.5.2 Στάθμιση Πιθανοφάνειας (Likelihood Weighting)*

Η προσεγγιστική συμπερασματολογία (approximate inference) είναι η απάντηση στην περίπτωση που οι υπολογισμοί που θα πρέπει να γίνουν για να εξαχθούν τα

συμπεράσματα είναι δυσεπίλυτοι και καθιστούν την όλη διαδικασία μη αποδοτική. Η στάθμιση πιθανοφάνειας (likelihood weighting) είναι μια μέθοδος συμπερασματολογίας που εντάσσεται στην κατηγορία της προσεγγιστικής συμπερασματολογίας. Η συγκεκριμένη μέθοδος θα χρησιμοποιηθεί σε επόμενο κεφάλαιο για την εκτίμηση νέων στιγμιοτύπων. Η likelihood weighting χρησιμοποεί έννοιες όπως η τοπολογική διάταξη (topological ordering) και η προς τα εμπρός δειγματοληψία (forward sampling), οι οποίες παρουσιάζονται στο συγκεκριμένο υποκεφάλαιο.

Αρχικά, ας υποτεθεί ότι έχουμε το DAG ενός μπεϋζιανού δικτύου και $X_i, X_j$ δύο διαφορετικές μεταβλητές του δικτύου. Αν υπάρχει μια διαδρομή με κατεύθυνση από τη $X_i$ στη $X_j$, τότε η $X_i$ προηγείται της $X_j$ στη διάταξη. Στην περίπτωση που δεν υπάρχει τέτοια διαδρομή τότε οποιαδήποτε διάταξη ανάμεσα στις δύο είναι ισοδύναμη. Για παράδειγμα έστω ένα DAG με μεταβλητές $A, B, C, D$ (Εικόνα 23). Τότε οι υποψήφιες τοπολογικές διατάξεις $A, B, C, D$ και $A, B, D, C$ είναι ισοδύναμες.




**Εικόνα 23. Παράδειγμα για την διασαφήνιση της τοπολογικής διάταξης.**

Στη συνέχεια ας υποθέσουμε πάλι ότι έχουμε ένα μπεϋζιανό δίκτυο με $X_1, \dots, X_n$ μεταβλητές και θέλουμε να βρούμε την πιθανότητα $P(x_1, \dots, x_n) = P(\boldsymbol{x})$. Τότε, η διαδικασία του forward sampling λειτουργεί ως εξής:

1. Έστω $X_1, X_2, \dots, X_n$ η τοπολογική διάταξη των μεταβλητών του δικτύου.
2. Για κάθε $i = 1, \dots, n$

- Τοποθέτησε $u_i \leftarrow \boldsymbol{x}\{pa_{X_i}\}$, δηλαδή τοποθετούμε στο $u_i$ τις τιμές των γονιών της $X_i$ που βρίσκονται στο $\boldsymbol{x}$ ανάμεσα στα $x_1, \dots, x_{i-1}$. Αν δεν έχει γονείς, γίνεται δειγματοληψία από την περιθώρια κατανομή του $\boldsymbol{x}$.
- Πάρε δείγμα $\hat{x}_i$ από την κατανομή $P(X_i \mid u_i)$, δηλαδή από τον CPT της $X_i$.

3. Επέστρεψε το δείγμα $(\hat{x}_1, \dots, \hat{x}_n)$.

Η διάταξη των μεταβλητών στο βήμα 1 μας επιτρέπει να έχουμε ήδη τιμές $u_i$ για τους γονείς της $X_i$ έτσι ώστε στο βήμα 2 να μπορεί να επιλεχθεί $\hat{x}_i$ από την κατανομή $P(X_i \mid u_i)$. Με το συγκεκριμένο τρόπο παράγεται ένα δείγμα ή αλλιώς ένα particle $(\hat{x}_1, \dots, \hat{x}_n)$ από την από κοινού κατανομή πιθανότητας του μπεϋζιανού δικτύου $P(X_1, \dots, X_n)$. Αν θέλουμε να παράξουμε $M$ το πλήθος δείγματα επαναλαμβάνουμε τη διαδικασία για $m = 1, \dots, M$ φορές και παράγουμε ένα σύνολο $(d[1], \dots, d[M])$ τέτοιων δειγμάτων. Τότε η πιθανότητα $P(x_1, \dots, x_n)$ μπορεί να βρεθεί ως εξής:

$$P(x_1, \dots, x_n) \approx \frac{\pi\lambda\acute{\eta}\theta o\varsigma\ \tau\omega\nu\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu\ \mu\epsilon\ X_1 = x_1, \dots, X_n = x_n}{\pi\lambda\acute{\eta}\theta o\varsigma\ \tau\omega\nu\ \delta\epsilon\iota\gamma\mu\acute{\alpha}\tau\omega\nu\ M}$$

Γενικά, από ένα σύνολο δειγμάτων $D = (d[1], \dots, d[M])$ που έχουν παραχθεί από την διαδικασία δειγματοληψίας forward sampling μπορούμε να βρούμε τη μέση τιμή οποιασδήποτε συνάρτησης $f$ ως εξής:

$$\hat{E}_D(f) = \frac{1}{M} \sum_{m=1}^{M} f(d[m])$$

Ενώ στην περίπτωση που θέλουμε να υπολογίσουμε την πιθανότητα $P(\boldsymbol{y})$ για το ενδεχόμενο να έχουμε τιμές $\boldsymbol{y}$ έχουμε το εξής:

$$\hat{P}_D(\boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{I}(\boldsymbol{y}[m] = \boldsymbol{y})$$

όπου $y[m] = \boldsymbol{y}$ δηλώνει το δείγμα $d[m]$ με τις τιμές $\boldsymbol{y}$ να έχουν τοποθετηθεί στο δείγμα και η συνάρτηση $I(y[m] = \boldsymbol{y})$ είναι:

$$\boldsymbol{I}(y[m] = \boldsymbol{y}) = \begin{cases} 1, \alpha\nu\ \tau o\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\ d[m]\ \pi\epsilon\rho\iota\acute{\epsilon}\chi\epsilon\iota\ \tau\iota\varsigma\ \tau\iota\mu\acute{\epsilon}\varsigma\ \boldsymbol{y} \\ 0, \alpha\nu\ \tau o\ \delta\epsilon\acute{\iota}\gamma\mu\alpha\ d[m]\ \delta\epsilon\nu\ \pi\epsilon\rho\iota\acute{\epsilon}\chi\epsilon\iota\ \tau\iota\varsigma\ \ \boldsymbol{y} \end{cases}$$

δηλαδή η $\hat{P}_D(\boldsymbol{y})$ είναι η σχετική συχνότητα εμφάνισης των τιμών $\boldsymbol{y}$ στο σύνολο των δειγμάτων $D = (d[1], \dots, d[M])$ που παράχθηκε με forward sampling. Ο τρόπος που επιλέγεται το δείγμα $\hat{x}_i$ από την κατανομή $P(X_i \mid u_i)$ στο βήμα 2 του forward sampling είναι ο εξής:

Αν υποτεθεί ότι η τυχαία μεταβλητή $X$ ακολουθεί πολυωνυμική κατανομή με τιμές $x^1, \dots, x^k$ και αντίστοιχες παραμέτρους $\theta_1, \dots, \theta_k$ τότε:

- Παράγεται ένα δείγμα $s$ με ομοιόμορφα τυχαίο τρόπο από το διάστημα $[0,1]$, δηλαδή το $s$ παράγεται με ψευδο-τυχαίο τρόπο από μια ομοιόμορφη κατανομή.
- Το διάστημα $[0,1]$ χωρίζεται σε $k$ το πλήθος υποδιαστήματα $[0,\ \theta_1], [\theta_1,\ \theta_1 + \theta_2], [\theta_1 + \theta_2,\ \theta_1 + \theta_2 + \theta_3], \dots$ έτσι ώστε το $i$ διάστημα να είναι το:

$$[\sum_{j=1}^{i-1} \theta_j\ , \sum_{j=1}^{i} \theta_j]$$

- Αν το δείγμα $s$ ανήκει στο $i$ διάστημα τότε επιλέγεται ώς δείγμα η τιμή $x^i$

Άρα σε κάθε βήμα παραγωγής δείγματος $d[m]$ με τη forward sampling, για κάθε μεταβλητή $X_i$ παράγεται ένα $s$ με ψευδο-τυχαίο τρόπο, το οποίο χρησιμοποιείται όπως αναλύθηκε για να παραχθεί το δείγμα $\hat{x}_i$ από το βήμα 2 που θα συμπληρώνει το $d[m]$.

Έχοντας πλέον υπόψιν την προηγούμενη ανάλυση, έστω ένα μπεϋζιανό δίκτυο με διάνυσμα μεταβλητών $\boldsymbol{X}$ και το γεγονός $\boldsymbol{Z} = \boldsymbol{z}$**,** όπου ένα υποσύνολο μεταβλητών $\boldsymbol{Z}$ έχει πάρει τις τιμές $\boldsymbol{z}$. Η μέθοδος likelihood weighting για την παραγωγή δειγμάτων είναι η εξής:

1. Έστω $X_1,\ X_2, \dots, X_n$ η τοπολογική διάταξη των μεταβλητών του δικτύου.
2. Τοποθέτησε $w \leftarrow 1$, όπου $w$ συμβολίζει την έννοια του βάρους.
3. Για κάθε $i = 1, \dots, n$
   3.1. Τοποθέτησε $u_i \leftarrow \boldsymbol{x}\{pa_{X_i}\}$, δηλαδή τοποθετούμε στο $u_i$ τις τιμές των γονιών της $X_i$ που βρίσκονται στο $\boldsymbol{x}$ ανάμεσα στα $x_1, \dots, x_{i-1}$.
   3.2. Εάν η μεταβλητή $X_i$ δεν ανήκει στο $\boldsymbol{Z}$
      3.2.1. Επιλέγεται δείγμα $x_i$ από την κατανομή $P(X_i \mid u_i)$
   3.3. Εάν η μεταβλητή $X_i \in \boldsymbol{Z}$
      3.3.1. Τοποθέτησε $x_i \leftarrow \boldsymbol{z}\{X_i\}$, δηλαδή τοποθέτησε την τιμή της $X_i$ από το $\boldsymbol{z}$ σαν $x_i$
      3.3.2. Τοποθέτησε $w \leftarrow w * P(x_i \mid u_i)$, δηλαδή το βάρος θα είναι η πιθανότητα της $X_i = x_i$ δεδομένου των τιμών $u_i$ των γονιών της.
4. Επέστρεψε το δείγμα $(x_1, \dots, x_n), w$

Δηλαδή αν $X_i$ δεν ανήκει στο $\boldsymbol{Z}$, το δείγμα $x_i$ παράγεται με forward sampling, ενώ αν η τιμή $x_i$ ανήκει στο γεγονός $\boldsymbol{Z} = \boldsymbol{z}$, θα επιλεχθεί η ίδια και θα προστεθεί στο συνολικό δείγμα $d[m]$ μαζί με την πιθανότητα να συμβεί η $x_i$ υπό τη συνθήκη των τιμών των γονιών της $u_i$, που την έχουμε έτοιμη από τον αντίστοιχο CPT. Η πιθανότητα αυτή θα είναι το βάρος $w$.

Επιπλέον, αν βρεθούν στο $\boldsymbol{z}$ οι τιμές από διάφορες μεταβλητές αυτό σημαίνει ότι το τελικό βάρος στο δείγμα $(x_1, \dots, x_n), w$ θα είναι το γινόμενο των πιθανοτήτων $P(x_i \mid u_i)$, όπως μπορεί κανείς να παρατηρήσει από το βήμα 3.3.2. Με αυτό το τρόπο παράγεται ένα δείγμα ή ένα particle με βάρος (weighted particle) και αν επαναληφθεί η διαδικασία $M$ το πλήθος φορές θα παραχθεί ένα σύνολο δεδομένων $D = ((d[1], w[1]), \dots, (d[M], w[m]))$ από δείγματα με βάρος. Συνεπώς, για να βρεθεί η πιθανότητα του $\boldsymbol{y}$ δεδομένου ότι έχει συμβεί το γεγονός $\boldsymbol{e}$, χρησιμοποιείται το σύνολο που παράχθηκε από την likelihood weighting ως εξής:

$$\hat{P}_D(\boldsymbol{y} \mid \boldsymbol{e}) = \frac{\sum_{m=1}^{M} w[m] * \boldsymbol{I}(\boldsymbol{y}[m] = \boldsymbol{y})}{\sum_{m=1}^{M} w[m]}$$

που αποτελεί γενίκευση της.

$$\hat{P}_D(\boldsymbol{y}) = \frac{1}{M} \sum_{m=1}^{M} \boldsymbol{I}(\boldsymbol{y}[m] = \boldsymbol{y})$$

Τέλος, η μέθοδος της likelihood weighting βελτιώνει τη μέθοδο forward sampling αφού αποφεύγει το εξής πρόβλημα: Με τη forward sampling θα μπορούσε να παραχθεί δείγμα $x_i$ το οποίο ανήκει στο υπό συνθήκη γεγονός, δηλαδή γνωρίζουμε ακριβώς σε εκείνο το στιγμιότυπο της τιμή που πήρε ο κόμβος $X_i$. Αυτό σημαίνει ότι το δείγμα θα απορρίπτεται αφού δεν θα είναι συμβατό με τα υπάρχοντα στοιχεία. Με τη likelihood weighting κάτι τέτοιο αποφεύγεται, αφού αν η τιμή $x_i$ ανήκει στο γεγονός που έχει συμβεί δεν επιλέγεται και παράγεται στη θέση της το αντίστοιχο βάρος (Koller and Friedman, 2009; Korb and Nicholson, 2011).

# 5 Μελέτη Περίπτωσης

Στο κεφάλαιο αυτό η περίπτωση που τίθεται προς μελέτη αφορά τη δουλειά των Marvin *et al.*(2017) κατά την οποία στόχος ήταν να γίνει πρόβλεψη κινδύνου νανοϋλικών με τη χρήση μπεϋζιανού δικτύου, λειτουργώντας έτσι υποστηρικτικά στην ανάλυση ρίσκου της ανθρώπινης υγείας. Χρησιμοποιώντας τα δεδομένα των Marvin *et al.*, δοκιμάστηκαν διαφορετικές στρατηγικές εκπαίδευσης τόσο των παραμέτρων όσο και της δομής μπεϋζιανών δικτύων, με σκοπό την εύρεση μίας λογικής που παράγει το βέλτιστο μοντέλο. Στην αρχή του κεφαλαίου γίνεται μία σύντομη παρουσίαση του σετ δεδομένων, στη συνέχεια παρουσιάζεται η μεθοδολογία και το κεφάλαιο κλείνει με παρουσίαση και ανάλυση των αποτελεσμάτων.

## 5.1 Περιγραφή του Προβλήματος

Αρχικά, στην μελέτη των Marvin *et al.*(2017) κατασκευάστηκε ένα μπεϋζιανό δίκτυο ύστερα από τη συμβολή ειδικών επιστημόνων (expert elicitation) στη νανοτοξικολογία. Από τις απαντήσεις που έδωσαν σε συγκεκριμένα ερωτηματολόγια καθορίστηκαν οι μεταβλητές του μπεϋζιανού δικτύου, το εύρος τιμών των μεταβλητών καθώς και οι σχέσεις ανάμεσα στις μεταβλητές. Έπειτα συγκεντρώθηκε ένα σύνολο δεδομένων χρησιμοποιώντας σχετική με το θέμα βιβλιογραφία. Με βάση αυτό το σύνολο δεδομένων, η αρχική δομή του δικτύου αναθεωρήθηκε με τη χρήση του δομικού αλγόριθμου EM. Το μπεϋζιανό δίκτυο με τη νέα δομή είναι εκείνο που χρησιμοποιήθηκε για να γίνουν οι προβλέψεις για τον κίνδυνο που δύνανται να έχουν για την υγεία του ανθρώπου διάφορα νανοϋλικα.

Το μπεϋζιανό δίκτυο αποτελείται από 20 διακριτές τυχαίες μεταβλητές, με 11 από αυτές να αντιστοιχούν σε αίτια που θα μπορούσαν να επηρεάσουν τη δυνατότητα ενός νανοϋλικού να είναι επικίνδυνο για τον άνθρωπο, 8 από αυτές να αντιστοιχούν σε πιθανά αποτελέσματα που θα μπορούσαν να υπάρξουν και 1 μεταβλητή που ομαδοποιεί τις μεταβλητές που αντιστοιχούν σε αποτελέσματα. Οι μεταβλητές που θεωρούνται αίτια, ή αλλιώς οι μεταβλητές εισόδου, χωρίζονται σε φυσικοχημικές ιδιότητες (physicochemichal properties), στον τύπο μελέτης (study type) και στον τρόπο χορήγησης (administration route) (Πίνακας 21).

**Πίνακας 21. Μεταβλητές εισόδου στο μπεϋζιανό δίκτυο.**

| **Μεταβλητές Εισόδου** |
|---|
| *Τύπος Μελέτης* |
| "Study type" |
| *Φυσικοχημικές Ιδιότητες (physicochemichal properties)* |
| "Shape" |
| "Nanoparticle" |
| "Dissolution" |
| "Surface area" |
| "Surface charge" |

| |
|---|
| “Surface coatings” |
| “Surface reactivity” |
| “Aggregation” |
| “Particle size” |
| *Τρόπος Χορήγησης* |
| “Administration route” |

Οι μεταβλητές που αντιστοιχούν σε πιθανά αποτελέσματα ή αλλιώς οι μεταβλητές εξόδου αφορούν 8 πιθανές βιολογικές συνέπειες (biological effects) και όλες έχουν το ίδιο εύρος τιμών. Δηλαδή οι πιθανές καταστάσεις κάθε τέτοιας μεταβλητής είναι “None”, “Low”, “Medium” και “High”, όπου αφορούν σε καθόλου, χαμηλές, μέτριες και υψηλές συνέπειες αντίστοιχα. Οι 8 μεταβλητές βιολογικών συνεπειών ομαδοποιούνται στη μεταβλητή “NM hazard”, που δηλώνει σε τι βαθμό ένα νανοϋλικό είναι επικίνδυνο για την υγεία. Η μεταβλητή “NM hazard” έχει το ίδιο εύρος και τις ίδες καταστάσεις με τις υπόλοιπες μεταβλητές εξόδου. Δηλαδή “None”, “Low”, “Medium” και “High”, για καθόλου, χαμηλό, μέτριο και υψηλό κίνδυνο αντίστοιχα. Η σχέση που συνδέει τη μεταβλητή “NM hazard” με τις υπόλοιπες μεταβλητές εξόδου είναι η εξής:

$$HR_i = \sum_{k=1}^{8} BE_{ik}$$

όπου $HR_i$ είναι η βαθμολογία του κινδύνου του νανοϋλικού για την περίπτωση (στιγμιότυπο) $i$ των δεδομένων και $BE_{ik}$ είναι η βαθμολογία της $k$ βιολογικής συνέπειας στην $i$ περίπτωση.

Η αντιστοιχία μεταξύ των καταστάσεων και των βαθμών είναι η παρακάτω:

- “None” → 0
- “Low” → 1
- “Medium” → 2
- “High” → 3

που σημαίνει ότι αν για παράδειγμα έχουμε $HR_i = 1$ τότε το νανϋλικό που εξετάζεται είναι χαμηλού κινδύνου σε σχέση με την υγεία του ανθρώπου και μόνο μία εκ των 8 βιολογικών συνεπειών είχε κατάσταση “Low”, με όλες τις υπόλοιπες να σημειώνουν

κατάσταση “None”. Οι μεταβλητές εξόδου του προβλήματος παρουσιάζονται στον Πίνακα 22.

**Πίνακας 22. Μεταβλητές εξόδου στο μπεϋζιανό δίκτυο.**

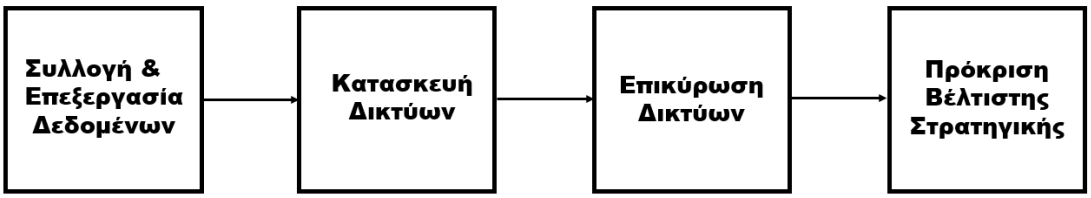
| **Μεταβλητές Εξόδου** |
|---|
| *Βιολογικές Συνέπειες* |
| “Neurological effects” |
| “RCNS effects” (reaches central nervous system) |
| “Pulmonary effects” |
| “Immunological effects” |
| “Inflammation” |
| “Fibrosis” |
| “Cytotoxicity” |
| “Genotoxicity” |
| *Κίνδυνος Νανοϋλικού* |
| “NM hazard” |

Το σύνολο των δεδομένων που χρησιμοποίηθηκε στην έρευνα αποτελείται από 559 στιγμιότυπα για τις 20 μεταβλητές που αναφέρθηκαν. Στην έρευνα των Marvin *et al* (2017), από αυτά τα στιγμιότυπα, τα 91 επιλέχθηκαν τυχαία και χρησιμοποιήθηκαν για την αξιολόγηση της ικανότητας πρόβλεψης του δικτύου ενώ τα υπόλοιπα 468 χρησιμοποιήθηκαν για την διαδικασία μάθησης. Δηλαδή, το σύνολο των δεδομένων χωρίστηκε σε ένα train set 468 στιγμιοτύπων και σε ένα test set 91 στιγμιοτύπων. Με βάση τις τιμές των 11 μεταβλητών εισόδου στο test set των 91 στιγμιοτύπων πραγματοποιήθηκε η εκτίμηση για τις 8 μεταβλητές εξόδου των βιολογικών συνεπειών. Το accuracy (στιγμιότυπα των οποίων η κλάση προβλέφτηκε σωστά προς συνολικά στιγμιότυπα) που επετεύχθη για τις 8 βιολογικές συνέπειες ήταν 71%, ενώ το accuracy που επετεύχθη για τη μεταβλητή NM hazard ήταν 72%, όπως αναφέρεται στην έρευνα (Marvin *et al*, 2017). Το τελικό δίκτυο που αναφέρουν στη μελέτη τους, παρουσιάζεται στην Εικόνα 24


**Εικόνα 24. Τελικό μπεϋζιανό δίκτυο της έρευνας των Marvin *et al.* (2017).**

## 5.2 Μεθοδολογία

Όπως αναφέρεται και στο διάγραμμα ροής της Εικόνας 25, αρχικά έγινε η συλλογή των δεδομένων και η επεξεργασία τους, στη συνέχεια ακολούθησε κατασκευή δικτύων και στατιστικών μοντέλων με διάφορες τεχνικές και τέλος πραγματοποιήθηκε η επικύρωση μοντέλων με διάφορα στατιστικά μέτρα. Στο τέλος, προκρίθηκε η καλύτερη στρατηγική δημιουργίας δικτύου και χρησιμοποιώντας αυτή, δομήθηκε ένα τελικό μοντέλο χρησιμοποιώντας τα ίδια train και test set που χρησιμοποιήθηκαν στους Marvin *et al*. (2017). Για τη δημιουργία και εκπαίδευση μπεϋζιανών δικτύων χρησιμοποιήθηκε η *R* (v.3.5.1) και πιο συγκεκριμένα οι ακόλουθες βιβλιοθήκες της R: *bnlearn* (Scutari and Denis, 2014), *Yardstick* (Kuhn and Vaughan, 2020), *Rgraphviz* (Hansen *et al*., 2019) και *Dplyr* (Wickham, *et al*. 2020).




**Εικόνα 25. Διάγραμμα ροής εργασίας.**

### *5.2.1 Επεξεργασία Δεδομένων*

Όσον αφορά τη συλλογή των δεδομένων, γνωρίζουμε ήδη ότι το σύνολο των δεδομένων αποτελείται από 559 στιγμιότυπα τα οποία αφορούν διάφορα νανοϋλικά καθώς και ότι τα δεδομένα δεν είναι πλήρη, δηλαδή υπάρχουν διάφορα στιγμιότυπα όπου οι τιμές των μεταβλητών δεν είναι συμπληρωμένες, κυρίως γιατί το πείραμα που παρήγαγε το στιγμιότυπο δεν αφορούσε τη συγκεκριμένη μεταβλητή (δεν περιελάμβανε μέτρησή της). Επιπλέον, υπάρχουν 20 διαφορετικές μεταβλητές, εκ των οποίων 11 μεταβλητές εισόδου και 9 μεταβλητές εξόδου. Η μεταβλητή για την οποία γίνεται η πρόβλεψη είναι η "NM hazard", η οποία αντιστοιχεί στην επικινδυνότητα ενός νανοϋλικού για την υγεία του ανθρώπου. Οι κλάσεις που θα εκτιμηθούν θα είναι οι δυνατές καταστάσεις της "NM hazard", δηλαδή "None", "Low", "Medium" και "High".

Οι μεταβλητές Dissolution και Immunological effects στο σύνολο των δεδομένων εμφανίζουν μια μοναδική τιμή. Συγκεκριμένα, η Dissolution εμφανίζει μόνο την τιμή

“0 – 25%” και η Immunological effects την τιμή “None”. Επιπλέον, οι μεταβλητές Fibrosis, RCNS effects και Genotoxicity δεν έχουν στιγμιότυπο με την τιμή “High” στο σύνολο των δεδομένων. Εφόσον δεν τροποποιείται η πληροφορία μεταξύ των στιγμιοτύπων για τις μεταβλητές Dissolution και Immunological effects, θα μπορούσαν να διαγραφούν αφού δεν προσφέρουν καμία προβλεπτική ικανότητα στο δίκτυο. Παρόλα αυτά, προτιμήθηκε να μπούνε με τεχνητό τρόπο τα επίπεδα που δεν εμφανίζονται στα στιγμιότυπα γιατί με αυτόν τον τρόπο διατηρείται το πλεονέκτημα που έχουν τα μπεϋζιανά δίκτυα τις επικαιροποίησης του δικτύου με εμφάνιση νέων δεδομένων (τα οποία θα μπορούσαν να καταγράφουν κατηγορίες που δεν εμφανίστηκαν σε αυτό το σετ δεδομένων).

Επίσης, στις 8 μεταβλητές των βιολογικών συνεπειών δεν υπάρχουν παντού τιμές στο σύνολο των δεδομένων. Οπότε το να προβλέψει κανείς τις βιολογικές συνέπειες έχει μεγαλύτερη αβεβαιότητα διότι λόγω των αραιών δεδομένων δεν βρίσκονται όλες οι σχέσεις αίτιου και αιτιατού. Αντιθέτως, λόγω του τρόπου που δομήθηκε η NM hazard, υπήρχε σε κάθε στιγμιότυπο τιμή της μεταβλητής. Άρα, προτιμήθηκε η πρόβλεψη της NM hazard άμεσα μέσω του δικτύου και όχι με τον έμμεσο τρόπο που αρχικά δομήθηκε η μεταβλητή, δηλαδή με κατασκευή της ως άθροισμα των βαθμολογιών των βιολογικών συνεπειών ύστερα από την πρόβλεψή τους. Τέλος, στο dataset υπήρχαν 306 διπλότυπα, δηλαδή στιγμιότυπα τον οποίων η πληροφορία επαναλαμβανόταν ακριβώς. Συνήθως, σε τεχνικές του πεδίου της μηχανικής μάθησης αφαιρούνται τα διπλότυπα, όμως εδώ διατηρήθηκαν για δύο λόγους: αρχικά γιατί αποτελούσαν πληροφορία από ξεχωριστά πειράματα και αφετέρου επειδή στα διακριτά μπεϋζιανά δίκτυα, όπως παρουσιάστηκε και στο κεφάλαιο 4, έχουν μεγάλη σημασία οι σχετικές συχνότητες εμφάνισης μίας κατηγορίας τιμών έναντι των υπόλοιπων.

### *5.2.2 Κατασκευή Μπεϋζιανών Δικτύων*

Γενικά, για την κατασκευή των δικτύων επιλέχθηκαν τέσσερις γενικές περιπτώσεις από διαφορετικές στρατηγικές. Συνοπτικά, η μία περίπτωση περιλαμβάνει χρήση του έτοιμου μπεϋζιανού δικτύου που παρουσιάζουν οι Marvin *et al.* (2017) (Εικόνα 24). Στη συνέχεια, στη δεύτερη περίπτωση το δίκτυο κατασκευάστηκε από τους constrained based αλγόριθμους Grow Shrink (gs) και IAMB. Ακολούθως, έγινε χρήση των score based αλγορίθμων Hill Climbing (hc) και Tabu seasrch (tabu) με βάση τον δομικό αλγόριθμο EM, και παράχθηκε δίκτυο με δύο τρόπους: ξεκινώντας από μία

τυχαία αρχική δομή και ορίζοντας ως πρότερη δομή αυτή που παρουσιάζουν οι Marvin et al. (2017) (Εικόνα 24). Οι συναρτήσεις βαθμολόγησης που χρησιμοποιήθηκαν για την βαθμολόγηση των υποψήφιων δικτύων είναι η BIC και η AIC. Στη συνέχεια, όταν γίνεται αναφορά σε ονομα μεθόδου, τότε γίνεται συσχέτιση με την αντίστοιχη μέθοδο της βιβλιοθήκης *bnlearn.*

Η διαδικασία εύρεσης των παραμέτρων έγινε χρησιμοιώντας τη μέθοδο μέγιστης πιθανοφάνειας (μέθοδος mle) ή τη μπεϋζιανή εκτίμηση παραμέτρων (μέθοδος bayes). Η πρόβλεψη για τα νέα στιγμιότυπα του test set γίνεται είτε με τη likelihood weighting (μέθοδος bayes-lw) είτε τοπικά βάζοντας τις νέες τιμές των γονιών της μεταβλητής που μας ενδιαφέρει και χρησιμοποιώντας την τοπική κατανομή πιθανότητας (μέθοδος parents). Πιο συγκεκριμένα, η parents πηγαίνει στο CPT της μεταβλητής $X_i$ και ανανεώνει τις τιμές των γονιών $pa_{X_i}$ της μεταβλητής βάση των νέων τιμών του προς πρόβλεψη στιγμιοτύπου. Τότε χρησιμοποιεί την τοπική δεσμευμένη κατανομή ώστε να βρεθεί η πιθανότητα $P(X_i \mid pa_{X_i})$. Επιπλεόν, η bayes-lw που αντιστοιχεί στη μέθοδο likelihood weighting, παράγει ένα σύνολο δεδομένων $D$ από στιγμιότυπα με βάρος (weighted samples). Η συγκεκριμένη μέθοδος έχει μία υπερπαράμετρο, τον αριθμό των δειγμάτων, η οποία ρυθμίστηκε κατά τη διαδικασία εκπαίδευσης.

Συγκεντρωτικά, οι μέθοδοι που χρησιμοποιήθηκαν στην μελέτη περίπτωσης είναι οι εξής:

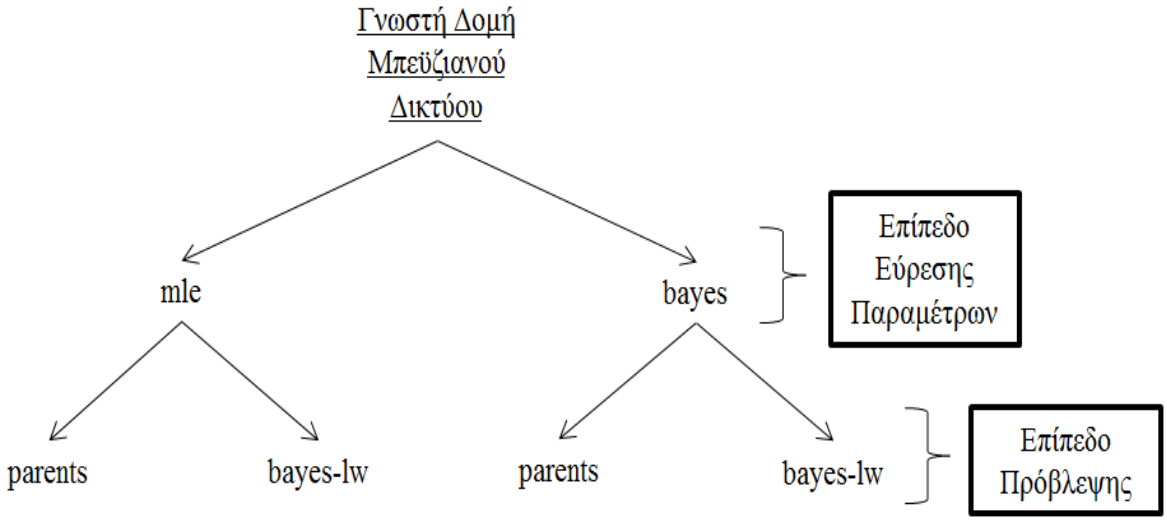
1. Μέθοδοι κατασκευής δικτύου
    1.1. Constrained based μέθοδοι
        1.1.1. gs → Grow Shrink
        1.1.2. iamb → IAMB
    1.2. Score based μέθοδοι
        1.2.1. Μέθοδοι εύρεσης υποψήφιων μπεϋζιανών δικτύων
            1.2.1.1. hc → Hill climbing
            1.2.1.2. tabu → Tabu search

            } Greedy search αλγόριθμοι με operators

        1.2.2. Συναρτήσεις βαθμολόγησης των δικτύων
            1.2.2.1. aic → AIC (Akaike Information Criterion)
            1.2.2.2. bic → BIC (Bayesian Information Criterion)
2. Μέθοδοι εύρεσης παραμέτρων
    2.1. mle → μέθοδος εκτίμησης μέγιστης πιθανοφάνειας
    2.2. bayes → μέθοδος μπεϋζιανής εύρεσης παραμέτρων

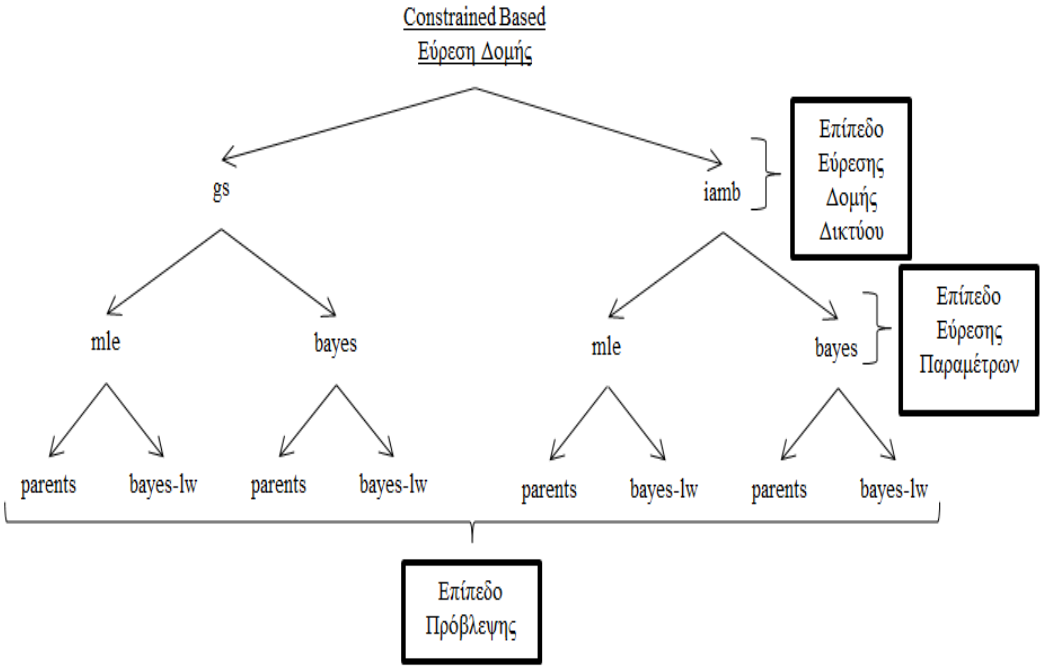
3. Μέθοδοι εκτίμησης νέων δεδομένων
    3.1. parents → εύρεση της εκτίμησης από την τοπική κατανομή
    3.2. bayes-lw → εύρεση της εκτίμησης με τη μέθοδο likelihood weighting

Ακολούθως θα παρουσιαστούν πιο αναλυτικά οι επιμέρους στρατηγικές που ακολουθήθηκαν. Στη πρώτη περιπτωση, έγινε χρήση του έτοιμου δικτύου με την εύρεση των παραμέτρων να γίνεται με δύο διαφορετικούς τρόπους. Με τη μέθοδο mlekαι με τη μέθοδο bayes. Η πρόβλεψη για τη μεταβλητή "NM hazard" έγινε επίσης με δύο τρόπους: με την τοπική προσέγγιση parents και με bayes-lw. Παρακάτω παρουσιάζεται διαγραμματικά όλη η διαδικασία στην Εικόνα 26.



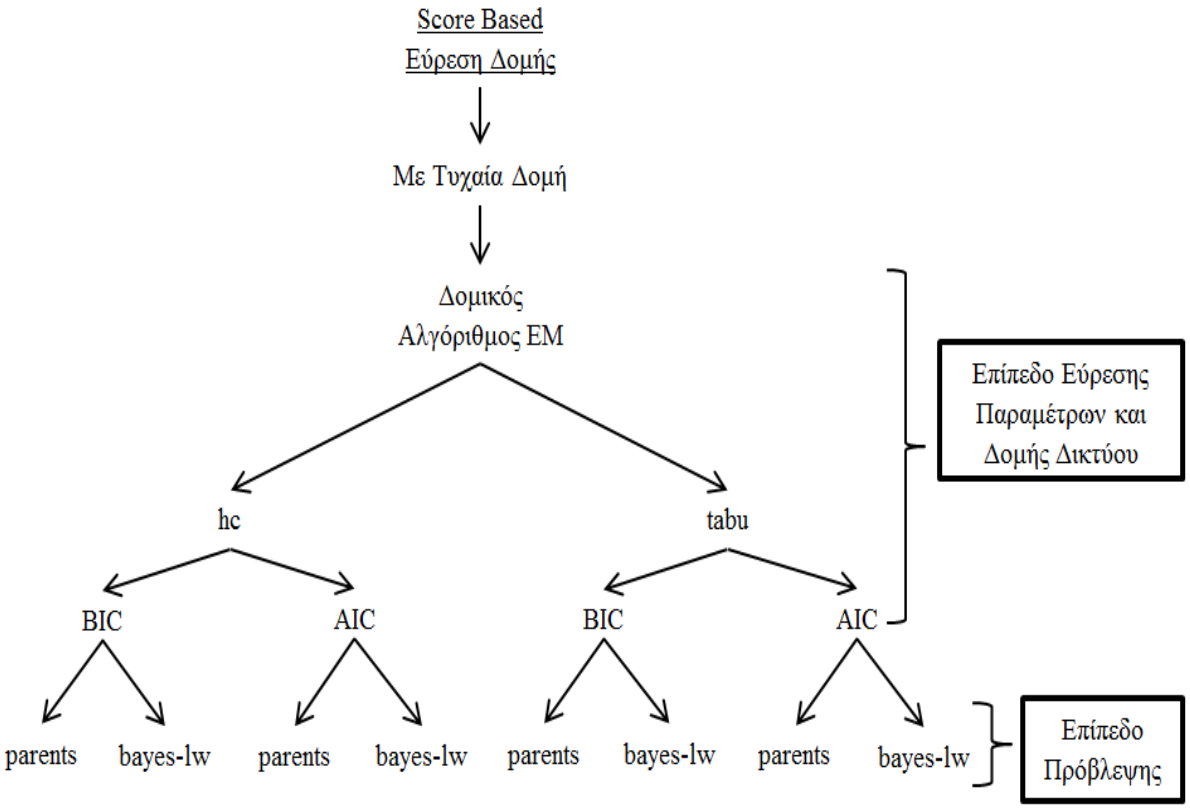

**Εικόνα 26. Στρατηγική εκτίμησης δεδομένων με βάση το γνωστό δίκτυο των Marvin *et al*. (2017).**

Στη δεύτερη περίπτωση έγινε εύρεση της δομής του δικτύου μέσα από την πληροφορία που παρείχαν τα δεδομένα, χρησιμοποιόντας τους costrained based αλγόριθμους gs και iamb. Η εύρεση παραμέτρων εγινε με τις μεθόδους mle και bayes. Η πρόβλεψη εγινε με την parents και την bayes-lw. Το διάγραμμα της Εικόνας 27 περιγράφει τη διαδικασία που ακολουθήθηκε.



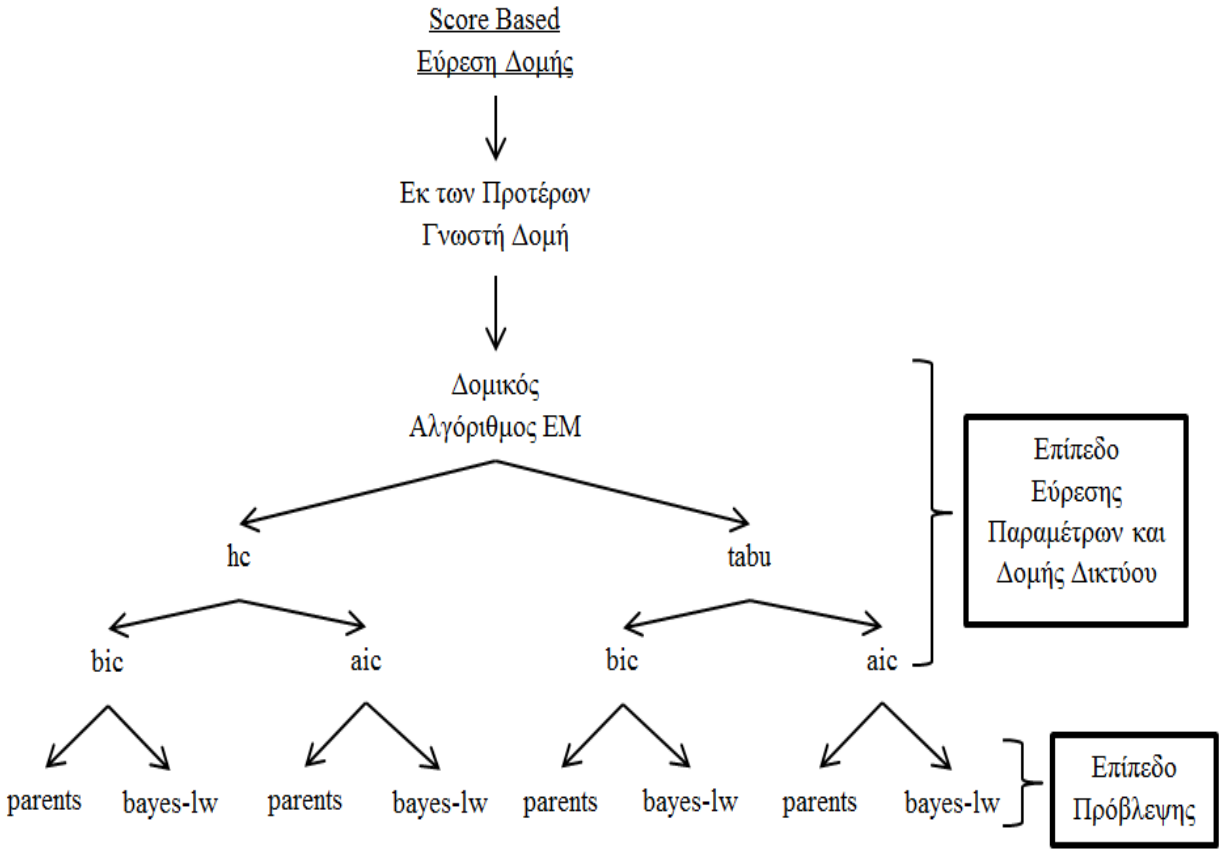

**Εικόνα 27. Στρατηγική κατασκευής δικτύου από τα δεδομένα με constrained based αλγόριθμους.**

Ακολούθως, στην τρίτη περίπτωση, λόγω του γεγονότος ότι τα δεδομένα δεν είναι πλήρη χρησιμοποιήθηκε ο δομικός αλγόριθμος EM, ο οποίος, όπως έχει αναφερθεί αναλυτικά στο Κεφάλαιο 4, υλοποιεί συμπλήρωση των ελλειπών δεδομένων (data imputation). Η κατασκευή της δομής έγινε με score based μεθόδους και συγκεκριμένα χρησιμοποήθηκαν οι greedy search αλγόριθμοι hc και tabu για την κατασκευή των υποψήφιων μπεϋζιανών δικτύων. Η εκκίνηση της εύρεσης της δομής του δικτύου από το σύνολο των δεδομένων έγινε με βάση ένα τυχαίο δίκτυο. Η επιλογή του κατάλληλου δικτύου πραγματοποιήθηκε μέσω των συναρτήσεων βαθμολόγησης bic και aic. Τέλος, η εκτίμηση των νέων στιγμιοτύπων πραγματοποιήθηκε με τις μεθόδους parents και bayes-lw. Η στρατηγική παρουσιάζεται διαγραμματικά στην Εικόνα 28.




**Εικόνα 28. Στρατηγική κατασκευής δικτύου από τα δεδομένα με score based αλγόριθμους και τυχαίο αρχικό δίκτυο.**

Στην τελευτάια περίπτωση χρησιμοποιήθηκαν πάλι score based μέθοδοι εύρεσης δομής μέσω του δομικου αλγόριθμου ΕΜ. Συγκερκιμένα, η εύρεση των υποψήφιων δομών έγινε από τις μεθόδους hc και tabu, αλλά αυτή τη φορά ξεκίνησαν τη διαδικασία έχοντας ως βάση τη γνωστή δομή του δικτύου όπως αυτή είχε διαμορφωθεί από τους Marvin *et al.* (2017) (Εικόνα 24). Η βαθμολόγηση των υποψήφιων δικτύων έγινε με τις bic και aic. Για την εύρεση των παραμέτρων δοκιμάστηκαν πάλι η μέθοδος mle και η bayes, ενώ, η εκτίμηση των νέων στιγμιοτύπων έγινε με την parents και την bayes-lw. Η στρατηγική παρουσιάζεται στην Εικόνα 29.




**Εικόνα 29. Στρατηγική κατασκευής δικτύου από τα δεδομένα με score based αλγόριθμους και εκκίνηση με το δίκτυο των Marvin *et al*. (2017).**

### *5.2.3 Αξιολόγηση Δικτύων*

Για την αξιολόγηση των δικτυων που κατασκευάστηκαν αλλά και για την επιλογή της καλύτερης στρατηγικής, χρησιμοποιήθηκαν οι στατιστικοί δείκτες της ακρίβειας (accuracy) και ο συντελεστής συσχέτισης του Matthews (Matthews Correlation Coefficient-MCC), ενώ για την περαιτέρω εξαγωγή συμπερασμάτων για το κάθε μοντέλο έγινε χρήση του πίνακα σύγχυσης (confusion matrix). Τέλος, για να αποφευχθεί όσο είναι δυνατόν η μη ισορροπημένη κατανομή των δεδομένων μεταξύ του training και του test set, χρησιμοποιήθηκε η μέθοδος του cross validation, και πιο συγκεκριμένα, το 10-fold cross validation, όπου κάθε φορά το ένα δέκατο των δεδομένων αποτελεί το test set και τα υπόλοιπα εννέα δέκατα το training set. Στα υποκεφάλαια που ακολουθούν γίνεται ανάλυση των προαναφερθέντων στατιστικών μέτρων και μεθοδολογίων.

#### *5.2.3.1 Πίνακας Σύγχυσης (Confusion Matrix)*

Ο πίνακας σύγχυσης είναι ένας πίνακας nxn διαστάσεων που περιέχει δεδομένα/παρατηρήσεις και βοηθάει στην αξιολόγηση της συμπεριφοράς ενός αλγόριθμου/μοντέλου μηχανικής μάθησης, που στόχο έχει να εκτιμήσει μια μεταβλητή. Ο πίνακας δείχνει με ποιό τρόπο ένας αλγόριθμος «συγχύζεται» όταν κάνει εκτιμήσεις, δηλαδή δίνει την γενική εικόνα των λαθών που γίνονται στις εκτιμήσεις αλλά και του είδους των λαθών.

Γενικά, τα δεδομένα του πίνακα χωρίζονται σε δύο κατηγορίες, που είναι τα πραγματικά δεδομένα (actual) και τα εκτιμώμενα δεδομένα (predicted). Στην περίπτωση ενός 2x2 confusion matrix, τα πραγματικά δεδομένα χωρίζονται σε δύο κλάσεις. Η πρώτη είναι η κλάση με τα αληθή δεδομένα (True) και η δεύτερη είναι κλάση με τα ψευδή δεδομένα (False). Από τις δύο κλάσεις διακρίνεται συνήθως μια κλάση αναφοράς, δηλαδή η κλάση που στην πραγματικότητα μας ενδιαφέρει περισσότερο να μάθουμε τι έκανε ο αλγόριθμος. Στη κατηγορία των predicted δεδομένων οι εκτιμήσεις διακρίνονται σε θετικές (positive) και αρνητικές (negative). Παρακάτω (Πίνακας 23) φαίνεται η γενική μορφή ενός πίνακα σύγχυσης για 2 κλάσεις.

**Πίνακας 23. Παράδειγμα πίνακα σύγχυσης για πρόβλημα κατηγοριοποίησης δύο κλάσεων(binary classification).**

| | | **Actual** | |
|---|---|---|---|
| **Predicted** | | True | False |
| | Positive | TP | FP |
| | Negative | FN | TN |

Η ορολογία που ακολουθείται είναι η εξής:

*Positive (P)* = Θετική→ Η εκτίμηση του αλγόριθμου είναι ότι το δεδομένο είναι αληθές.

*Negative (N)* = Αρνητική→ Η εκτίμηση του αλγόριθμου είναι ότι το δεδομένο θα είναι ψευδές.

*True Positive (TP)* = Αληθής Θετική→ Η εκτίμηση του αλγόριθμου είναι ότι το δεδομένο είναι αληθές και πράγματι είναι.

*True Negative (TN)* = Αληθής Αρνητική→ Η εκτίμηση του αλγόριθμου είναι ότι το δεδομένο είναι ψευδές και πράγματι είναι.

*False Positive (FP)* = Ψευδής Θετική→ Η εκτίμηση του αλγόριθμου είναι ότι το δεδομένο είναι αληθές και δεν είναι, δηλαδή στη πραγματικότητα είναι ψευδές.

*False Negative (FN)* = Ψευδής Αρνητική→ Η εκτίμηση του αλγόριθμου είναι ότι το δεδομένο είναι ψευδές και δεν είναι, δηλαδή στη πραγματικότητα είναι αληθές.

*TP+FP* = Το σύνολο των δεδομένων που ο αλγόριθμος εκτιμά οτι είναι αληθή.

*FN+TN*= Το σύνολο των δεδομένων που ο αλγόριθμος εκτιμά οτι δεν είναι αληθή, δηλαδή είναι ψευδή.

*TP+FN*= Το σύνολο των δεδομένων που στην πραγματικότητα είναι αληθή.

*FP+TN*= Το σύνολο των δεδομένων που στην πραγματικότητα είναι ψευδή.

Ο πίνακας σύγχυσης γενικεύεται και για πρόβλημα κατηγοριοποίησης με περισσότερες από δύο κλάσεις (multiclass classification) εάν θεωρήσουμε την κάθε κλάση μεμονωμένα ως True και positive και όλες τις άλλες ως False και negative, και αυτό το κάνουμε για κάθε κλάση ξεχωριστά.

Τέλος, αξίζει να αναφερθεί ότι τα FP και FN είναι στενά συνδεδεμένα με την έννοια της μηδενικής υπόθεσης και με των τύπων σφαλμάτων Ί & II. Η μηδενική υπόθεση, που συμβολίζεται με $H_0$, είναι μια υπόθεση που με βάσει τις πληροφορίες και τα δεδομένα που υπάρχουν ελέγχεται για το αν θα γίνει αποδεκτή ή θα απορριφθεί. Η εναλλακτική υπόθεση $H_1$, είναι η υπόθεση που γίνεται αποδεκτή στην περίπτωση που απορριφθεί η μηδενική υπόθεση. Αν η μηδενική υπόθεση γίνει αποδεκτή δεν σημαίνει απαραίτητα ότι στην πραγματικότητα ισχύει, αλλά ότι τα δεδομένα δεν ήταν αρκετά ώστε να απορριφθεί. Αντίθετα αν απορριφθεί, τότε κρίθηκε ότι τα δεδομένα που ήταν διαθέσιμα ήταν ικανά να την απορρίψουν.Ο τύπος σφάλματος Ι είναι η περίπτωση που έχει απορριφθεί η μηδενική υπόθεση ενώ στην πραγματικότητα ισχύει. Για αυτό το λόγο οι περιπτώσεις που αφορούν τα False Positive εμπίπτουν στα σφάλματα τύπου Ι. Ο τύπος σφάλματος ΙΙ είναι η περίπτωση που έχει γίνει αποδεκτή η μηδενική υπόθεση ενώ στην πραγματικότητα δεν ισχύει. Για αυτό το λόγο οι περιπτώσεις που αφορούν τα False Negatives εμπίπτουν στα σφάλματα τύπου 2.

#### *5.2.3.2 Στατιστικοί Δείκτες*

Εφόσον ορίστηκε ο πίνακας σύγχυσης, στη συνέχεια χρησιμοποιούνται οι ορισμοί της προηγούμενης υποενότητας για να οριστούν οι σημαντικότεροι στατιστικοί δείκτες που χρησιμοποιήθηκαν τόσο για την επιλογή της καλύτερης στρατηγικής, όσο και για την εξαγωγή συμπερασματολογίας για το τελικό μοντέλο.

*Sensitivity ή Recall ή True Positive Rate*

$$Sensitivity = \frac{TP}{TP + FN}$$

Από το σύνολο των παρατηρήσεων που στην πραγματικότητα είναι αληθείς (TP+FN), το sensitivity αποκαλύπτει πόσο συχνά ο αλγόριθμος εκτιμά σωστά οτι είναι αληθείς (TP). Με άλλα λόγια, είναι ο ρυθμός με τον οποίο ο αλγόριθμος εκτιμά οτι η παρατήρηση θα είναι αληθής (θετική/positive) σε σχέση με αυτές που στην πραγματικότητα είναι αληθείς (True Positive Rate). Συγκεκριμένα, ο δείκτης Sensitivity βοηθάει όταν το ενδιαφέρον μας επικεντρώνεται στα False Negative. Αύξηση των FN προκαλεί μείωση του sensitivity και μείωση των FN προκαλεί αύξηση του sensitivity. Δηλαδή, μικρό sensitivity σημαίνει οτι υπάρχουν πολλά δεδομένα που εκτιμήσε ο αλγόριθμος οτι ειναι ψευδή και στη πραγματικότητα ήταν αληθή. Στην ουσία επικεντρώνεται στην εκτίμηση σφαλμάτων τύπου 2 (Type 2 Error). Συνεπώς, αν το ενδιαφέρον μας ήταν η πρόβλεψη των False Negative, τότε θα επιλέγαμε τον αλγόριθμο/μοντέλο που θα έδινε το μεγαλύτερο Sensitivity.

*Precision*

$$Precision = \frac{TP}{TP + FP}$$

Από το σύνολο των εκτιμήσεων που υποδεικνύουν οτι τα δεδομένα είναι αληθή (TP+FP), το precision αποκαλύπτει πόσο συχνά ο αλγόριθμος εκτιμά σωστά οτι είναι αληθή (TP). Συγκεκριμένα, ο δείκτης precision βοηθάει όταν το ενδιαφέρον μας επικεντρώνεται στα False Positive. Αύξηση των FP προκαλεί μείωση του Precision και μείωση των FP προκαλεί αύξηση του precision. Δηλαδή, μικρό precision σημαίνει οτι υπάρχουν πολλά δεδομένα που εκτιμήσε ο αλγόριθμος οτι ειναι αληθή και στη πραγματικότητα ήταν ψευδή. Στην ουσία επικεντρώνεται στην εκτίμηση σφαλμάτων τύπου I. Συνεπώς, αν το ενδιαφέρον μας ήταν πρόβλεψη των False Positive τότε θα επιλέγαμε τον αλγόριθμο/μοντέλο που θα έδινε το μεγαλύτερο Precision.

*Accuracy*

$$Accuracy = \frac{TP + TN}{TOTAL} = \frac{TP + TN}{TP + TN + FP + FN}$$

Από το σύνολο όλων των δεδομένων, το accuracy δείχνει πόσο συχνά ο αλγόριθμος εκτιμά σωστά οτι τα δεδομένα είναι αληθή ή είναι ψευδή. Ο δείκτης accuracy βοηθάει περισσότερο όταν μας ενδιαφέρουν παραπάνω οι σωστές εκτιμήσεις δηλαδή TP+TN και όχι τόσο οι λάθος FP+FN. Ωστόσο, παρουσιάζει αδυναμία όταν τα δεδομένα δεν είναι συμμετρικά κατανεμημένα, δηλαδή όταν το πλήθος των δεδομένων των κλάσεων True και False παρουσιάζει μεγάλη διαφορά και τελικά μας ενδιαφέρει να μάθουμε για τις λάθος εκτιμήσεις του αλγόριθμου, τις FP και FN. Συνεπώς, σε αυτή την περίπτωση το accuracy δεν αποτελεί την καλύτερη επιλογή δείκτη για να συγκρίνουμε αλγόριθμους. Σημειώνεται ότι τύπος που δίνεται παραπάνω ισχύει για πρόβλημα κατηγοριοποίησης δύο κλάσεων. Η γενίκευσή του είναι το άθροισμα της διαγωνίου του πίνακα σύγχυσης προς το άθροισμα όλων των στοιχείων .

*Συντελεστής Συσχέτισης του Matthew - Matthews Correlation Coefficient (MCC)*

$$MCC = \frac{TP * TN - FP * FN}{\sqrt{(TP + FP) * (TP + FN) * (TN + FP) * (TN + FN)}}$$

Ο συντελεστής συσχέτισης MCC, σε αντίθεση με τους προηγούμενους δείκτες, θεωρεί ως κλάσεις τα πραγματικά δεδομένα (Actual) και τις εκτιμήσεις για τα δεδομένα (Predicted). Οι τιμές που μπορεί να λάβει ο MCC κυμαίνονται από το -1 εώς το 1. Συγκεκριμένα, για $MCC = 1$ σημαίνει οτι υπάρχει ταύτιση μεταξύ των εκτιμήσεων του αλγόριθμου και των πραγματικών δεδομένων, για $MCC = 0$ σημαίνει οτι οι εκτιμήσεις είναι τυχαίες σε σχέση με τα πραγματικά δεδομένα και για $MCC = -1$ σημαίνει οτι οι εκτιμήσεις βρίσκονται σε πλήρη αντίθεση με τα πραγματικά δεδομένα. Τα θετικά στοιχεία του MCC σαν δείκτη για την αξιολόγηση της αξιοπιστίας αλγόριθμων είναι ότι λαμβάνει υπόψη του τα TN, δεν μεταβάλλεται σε περίπτωση ανταλλαγής των κλάσεων και δίνει μια συσχέτιση μεταξύ των εκτιμήσεων και των πραγματικών δεδομένων.

#### *5.2.3.3 Cross Validation*

Για την διαδικασία εκπαίδευσης επιλέχθηκε η μέθοδος του cross validation σε αντίθεση με τον τυχαίο διαχωρισμό των δεδομένων σε ένα training set και σε ένα test set. Γενικά, όσα περισσότερα δεδομένα υπάρχουν για να εκπαιδευτεί ο αλγόριθμος, τόσο καλύτερες θα είναι οι εκτιμήσεις που θα δίνει στις κλάσεις ενδιαφέροντος.

Ωστόσο, δεν είναι παντα εφικτή η ύπαρξη πολλών δεδομένων και πολλές φορές θα πρέπει να χωριστούν τα υπάρχοντα δεδομένα σε training set και σε test set, όπώς έγινε και στην περίπτωση που τίθεται προς μελέτη. Συνήθως, από το σύνολο των δεδομένων το ένα τρίτο επιλέγεται ως test set και τα υπόλοιπα δύο τρίτα επιλέγονται ως training set. Κάτι τέτοιο όμως μπορεί να οδηγήσει σε χαμηλή ικανότητα εκτίμησης του αλγόριθμου. Αυτό μπορεί να συμβεί όταν η επιλογή του training set δεν περιέχει πολλά δεδομένα που να αφορούν μια συγκεκριμένη κλάση και συγχρόνως αυτά βρίσκονται στο test set που επιλέχθηκε. Τότε είναι αναμενόμενο ότι αλγόριθμος δεν θα εκτιμά σωστά τη συγκεκριμένη κλάση, διότι το training set δεν θα είναι αντιπροσωπευτικό για όλες τις κλάσεις. Το πρόβλημα αυτό μπορεί να αποφευχθεί με την τεχνική του cross validation.

Το cross validation είναι μια μέθοδος η οποία στόχο έχει τη αποφυγή του overfitting κατά τη διαδικασία της εκπαίδευσης του αλγόριθμου και είναι ιδιαιτέρως χρήσιμη ειδικά στην περιπτωση που υπάρχουν λίγα δεδομένα. Ένα από τα είδη της μεθόδου είναι η k-fold cross validation. Στη περίπτωση αυτή το σύνολο των δεδομένων χωρίζεται σε k το πλήθος υποσύνολα δεδομένων, όπου τα k-1 το πλήθος από αυτά χρησιμοποιούνται ως training set για την εκαπίδευση του αλγόριθμου και το ένα που μένει αφήνεται ως test set. Η διαδικασία αυτή επαναλαμβάνεται μέχρι όλα τα k υποσύνολα των δεδομένων να χρησιμοποιηθούν μια φορά ως test set. Τότε, η εκτίμηση του αλγόριθμου θα είναι ο μέσος όρος των k το πλήθος εκτιμήσεων που παράχθηκαν από τη διαδικασία. Έχει βρεθεί πειραματικά ότι ο διαμερισμός των δεδομένων σε 10 το πλήθος υποσύνολα συνήθως δίνει καλές εκτιμήσεις χωρίς να σημαίνει απαραίτητα ότι ο διαμερισμός σε λίγο λιγότερα ή περισσότερα υποσύνολα δεν θα έχει καλές εκτιμήσεις.

Μια ακόμα περίπτωση αποτελεί η μέθοδος leave one out cross validation κατά την οποία αφήνεται ένα στιγμιότυπο εκτός και τα υπόλοιπα αποτελούν το training set. Άρα αν τα στιγμιότυπα των δεδομένων είναι n το πλήθος η διαδικασία επαναλαμβάνεται μέχρι όλα τα στιγμιότυπα δεδομένα να χρησιμοποιηθούν μια φορά ως test set (Witten, *et al.*, 2011).

Φυσικό επακόλουθο της χρήσης 10-fold cross validation στην παρούσα εργασία είναι να γίνεται αναφορά των μέσων όρων των στατιστικών δεικτών μεταξύ της κάθε επανάληψης. Επίσης, ο τελικός πίνακας σύγχυσης αποτελεί το μέσο όρο των πινάκων σύγχυσης των 10 επαναλήψεων.

## 5.3 Αποτελέσματα

### *5.3.1 Επιλογή Βέλτιστης Στρατηγικής Εκπαίδευσης*

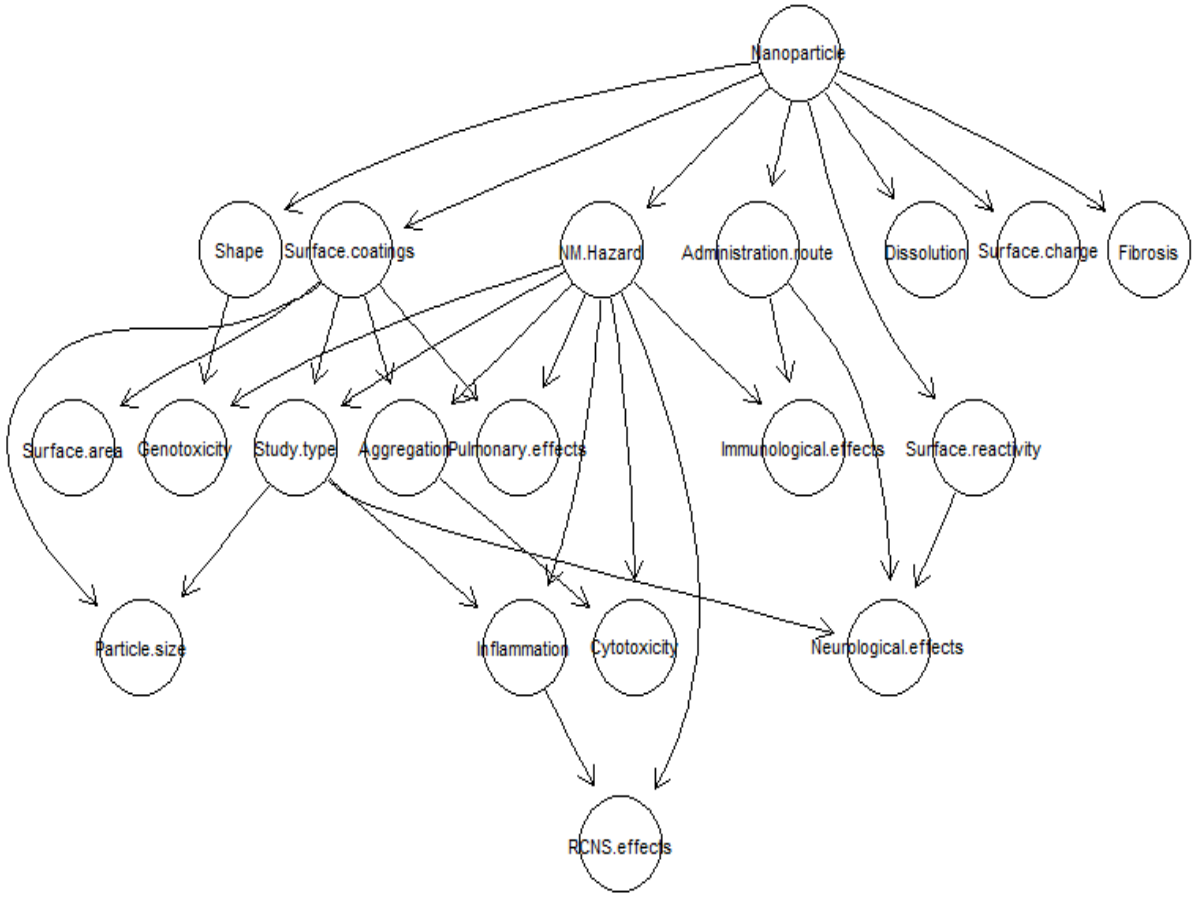
Όπως έχει ήδη αναφερθεί, χρησιμοποιήθηκε η μέθοδος του cross validation και συγκεκριμένα η 10-fold cross validation. Για κάθε μία από τις 10 επαναλήψεις υπολογίστηκαν οι δείκτες accuracy, MCC καθώς και οι αντίστοιχοι πίνακες σύγχυσης. Σε κάθε cross validation, για κάθε υποπερίπτωση στρατηγικής, υπολογίστηκε ο μέσος όρος του Accuracy, του MCC καθώς και ο μέσος όρος των πινάκων σύγχυσης. Στα αποτελέσματα παρουσιάζονται ακριβώς αυτοί οι μέσοι. Τα αναλυτικά αποτελέσματα των τεσσάρων περιπτώσεων στρατηγικής που αφορούν στα μπεϋζιανά δίκτυα βρίσκονται στο Παράρτημα, εκτός αυτών που κρίθηκε σκόπιμο να παρουσιαστούν σε αυτό το υποκεφάλαιο.

Όσον αφορά το δείκτη MCC, σε αρκετές από τις 10 επαναλήψεις του cross validation, για κάθε υποπερίπτωση στρατηγικής, εμφανίστηκαν τιμές τύπου NaN (Not a Number) που σημαίνει ότι υπήρξε απροσδιοριστία του τύπου 0/0. Αυτό είχε ως αποτέλεσμα ο μέσος όρος του MCC να παίρνει την τιμή NaN σε πολλές υποπεριπτώσεις. Για αυτό το λόγο δόθηκε περισσότερη βαρύτητα στο δείκτη Accuracy και ο δείκτης MCC χρησιμοποιήθηκε επικουρικά σε περιπτώσεις όπου ο μέσος όρος δεν είχε την τιμή NaN.

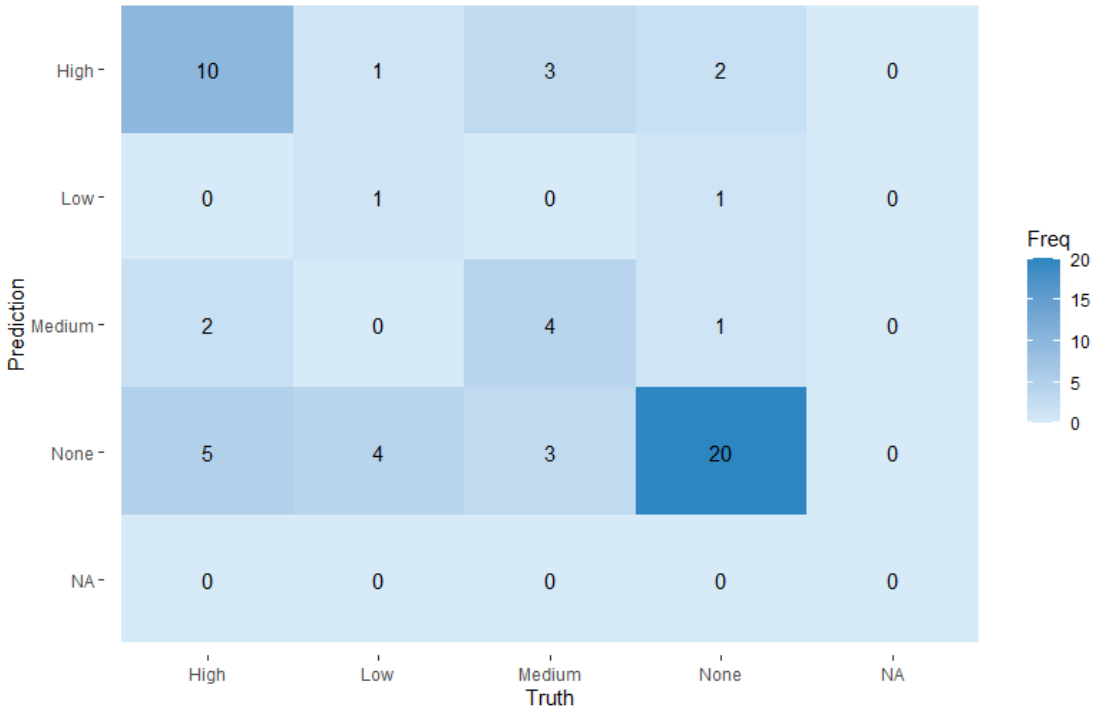
Γενικά σε όλες τις υποπεριπτώσεις των τεσσάρων περιπτώσεων στρατηγικής όπου χρησιμοποιήθηκε η bayes-lw για την εκτίμηση νέων δεδομένων, έγινε ρύθμιση της υπερπαραμέτρου $n$ που καθορίζει το πλήθος των παραγόμενων στιγμιοτύπων με βάρος που θα χρησιμοποιηθούν για την εκτίμηση. Η βέλτιστη τιμή της υπερπαραμέτρου ήταν $n = 40000$ δείγματα. Για μεγαλύτερο αριθμό δεν αύξανε επαρκώς το accuracy, αντίθετα με το υπολογιστικό κόστος, ενώ για μικρότερο $n$ μειωνόταν το accuracy. Συνεπώς η καλύτερη ανταλλαγή (trade-off) μεταξύ accuracy και υπολογιστικού κόστους γινόταν για $n = 40000$ στιγμιότυπα.

Τα καλύτερα αποτελέσματα με βάση το accuracy συνεπικουρούμενο και από τον δείκτη MCC βρέθηκαν στην $4^{η}$ περίπτωση που αφορά την κατασκευή του δικτύου από τα δεδομένα με τη χρήση score based μεθόδων εύρεσης δικτύου. Συγκεκριμένα, πρόκειται για την περίπτωση που χρησιμοποιήθηκε ο δομικός αλγόριθμος ΕΜ με τον

score based αλγόριθμο tabu και η βαθμολόγηση των υποψήφιων δικτύων έγινε με τη συνάρτηση βαθμολόγησης aic. Το αρχικό δίκτυο που χρησιμοποιήθηκε για την έναρξη της διαδικασίας είναι το μπεϋζιανό δίκτυο των Marvin *et al*. (2017) (Εικόνα 24) και η εκτίμηση των νέων δεδομένων του test set έγινε με τη μέθοδο bayes-lw. Το μπεϋζιανό δίκτυο στο οποίο κατέληξε η διαδικασία παρουσιάζεται στην Εικόνα 30 και ο προκύπτων πίνακας σύγχυσης παρουσιάζεται στην Εικόνα 31.




**Εικόνα 30. Το μπεϋζιανό δίκτυο της μεθόδου με τις καλύτερες εκτιμήσεις δεδομένων.**




**Εικόνα 31. Ο πίνακας σύγχυσης της μεθόδου με τις καλύτερες εκτιμήσεις δεδομένων.**

Επιπλέον οι στατιστικοί δείκτες των Accuracy και MCC είχαν τα εξής αποτελέσματα:

$$Accuracy = 0.612$$

$$MCC = 0.366$$

Τα συνολικά αποτελέσματα που αφορούν στους δείκτες Accuracy και MCC παρουσιάζονται συγκντρωτικά σε πίνακες για κάθε υποπερίπτωση των τεσσάρων περιπτώσεων στρατηγικής (Πίνακες 24-27).

**Πίνακας 24. Αποτελέσματα των accuracy και MCC για την 1η περίπτωση.**

| $1^{η}$ περίπτωση στρατηγικής | | | |
|---|---|---|---|
| Έυρεση Παραμέτρων | Εκτίμηση Δεδομένων | Accuracy | MCC |
| bayes | bayes-lw | 0.591 | 0.394 |
| bayes | parents | 0.419 | NaN |
| mle | bayes-lw | 0.264 | 0.141 |
| mle | parents | 0.419 | NaN |

Όπως φαίνεται και στον πίνακα 24, τα καλύτερα αποτελέσματα στην 1η περίπτωση βρέθηκαν όταν χρησιμοποιήθηκε η bayes για την εύρεση των παραμέτρων και η εκτίμηση έγινε με την bayes-lw. Το Accuracy ήταν ίσο με 0.591 και το MCC ήταν ίσο με 0.394. Το μπεϋζιανό δίκτυο των Marvin *et al*. (2017) και ο αντίστοιχος μέσος όρος των πινάκων σύγχυσης βρίσκονται στις Εικόνες 1 και 2 του Παραρτήματος αντίστοιχα.

Στο σημείο αυτό θα πρέπει να σημειωθεί ότι στην υποπερίπτωση που χρησιμοποιήθηκε η mle και η bayes-lw εμφανίστηκαν πολλές τιμές NA (Not Available), όπως φαίνεται για παράδειγμα στην Εικόνα 4 του Παραρτήματος. Γενικά αυτό το πρόβλημα παρουσιάστηκε και σε άλλες περιπτώσεις στρατηγικής που η εύρεση των παραμέτρων γινόταν με την μέθοδο mle Πιο συγκεκιμένα, η διαδικασία εύρεσης των παραμέτρων με τη μέθοδο της μέγιστης πιθανοφάνειας (mle), λόγω του μη πλήρους συνόλου δεδομένων, δεν μπορούσε πάντα να εκτιμήσει τις παραμέτρους $\theta$ του μοντέλου ώστε να συμπληρωθούν παντού οι τιμές στα CPTs των μεταβλητών. Έτσι, η εντολή predict (με όρισμα parents ή bayes-lw) μετέφερε τα NAs που παρουσιάζονταν στις τελικές εκτιμήσεις της μεταβλητής NM hazard. Οι εκτιμήσεις με τα NAs με τη σειρά τους εμπεριέχονταν στον διάνυσμα των προβλέψεων το οποίο φτιάχνεται για να αποθηκεύει τις εκτιμήσεις σε κάθε επανάληψη του cross validation. Για να υπολογιστεί το Accuracy για κάθε μια επανάληψη του cross validation έγινε σύγκριση του διανύσματος των παρατηρήσεων που περιείχε τα πραγματικά δεδομένα και του διανύσματος των προβλέψεων που περιείχε τις εκτιμήσεις του μοντέλου. Συνεπώς, λόγω των NAs που εμφανίζονταν σε κάποια σημεία στο διάνυσμα των προβλέψεων, εμφανίζονταν NAs στα αντίστοιχα σημεία του διανύσματος του accuracy. Για αυτό το λόγο προστέθηκε στα επίπεδα των διανυσμάτων των παρατηρήσεων και των εκτιμήσεων ένα ακόμα επίπεδο, το NA, όπως φαίνεται και στους παρουσιαζόμενους πίνακες σύγχυσης. Έτσι, κατά τη διαδικασία της πρόβλεψης, όπου εμφανίστηκε η τιμή NA αντικαταστάθηκε με το level NA.

Στη 2η περίπτωση στρατηγικής που αφορά στη κατασκευή του μπεϋζιανού δικτύου από τα δεδομένα με βάση τους constrained based αλγόριθμους gs και iamb, κατοχυρώθηκαν τα αποτελέσματα που παρουσιάζονται στον πίνακα 25.

**Πίνακας 25. Αποτελέσματα των accuracy και MCC για την 2η περίπτωση.**

| 2η περίπτωση στρατηγικής | | | | |
|---|---|---|---|---|
| Constrained based αλγόριθμοι | Έυρεση Παραμέτρων | Εκτίμηση Δεδομένων | Accuracy | MCC |
| gs | bayes | bayes-lw | 0.419 | NaN |
| | bayes | parents | 0.419 | NaN |
| | mle | bayes-lw | 0.360 | NaN |
| | mle | parents | 0.419 | NaN |
| iamb | bayes | bayes-lw | 0.419 | NaN |
| | bayes | parents | 0.375 | NaN |
| | mle | bayes-lw | 0.401 | NaN |
| | mle | parents | 0.375 | NaN |

Σε αυτή την περίπτωση, τόσο ο αλγόριθμος gs όσο και ο iamb επέστρεφαν ένα μερικώς κατευθυνόμενο δίκτυο δηλαδή κάποιες από τις ακμές ανάμεσα στους κόμβους δεν είχαν κατεύθυνση. Συνεπώς, δεν μπορούσαν να διαμορφωθούν τα CPTs των μεταβλητών του δικτύου κατά τη διαδικασία εύρεσης των παραμέτρων. Για αυτό το λόγο χρησιμοποιήθηκε η εντολή cextend της βιβλιοθήκης *Bnlearn*, η οποία πηγαίνει στο γράφημα και δίνει κατεύθυνση στις ακμές που δεν είχαν έτσι ώστε να παραχθεί το γράφημα τύπου DAG του μπεϋζιανού δικτύου. Επιπλέον, όπως φαίνεται στις Εικόνες 6 και 11 του Παραρτήματος, οι constrained based αλγόριθμοι gs και iamb δεν κατάφεραν να βρούν πολλές εξαρτήσεις μεταξύ των μεταβλητών σε αντίθεση με τους score based αλγόριθμους. Τα αποτελέσματα στο accuracy δεν ξεπέρασαν το 0.419 σε οποιαδήποτε υποπερίπτωση.

Η 3η περίπτωση στρατηγικής ήταν εκείνη στην οποία έγινε κατασκευή του μπεϋζιανού δικτύου από τα δεδομένα με χρήση των score based μεθόδων hc και tabu καθώς και χρηση του δομικού αλγόριθμου EM, όπου η εκκίνηση της διαδικασίας έγινε σε κάθε υποπερίπτωση με τυχαία δομή δικτύου. Τα αποτελέσματα για την κάθε υποπερίπτωση παρουσιάζονται στον πίνακα 26.

**Πίνακας 26. Αποτελέσματα των accuracy και MCC για την 3η περίπτωση.**

| 3η περίπτωση στρατηγικής | | | | |
|---|---|---|---|---|
| Score based αλγόριθμοι | Συνάρτηση Βαθμολόγησης | Εκτίμηση Δεδομένων | Accuracy | MCC |
| hc | aic | bayes-lw | 0.564 | NaN |
| | aic | parents | 0.558 | NaN |
| | bic | bayes-lw | 0.392 | NaN |
| | bic | parents | 0.326 | NaN |
| tabu | aic | bayes-lw | 0.555 | NaN |
| | aic | parents | 0.558 | NaN |
| | bic | bayes-lw | 0.385 | NaN |
| | bic | parents | 0.326 | NaN |

Η χρήση της score based μεθόδου hc, με τη συνάρτηση βαθμολόγησης aic και την μέθοδο της bayes-lw είναι η υποπερίπτωση που έδωσε το καλύτερο Accuracy (0.564). Επίσης, μπορεί να παρατηρηθεί ότι η χρήση της συνάρτησης βαθμολόγησης aic επηρέασε ελαφρώς προς το καλύτερο τα αποτελέσματα του Accuracy, κάτι που ενδεχομένως οφείλεται στο γεγονός ότι η aic ποινικοποιεί λιγότερο την πολυπλοκότητα του δικτύου σε σχέση με τη συνάρτηση βαθμολόγησης bic.

Στην 4η περίπτωση παρουσιάστηκε, όπως ήδη έχει αναφερθεί, το μεγαλύτερο accuracy (Πίνακας 27) από όλες τις στρατηγικές που χρησιμοποιήθηκαν. Η διαφορά σε αυτή την στρατηγική σε σχέση με την προηγούμενη είναι ότι η εκκίνηση του δομικού αλγόριθμου ΕΜ έγινε με βάση το μπεϋζιανό δίκτυο στο οποίο κατέληξαν στην έρευνά τους οι Marvin *et al*. (2017).

**Πίνακας 27. Αποτελέσματα των accuracy και MCC για την 4η περίπτωση.**

| 4η περίπτωση στρατηγικής | | | | |
|---|---|---|---|---|
| Score based αλγόριθμοι | Συνάρτηση Βαθμολόγησης | Εκτίμηση Δεδομένων | Accuracy | MCC |
| hc | aic | bayes-lw | 0.585 | 0.333 |
| | aic | parents | 0.419 | NaN |
| | bic | bayes-lw | 0.544 | 0.256 |
| | bic | parents | 0.475 | NaN |
| tabu | aic | bayes-lw | 0.612 | 0.366 |
| | aic | parents | 0.585 | NaN |
| | bic | bayes-lw | 0.546 | 0.264 |
| | bic | parents | 0.496 | NaN |

Συνεπώς, από την 3η και 4η περίπτωση μπορεί να παρατηρηθεί ότι το αρχικό δίκτυο με το οποίο ξεκινάει ο δομικός αλγόριθμος ΕΜ επηρεάζει τα αποτέλεσματα. Αυτό συμβαίνει διότι ο δομικός αλγόριθμος ΕΜ εξασφαλίζει σε κάθε επανάληψη ότι το δίκτυο το οποίο κατασκευάζεται είναι καλύτερο από το προηγούμενο, με βάση κάποια συνάρτηση βαθμολόγησης. Αυτό φυσικά δεν εξασφαλίζει ότι το δίκτυο στο οποίο καταλήγει θα είναι επαρκώς κοντά στη καλυτερη δυνατή περίπτωση δικτύου.

Γενικά, τόσο στη 3η όσο και στη 4η περίπτωση, που αφορούν την κατασκευή του μπεϋζιανού δικτύου από τα δεδομένα με score based αλγόριθμους, παρουσιάστηκε πρόβλημα στην εκτίμηση νέων δεδομένων όταν χρησιμοποιήθηκε η μέθοδος parents. Το πρόβλημα αφορούσε στο γεγονός ότι δεν υπήρχαν οι τιμές των γονιών της μεταβλητής NM hazard. Η συγκεκριμένη μέθοδος χρειάζεται τις τιμές των γονιών της μεταβλητής, γιατί όπως έχει αναφερθεί χρησιμοποιεί την τοπική κατανομή στο CPT. Για το λόγο αυτό δημιουργήθηκε ένας πίνακας με ακμές που δεν πρέπει να συμπεριληφθούν στη διαδικασία μάθησης του δικτύου και δόθηκαν στον αλγόριθμο με τη μορφή black list, δηλαδή απαγορευόταν η επιλογή μίας τέτοιας ακμής. Οι ακμές αυτές περιλαμβάνουν όλες εκείνες τις ακμές που είχαν ως γονέα τις βιολογικές συνέπειες και ως παιδί τον κόμβο NM hazard. Η λογική πίσω από αυτή τη λίστα είναι ότι οι βιολογικές συνέπειες είχαν πάρα πολλά NAs, οπότε οποιαδήποτα σύνδεση μεταξύ των βιολογικών συνεπειών και του NM Hazard με κατεύθυνση από τις πρώτες

στη δεύτερη θα δημιουργούσε πρόβλημα. Στη συνέχεια, επειδή σε κάποια είσοδο μπορεί να έλειπε η τιμή (δηλαδή να εμφανιζόταν η τιμή NA) χρησιμοποιήθηκε μια απλή μέθοδος συμπλήρωσης δεδομένων (data imputation) βάση της οποίας όποτε εντοπιζόταν η τιμή NA γινόταν αντικατάστασή της με εκείνη την τιμή της υπό εξέτασης μεταβλητής εισόδου που εμφάνιζε την υψηλότερη συχνότητα στο αρχικό σύνολο δεδομένων.

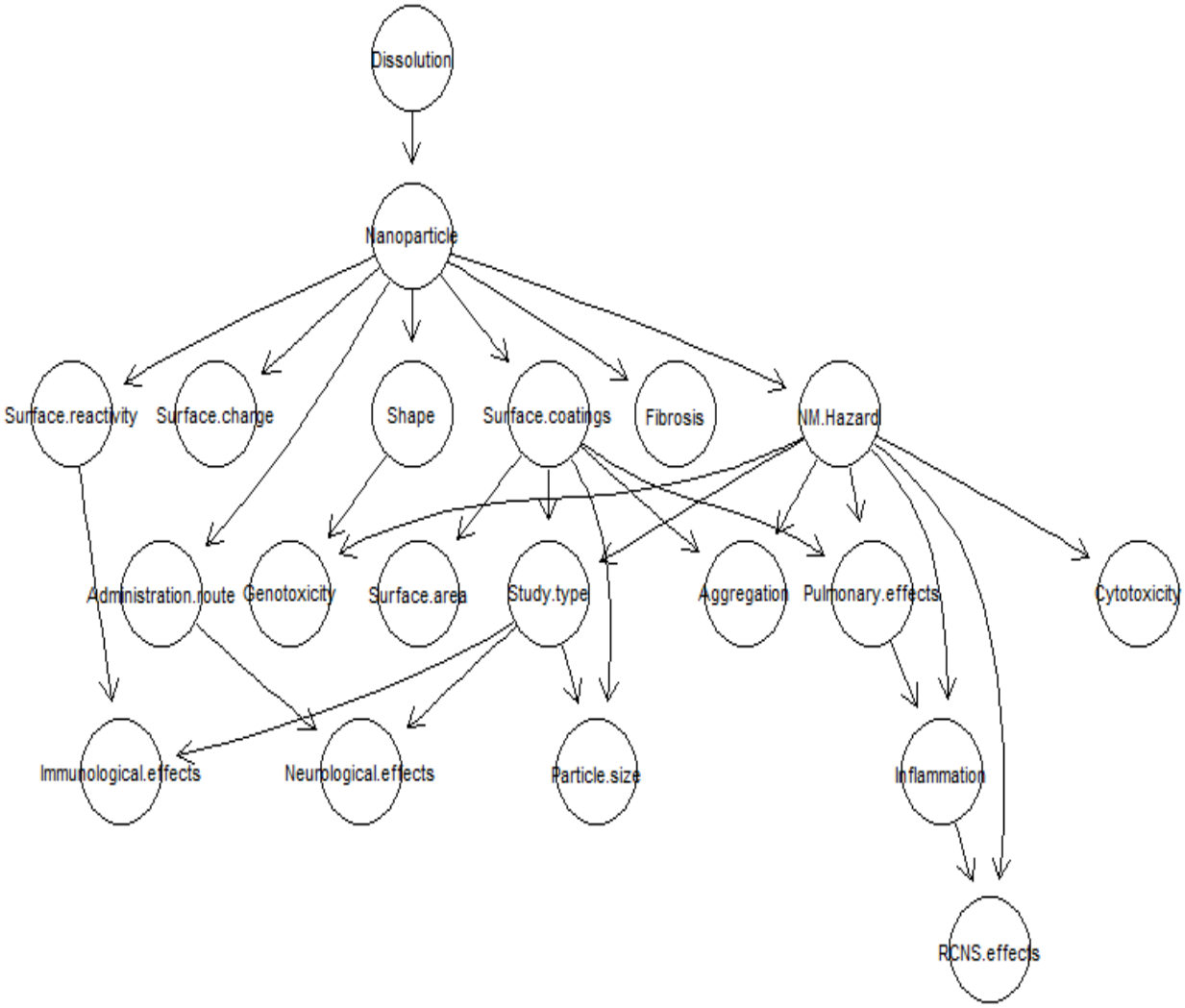
### *5.3.2 Σύγκριση Αποτελεσμάτων με Marvin et al. (2017)*

Στη συνέχεια, επιλέχθηκε η καλύτερη στρατηγική της προηγούμενης ενότητας και έγινε εκπαίδευση του δικτύου από την αρχή, χρησιμοποιώντας ακριβώς τα ίδια training και test set που χρησιμοποίησαν στο μοντέλο τους οι Marvin *et al*. (2017). Ο συνδυασμός που χρησιμοποιήθηκε είναι ο EM με tabu και aic score και πρόβλεψη με bayes-lw, με εκκίνηση από το τελικό δίκτυο που αναφέρουν οι οι Marvin *et al*. (2017) (Εικόνα 24). Τα αποτελέσματα που αφορούν στο accuracy και στο MCC για την εκτίμηση των κλάσεων της μεταβλητής NM hazard είναι τα εξής:

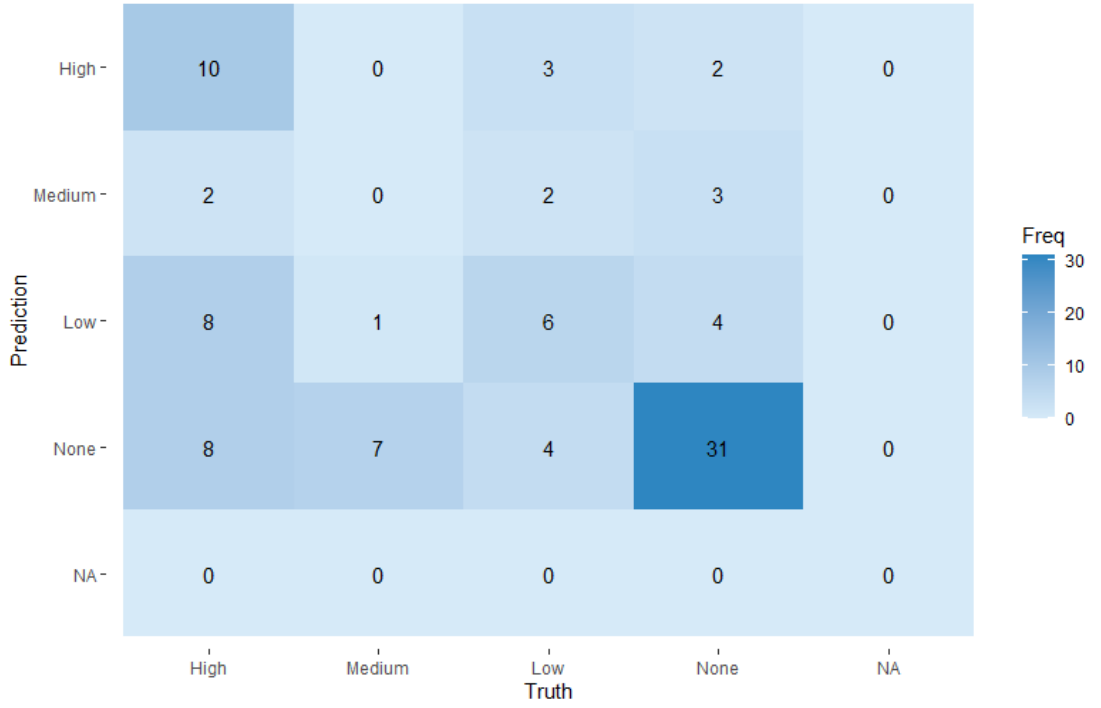
$$Accuracy = 0.516 \text{ ή } 51.6\%$$

$$MCC = 0.282$$

Το τελικό δίκτυο που παρήγαγε ο αλγόριθμος παρουσιάζεται στην Εικόνα , ενώ ο πίνακας σύγχυσης που προέκυψε φαίνεται στην Εικόνα




**Εικόνα 32. Η δομή του μπεϋζιανού δικτύου με εφαρμογή της βέλτιστης στρατηγικής και χρήση ίδιων δεδομένων εκπαίδευσης με τους Marvin et al. (2017).**




**Εικόνα 33: Ο πίνακας σύγχυσης που προέκυψε με εφαρμογή της βέλτιστης στρατηγικής και χρήση ίδιων δεδομένων εκπαίδευσης με τους Marvin et al. (2017).**

Στο σημείο αυτό επαναλαμβάνεται ότι το αντίστοιχο αποτέλεσμα για το accuracy από τους Marvin *et al*. (2017) ήταν 72%. Η μεγάλη αυτή διαφορά οφείλεται στο γεγονός ότι δεν ξέρουμε ποιά ήταν η αρχική δομή με την οποία έτρεξε ο δομικός αλγόριθμος ΕΜ των Marvin *et al*. (2017). Όπως αναφέρθηκε και σε προηγούμενο μέρος του κεφαλαίου, ο δομικός αλγόριθμος ΕΜ εξασφαλίζει την κατασκευή ενός δικτύου με καλύτερη βαθμολογία σε κάθε επανάληψη από το προηγούμενο δίκτυο. Ωστόσο δεν εξασφαλίζει το αν το δίκτυο πλησιάζει κοντά στο βέλτιστο δυνατό, κάτι που σημαίνει ότι μπορεί για τον αριθμό των επαναλήψεων που έγιναν να παράχθηκε ένα δίκτυο με καλύτερη βαθμολόγηση από την αρχική ωστόσο δεν σημαίνει ότι η πορεία ήταν τέτοια ώστε να πλησιάσει το βέλτιστο δυνατό δίκτυο. Ένα ακόμα γεγονός που εξηγεί τη διαφορά στην ακρίβεια είναι ότι ο δομικός αλόγοριθμος ΕΜ χρειάζεται ένα αρχικό σύνολο παραμέτρων $\boldsymbol{\theta}$ για να ξεκινήσει τη διαδικασία, το οποίο αν δεν δωθεί λαμβάνεται με τυχαίο τρόπο. Στη συγκεκριμένη περίπτωση δεν υπήρχε πρότερη γνώση των παραμέτρων και άρα ο αλγόριθμος ξεκίνησε με ένα τυχαίο σύνολο $\boldsymbol{\theta}$, ενώ στην περίπτωση των Marvin *et al.* (2017) μπορεί να δόθηκαν πρότερες κατανομές (δεν διευκρινίζεται μέσα στο κείμενο).

Στη συνέχεια, μπορεί κανείς να παρατηρήσει στον πίνακα σύγχυσης (Εικόνα 33) ότι ο αλγόριθμος δεν παρουσιάζει ιδιαίτερη προβλεπτική ικανότητα για της κλάσεις Low,

Medium και High, σε αντίθεση με την κλάση None, για την οποία παρουσίασε τα περισσότερα True Positives. Συγκεκριμένα, ο αλγόριθμος προέβλεψε για 31 περιπτώσεις ότι δεν θα αποτελούσαν κάποιο κίνδυνο για το άνθρωπο και πράγματι αυτό ήταν αλήθεια. Για την κλάση None έχουμε τα εξής αποτελέσματα:

$$TP_{None} = 31$$

$$FP_{None} = 8 + 7 + 4 = 19$$

$$FN_{None} = 2 + 3 + 4 = 9$$

$$TN_{None} = 10 + 3 + 2 + 2 + 8 + 1 + 6 = 32$$

Όπου πλέον μπορούμε να υπολογίσουμε τους δείκτες Sensitivity και Precision της κλάσης None ως εξής:

$$Sensitivity_{None} = \frac{TP_{None}}{TP_{None} + FN_{None}} = 0.775$$

$$Precision_{None} = \frac{TP_{None}}{TP_{None} + FP_{None}} = 0.62$$

Το γεγονός ότι το $Sensitivity_{None}$ της κλάσης None είναι κοντά στο 78% σημαίνει ότι από το σύνολο των περιπτώσεων που στην πραγματικότητα ήταν None ο αλγόριθμος προέβλεψε ότι θα είναι None το 78% των περιπτώσεων. Αυτό ουσιαστικά μας δείχνει ότι ο αλγόριθμος κατάφερε να περιορίσει τα $FN_{None}$ και άρα να περιορίσει τις περιπτώσεις που προέβλεπε ότι ένα ναναοϋλικό δεν θα είναι ακίνδυνο και τελικά αυτό ήταν ακίνδυνο. Επίσης, το γεγονός ότι το $Precision_{None}$ της κλάσης None είναι κοντά στο 62% σημαίνει ότι από το σύνολο των περιπτώσεων που ο αλγόριθμος εκτίμησε ότι θα είναι None, τελικά το 62% ήταν πράγματι None. Αυτό ουσιαστικά μας δείχνει ότι ο αλγόριθμος περιόρισε σε σχετικά καλό βαθμό τις περιπτώσεις των $FP_{None}$, δηλαδή τις περιπτώσεις όπου προέβλεπε ότι το νανοϋλικο θα ήταν ακίνδυνο και τελικά δεν ήταν. Αυτού του είδους η λάθος πρόβλεψη είναι σημαντική διότι αν για παράδειγμα ο σκοπός ήταν η παραγωγή νανοϋλικών για την αγορά, το 38% των περιπτώσεων θα αποτελούσαν νανοϋλικά που προβλέφθηκαν ότι δεν είναι επικίνδυνα για την ανθρώπινη υγεία ενώ στην πραγματικότητα θα ήταν.

Από τα αποτελέσματα για την κλάση της None διαπιστώνει κανείς ότι ο αλγόριθμος θα μπορούσε να χρησιμοποιηθεί για να προβλέπει ανάμεσα σε δύο κλάσεις, όπου η μια θα αφορούσε στην περίπτωση που το νανοϋλικό θα ήταν επικίνδυνο για την ανθρώπινη υγεία, ομαδοποιώντας τις κλάσεις Low, Medium, High, και η άλλη θα αφορούσε την περίπτωση που δεν θα ήταν. Τότε, σύμφωνα με τα αποτελέσματα που

ήδη έχουμε για τα TP, FP, FN και TN, ο πίνακας σύγχυσης θα μπορούσε να διαμορφωθεί όπως φαίνεται στον Πίνακα 28.

**Πίνακας 28. Αναγωγή του προβλήματος κατηγοριοποίησης τεσσάρων κλάσεων σε πρόβλημα δύο κλάσεων.**

| | | | |
|---|---|---|---|
| Predicted | 0 | TN=32 | FN=9 |
| | 1 | FP=19 | TP=31 |
| | | 0 | 1 |
| | | Truth | |

Στον Πίνακα 28 το 0 αντιστοιχεί στην περίπτωση που το νανοϋλικό δεν είναι ακίνδυνο και το 1 στην περίπτωση που ετο ναναϋλικό είναι ακίνδυνο. Ο νέος δείκτης Accuracy υπλογίζεται:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} = 0.692$$

Δηλαδή ο αλγόρθμος θα είχε ακρίβεια 69%.

Επιστρέφοντας στο αρχικό πρόβλημα κατηγοριοποίησης τεσσάρων κλάσεων, ξεκινώντας με την κλάση Low έχουμε:

$$TP_{Low} = 6$$

$$FP_{Low} = 13$$

$$FN_{Low} = 9$$

$$Sensitivity_{Low} = 0.4$$

$$Precision_{Low} = 0.315$$

Για την κλάση Medium:

$$TP_{Medium} = 0$$

$$FP_{Medium} = 7$$

$$FN_{Medium} = 8$$

$$Sensitivity_{Medium} = 0$$

$$Precision_{Medium} = 0$$

Και τέλος για την κλάση High:

$$TP_{High} = 10$$

$$FP_{High} = 5$$

$$FN_{High} = 18$$

$$Sensitivity_{High} = 0.357$$

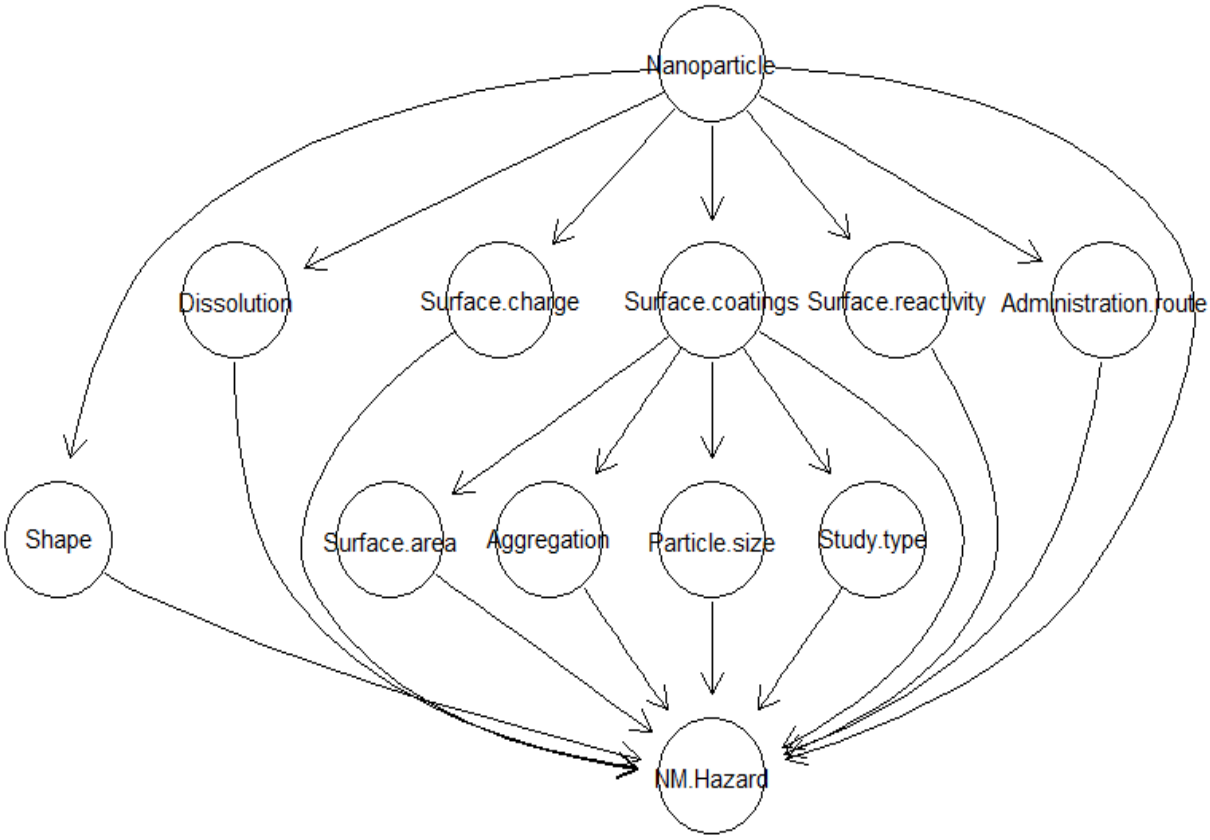
$$Precision_{High} = 0.666$$

Μπορεί να παρατηρήσει κανείς ότι ο αλγόριθμος δεν παρουσιάζει καλά αποτελέσματα για τις κλάσεις Low, Medium και High και ειδικά για τις κλάσεις Low και Medium. Συγκεκριμένα, για την κλάση Medium παρουσίασε πλήρη αδυναμία στο να βρεθούν εκείνες οι περιπτώσεις νανοϋλικών που θα είχαν μέτρια επίδραση στην ανθρώπινη υγεία, αφού το $Sensitivity$ που ήταν ίσο με 0 φανερώνει ότι από όλες εκείνες τις περιπτώσεις νανοϋλικών που πραγματικά θα είχαν μέτρια επίδραση στον οργανισμό δεν μπόρεσε να προβλέψει καμία σωστά. Αυτό σημαίνει ότι το δίκτυο δεν κατάφερε να εγκαταστήσει κάποια προβλεπτική σχέση μεταξύ των μεταβλητών εισόδου και της κλάσεις Medium.

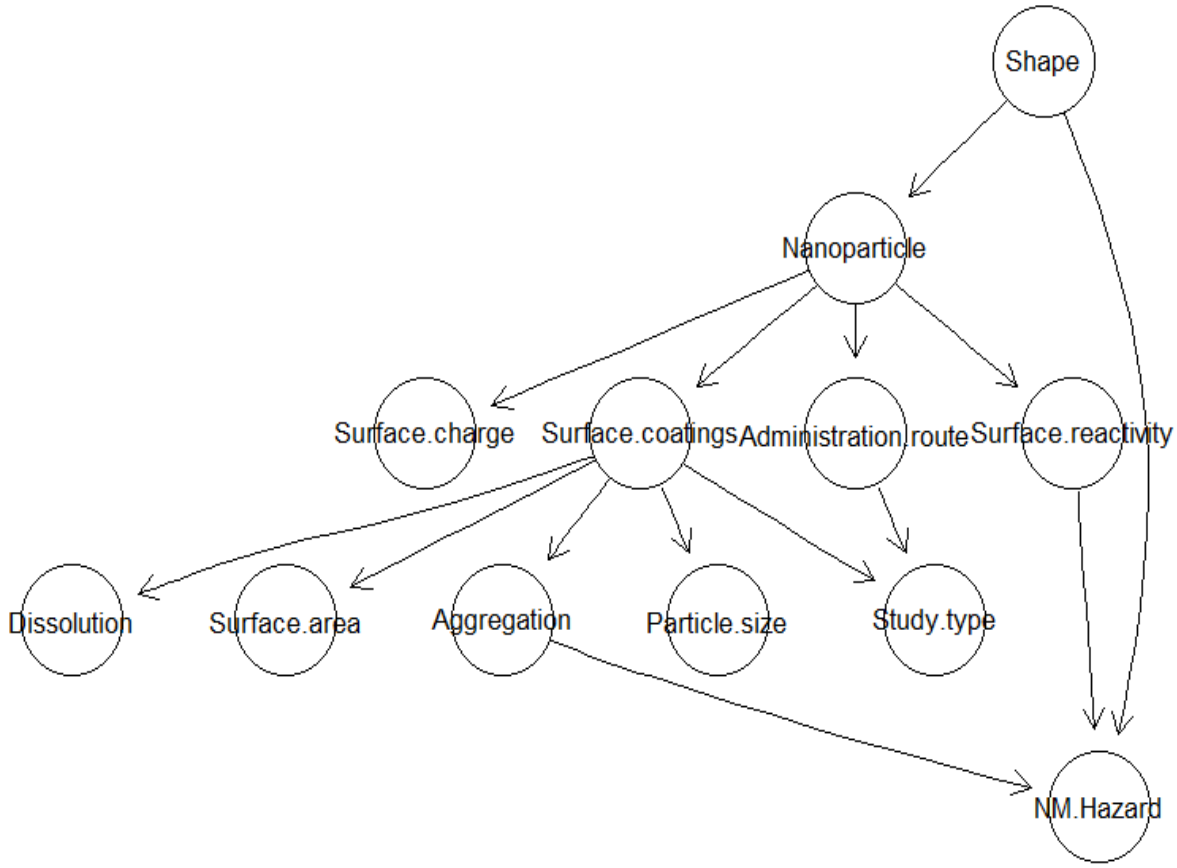
Η κακή προβλεπτική ικανότητα για τις κλάσεις Medium και Low μπορεί να οφείλεται στο ότι υπήρχαν λιγότερα στιγμιότυπα στο training set με αυτές τις κλάσεις σε σχέση με τις άλλες δύο. Συγκεκριμένα, οι σχετικές συχνότητες για τις κλάσεις High, Medium, Low και None στο training set ήταν 29%, 11%, 18% και 42% αντίστοιχα.

#### *5.3.2.1 Χρήση Διαφορετικού Αρχικού Δικτύου*

Όπως διαπιστώθηκε από τα αποτελέσματα του Κεφαλαίου 5.3.1, το αρχικό δίκτυο με το οποίο θα ξεκινήσει την διαδικασία εύρεσης ο δομικός αλγόριθμος ΕΜ επηρεάζει το τελικό αποτέλεσμα. Για να επαληθευτεί αυτό το συμπέρασμα, δομήθηκε ένα νέο δίκτυο το οποίο τροφοδοτήθηκε στην βέλτιστη στρατηγική. Στο νέο δίκτυο αφαιρέθηκαν οι κόμβοι των βιολογικών συνεπειών και ενώθηκαν οι 11 μεταβλητές εισόδου με την μεταβλητή NM Hazard με τέτοιο τρόπο ώστε οι μεταβλητές εισόδου να είναι οι γονείς και η NM Hazard να είναι το παιδί, ενώ διατηρήθηκαν οι σχέσεις μεταξύ των μεταβλητών εισόδου. Το δίκτυο εκκίνησης παρουσιάζεται στην Εικόνα 34, ενώ το δίκτυο που παρήχθη φαίνεται στην Εικόνα 35.




**Εικόνα 34. Δομή νέου αρχικού δικτύου με τη μεταβλητή NM Hazard ως παιδί.**



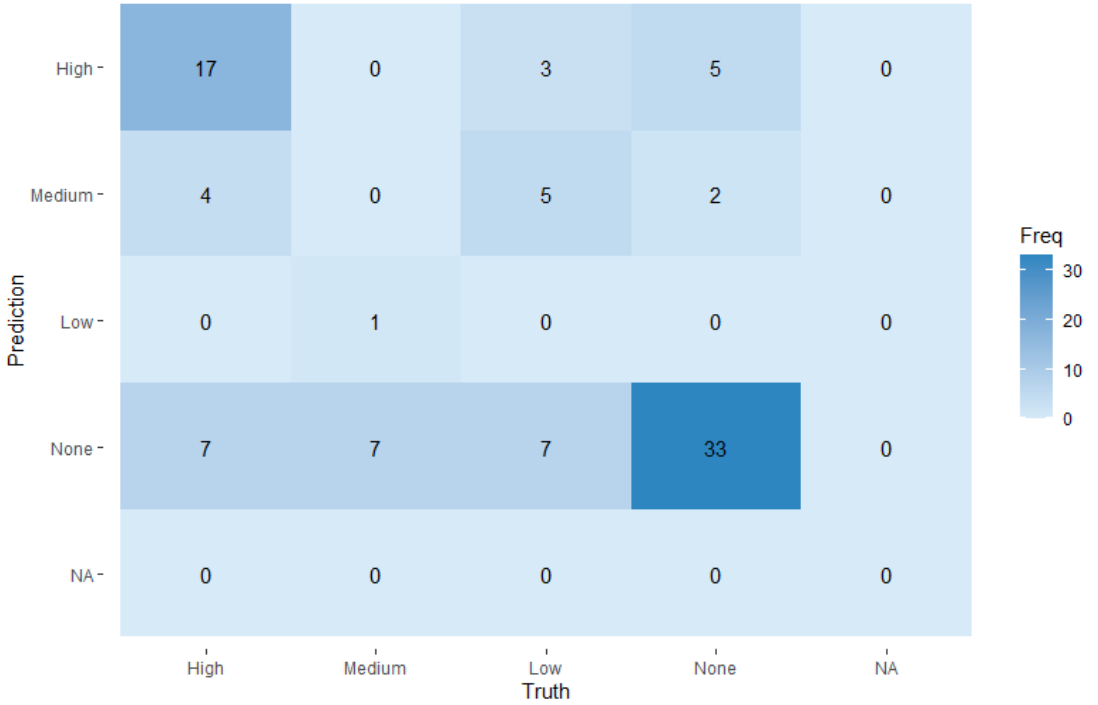

**Εικόνα 35. Η τελική δομή δικτύου όταν το δίκτυο εκκίνησης αυτό που είχε την NM Hazard ως παιδί.**

Τα αποτελέσματα σε σχέση με τους δείκτες Accuracy και MCC ήταν τα εξής:

$$Accuracy = 0.549$$

$$MCC = 0.311$$

Ενώ ο αντίστοιχος πίνακας σύγχυσης παρουσιάζεται στην Εικόνα 36.




**Εικόνα 36. Ο πίνακας σύγχυσης για τη βέλτιστη στρατηγική με δίκτυο εκκίνησης αυτό που είχε την NM Hazard ως παιδί.**

Όπως μπορεί να παρατηρήσει κανείς, τα αποτελέσματα για την ακρίβεια βελτιώθηκαν, αφού στην περίπτωση της χρήσης του μπεϋζιανού δικτύου των Marvin *et al.* (2017) είχαμε ακρίβεια 51.6% ενώ στη συγκεκριμένη περίπτωση η ακρίβεια ανέβηκε κοντά στο 55%. Το αποτέλεσμα αυτό ενισχύει ακόμα περαιτέρω το συμπέρασμα ότι το αρχικό δίκτυο που θα χρησιμοποιηθεί επηρέαζει τα αποτελέσματα της βέλτιστης στρατηγικής. Ενδεικτικά, οι δείκτες Sensitivity και Precision για την κλάση None είναι:

$$TP_{None} = 33$$

$$FP_{None} = 5 + 2 = 7$$

$$FN_{None} = 7 + 7 + 7 = 21$$

$$Sensitivity_{None} = \frac{TP_{None}}{TP_{None} + FN_{None}} = 0.825$$

$$Precision_{None} = \frac{TP_{None}}{TP_{None} + FP_{None}} = 0.611$$

Σημειώθηκε βελτίωση στο Sensitivity σε σχέση με την προηγούμενη αρχική δομή που το αποτέλεσμα ήταν 0.775, ενώ το Precision παρέμεινε σχεδόν σταθερό. Άρα

βελτιώθηκε ο αλγόριθμος στο να αποφεύγει λάθη FN, δηλαδή μέιωσε τις περιπτώσεις που θεώρησε ότι το νανοϋλικό δεν ήταν ακίνδυνο ενώ στην πραγματικότητα ήταν.

Όλοι οι κώδικες που χρησιμοποιήθηκαν για την παραγωγή των αποτελεσμάτων που παρουσιάστηκαν έχουν μεταφορτωθεί στο ηλεκτρονικό αποθετήριο της Μονάδας Αυτόματης Ρύθμισης και Πληροφορικής της Σχολής Χημικών Μηχανικών, στην παρακάτω διεύθυνση:

https://github.com/ntua-unit-of-control-and-informatics/BN_NM_HAZARD

# 6 Συμπεράσματα

Κατά τη διάρκεια της παρούσας διπλωματικής διαπιστώθηκε η δυνατότητα που έχουν τα μπεϋζιανά δίκτυα να αναπαριστούν με γραφικό τρόπο τις σχέσεις εξάρτησης μεταξύ των μεταβλητών, όπως αυτές εκφράζονται μέσω γραφημάτων τύπου DAG. Το πλεονέκτημα που παρέχεται είναι ότι αυτές οι γραφικές σχέσεις εξάρτησης είναι σχέσεις αίτιου-αιτιατού, οι οποίες με τη σειρά τους αντιστοιχούν σε στοχαστικού τύπου σχέσεις. Στο κομμάτι της στοχαστικής έκφρασης αυτών των σχέσων, τα μπεϋζιανά δίκτυα παρέχουν τη δυνατότητα μιας κλειστού τύπου μαθηματικής έκφρασης η οποία αντανακλά τις εξαρτήσεις μεταξύ των μεταβλητών ως δεσμευμένες κατανομές. Το γεγονός αυτό διευκολύνει την εύρεση της γενικής κατανομής που ακολουθούν οι μεταβλητές του δικτύου, αφού μετά από αποσύνθεση, δεδομένων των υποθέσεων ανεξαρτησίας που υποδεικνύονται από το δίκτυο, δεν είναι τίποτα άλλο παρά το γινόμενο των τοπικών δεσμευμένων κατανομών, όπως αυτές διαμορφώνονται από τις σχέσεις εξάρτησης των μεταβλητών.

Επίσης, φάνηκε η δυνατότητα που έχουν τα μπεϋζιανά δίκτυα να ανανεώνουν την πεποίθηση μας για την κατάσταση των μεταβλητών υπό το φώς νέων δεδομένων. Πράγματι η δυνατότητα αυτή σημαίνει ότι δίνονται απαντήσεις σε ερωτήσεις (queries) που γίνονται στο δίκτυο και οι οποίες αφορούν την πιθανότητα να συμβεί κάτι σε μια μεταβλητή υπό τη συνθήκη ότι έχουμε νέα στοιχεία για τις υπόλοιπες μεταβλητές. Κάτι τέτοιο φάνηκε χαρακτηριστικά κατά τη διάρκεια των ερωτήσεων που έγιναν για την κατάσταση της μεταβλητής NM Hazard στην εφαρμογή, όπου δίνονταν απαντήσεις για την κατάσταση της μεταβλητής για κάθε νέο στιγμιότυπο το test set.

Επιπλέον, επιβεβαιώθηκε η ικανότητα των μπεϋζιανων δικτύων να παρέχουν αποτελέσματα ακόμα και στην περίπτωση που το σύνολο των δεδομένων δεν είναι πλήρες. Κάτι τέτοιο παρατηρήθηκε στην $1^{η}$ στρατηγική, όπου χρησιμοποιήθηκε η δομή που κατέληξαν στην έρευνά τους οι Marvin *et al.* (2017) και πραγματοποιήθηκε η εύρεση του στατιστικού μοντέλου με βάση τη συγκεκριμένη δομή. Τα δεδομένα που χρησιμοποιήθηκαν δεν ήταν πλήρη και δεν χρησιμοποιήθηκε κάποια μέθοδος imputation για την συμπλήρωση τους πριν τη διαδικασία της εκπαίδευσης. Ωστόσο παρήχθησαν αποτελέσματα και μάλιστα στην υποπερίπτωση που η εύρεση του στατιστικού μοντέλου έγινε με την μέθοδο bayes και η εκτίμηση των δεδομένων έγινε με την likelihood weighting, η ακρίβεια έφτασε το 59%. Αποτελέσματα παράχθηκαν

και στη $2^{η}$ στρατηγική, όπου πάλι χρησιμοποιήθηκαν τα μη πλήρη δεδομένα ώστε να κατασκευαστεί το μπεϋζιανό δίκτυο με τους constrained based αλγόριθμους gs και iamb. Ωστόσο, στη $2^{η}$ στρατηγική τα αποτελέσματα οσον αφορά την ακρίβεια ήταν αισθητά χειρότερα και αυτό συνέβη διότι οι constrained based αλγόριθμοι δεν κατάφεραν να βρούν στατιστικά σημαντικές σχέσεις εξάρτησης ανάμεσα στις μεταβλητές.

Η ακρίβεια των αποτελεσμάτων βελτιώθηκε στην $3^{η}$ και την $4^{η}$ στρατηγική, όταν χρησιμοποιήθηκε ο δομικός αλγόριθμος EM ο οποίος καλεί τον αλγόριθμο EM που πραγματοποιεί imputation στο σύνολο των δεδομένων. Τα καλύτερα αποτελέσματα, χρησιμοποιώντας το accuracy ως στατιστικό μέτρο, βρέθηκε να έχει η υποπερίπτωση της $4^{ης}$ στρατηγικής όπου χρησιμοποιείται η score based μεθοδος tabu, η συνάρτηση βαθμολόγησης aic και η likelihood weighting, ενώ το αρχικό δίκτυο εκκίνησης του αλγόριθμου EM ήταν το μπεϋζιανό δίκτυο στο οποίο κατέληξε η έρευνα των Marvin *et al*. (2017).

Η υποπερίπτωση αυτή χρησιμοποιήθηκε για να γίνει σύγκριση των αποτελεσμάτων με εκείνα που παράχθηκαν στην συγκεκριμένη έρευνα, στην οποία επίσης χρησιμοποιήθηκε ο δομικός αλγόριθμος EM. Η μεγάλη διαφορά στο accuracy οφείλεται κατά πάσα πιθανότητα στο ότι δεν υπήρχε η γνώση του αρχικού δικτύου με το οποίο ξεκίνησε ο δομικός αλγόριθμος EM στην έρευνα των Marvin *et al*. (2017). Επιπλέον, δεν ήταν γνωστό αν δόθηκε κάποια συγκεκριμένη αρχικοποίηση των παραμέτρων στην έρευνα των Marvin *et al.* (2017), κάτι που θα επηρέαζε αρκετά τα τελικά τους αποτελέσματα. Συνεπώς, ο δομικός αλγόριθμος EM της παρούσας διπλωματικής εργασίας, που ξεκίνησε την διαδικασία με το τελικό μπεϋζιανό δίκτυο της έρευνας των Marvin *et al.* (2017) και τυχαία παραμετρική αρχικοποίηση, δεν είναι παράλογο που κατέληξε σε διαφορετικά αποτελέσματα. Το συμπέρασμα που εξάγεται τελικά είναι ότι η αρχική δομή και το αρχικό σύνολο παραμέτρων με το οποίο θα ξεκινήσει τη διαδικασία ο δομικός αλγόριθμος EM αποτελούν στημαντικό κομμάτι επιρροής στην ακρίβεια των προβλέψεων. Για να επιβεβαιωθεί αυτό το συμπέρασμα, επιλέχθηκε ένα διαφορετικό αρχικό δίκτυο στο οποίο η μεταβλητή NM hazard ήταν το παιδί των μεταβλητών εισόδου και το αρχικό σύνολο των παραμέτρων του δικτύου ήταν πάλι τυχαίο. Παρατηρήθηκε ότι τα αποτελέσματα στην ακρίβεια βελτιώθηκαν, επιβεβαιώνοντας το συμπέρασμα πως η αρχική δομή δικτύου επηρέαζει την τελική ακρίβεια ενός μοντέλου που παράγεται με τον δομικό EM.

Σε μελλοντική επέκταση αυτής της εργασίας θα μπορούσε να γίνει εξέταση διαφορετικών αρχικών δομών δικτύων ως αρχικές δόμές εκκίνησης του δομικού αλγόριθμου EM, ώστε να διερευνηθεί περαιτέρω ο ρόλος που διαδραματίζει η αρχική δομή στην ακρίβεια των προβλέψεων. Συμπληρωματικά σε αυτό, θα μπορούσε να εξεταστεί επίσης η επιρροή του αρχικού συνόλου των παραμέτρων στα τελικά αποτελέσματα. Τέλος, θα μπορούσαν να ερευνηθούν διαφορετικές τεχνικές imputation του συνόλου των δεδομένων, όπως είναι η συμπλήρωση κενών με τη μέθοδο των κοντινότερων γειτόνων (Nearest Neighbors) (Beretta and Santaniello, 2016) και να επαναληφθεί η $1^{η}$ στρατηγική με βάση το συμπληρωμένο σύνολο δεδομένων, με σκοπό να φανεί σε ποιό βαθμό δύναται να βελτιωθεί η ακρίβεια των αποτελεσμάτων. Το ίδιο μπορεί να γίνει και εφαρμόζοντας τη $2^{η}$ στρατηγική, ώστε να διαπιστωθεί ο βαθμός στον οποίο επηρεάζουν τα μη πλήρη δεδομένα τη δυνατότητα εύρεσης εξαρτήσεων ανάμεσα στις μεταβλητές των constrained based αλγόριθμων gs και iamb.

# Βιβλιογραφία


Κοκολάκης, Γ. και Σπηλιώτης, Ι. (2002) Εισαγωγή στις πιθανότητες. Αθήνα: Εκδόσεις Συμεών.

Κοκολάκης, Γ. και Φουσκάκης, Δ. (2009) Στατιστική θεωρία και εφαρμογές. Αθήνα: Εκδόσεις Συμεών.

Χαραλαμπίδης, Α.Χ. (2009) Θεωρία πιθανοτήτων και εφαρμογές. Αθήνα: Εκδόσεις Συμμετρία.

Beretta, L. and Santaniello, A. (2016). Nearest neighbor imputation algorithms: a critical evaluation. *BMC medical informatics and decision making*, *16 Suppl 3*(Suppl 3), 74. https://doi.org/10.1186/s12911-016-0318-z

Bernardo, J.M. and Smith, A.F.M. (2000) Bayesian theory. West Sussex: John Wiley and Sons.

Bozdogan, H. (1987) Model selection and Akaike's Information Criterion (AIC): The general theory and its analytical extensions. Psychometrika 52, 345–370. https://doi.org/10.1007/BF02294361

Gelman, A., Carlin, J., Stern, H., Dunson, D., Vehtari, A., Rubin, D. (2013). Bayesian data analysis. 3rd edn. Florida: CRC Press.

Cowel, R.G., Dawid, P.A., Lauritzen, S.L. and Spiegelhalter, D.J. (1999) Probabilistic networks and expert systems. New York, NY: Springer-Verlag.

Darwiche, A. (2009) Modeling and reasoning with bayesian networks. New York, NY: Cambridge University Press.

Hansen, K. D., Gentry, J., Long, L., Gentleman, R., Falcon, S., Hahne, F. and Sarkar. D. (2019) Rgraphviz: Provides plotting capabilities for R graph objects. R package version 2.30.0.

Jensen, F.V. and Nielsen, T.D. (2007) Bayesian networks and decision graphs. 2nd edn. New York, NY: Springer.

Keller, D.J., Mac Namee, B. and D'Arcy, A. (2015) Fundamentals of machine learning for predictive data analytics: algorithms, worked examples, and case studies. Massachusett: MIT Press.

Kjaerulff, U.B. and Madsen A.L. (2013) Bayesian networks and influence diagrams: a guide to construction and analysis. 2nd edn. New York: Springer Science and Business Media.

Koller, D. and Friedman, N. (2009) Probabilistic graphical models: principles and techniques. Cambridge: MIT Press.

Korb, K.B. and Nicholson, A.E. (2011) Bayesian artificial intelligence. 2nd edn. New York: CRC Press.

Kuhn and Vaughan (2020) yardstick: Tidy Characterizations of Model Performance. R package version 0.0.5. https://CRAN.R-project.org/package=yardstick

Margaritis, D (2003) Learning bayesian network model structure from data. Ph.D. Thesis, School of Computer Science, Carnegie Mellon University, Pittsburgh, PA. Available at: http://reports-archive.adm.cs.cmu.edu/anon/2003/abstracts/03-153.html

Marvin, J.P.H., Bouzembrak, Y., Janssen, M.E., Van der Zande, M., Murphy, F., Sheehan, B., Mullins, M. and Bouwmeester, H. (2017) Application of Bayesian networks for hazard ranking of nanomaterials to support human health risk assessment. Available at: https://doi.org/10.1080/17435390.2016.1278481

Nagarajan, R., Scutari, M. and Lèbre, S. (2013) Bayesian networks in R. New York, NY: Springer.

Neapolitan, R.E. (2003) Learning bayesian networks. New Jersey: Pearson Prentice Hall

Pearl, J. (2009) Causality: models, reasoning and inference. 2nd edn. Cambridge: Cambridge University Press.

Schwarz, G.E. (1978), "Estimating the dimension of a model", Annals of Statistics, 6 (2): 461–464, doi:10.1214/aos/1176344136

Scutari, M. and Denis, J. (2014) Bayesian networks with examples in R. Boca Raton: Chapman and Hall/CRC

Tsamardinos, I., Aliferis, C.F. and Statnikov, A. (2003). Algorithms for large scale markov blanket discovery. In Proceedings of the 16th international Florida artificial intelligence research society conference. AAAI Press, Vanderbilt University, Nashville.

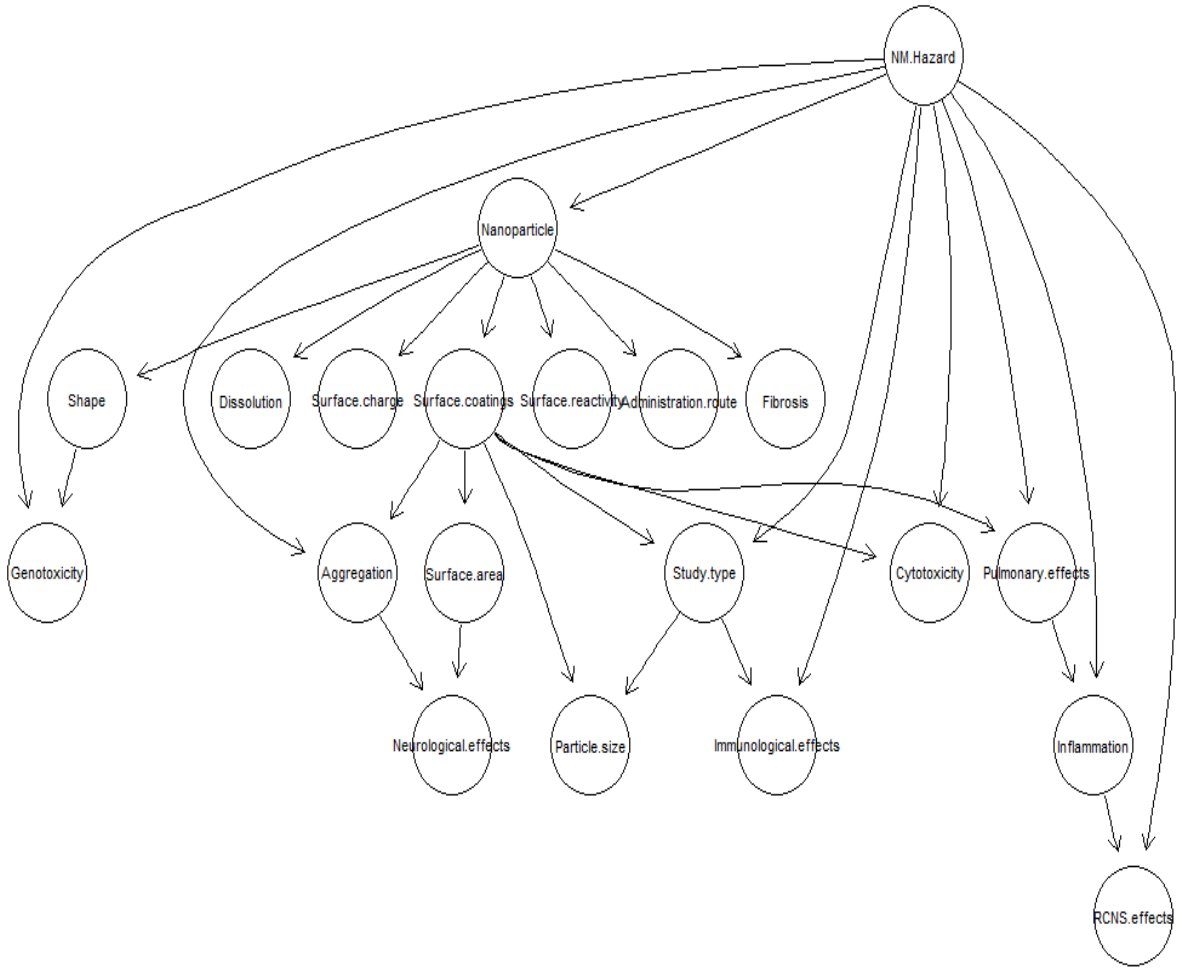
Wickham, H., Francois, R., Henry, L. and Muller, K. (2020). dplyr: A Grammar of Data Manipulation. R package version 0.8.4. https://CRAN.R-project.org/package=dplyr

Witten, I.H., Frank, E. and Hall, M.A. (2011) Data Mining: practical machine learning tools and techniques. 3rd edn. San Fransisco: Morgan Kaufmann.

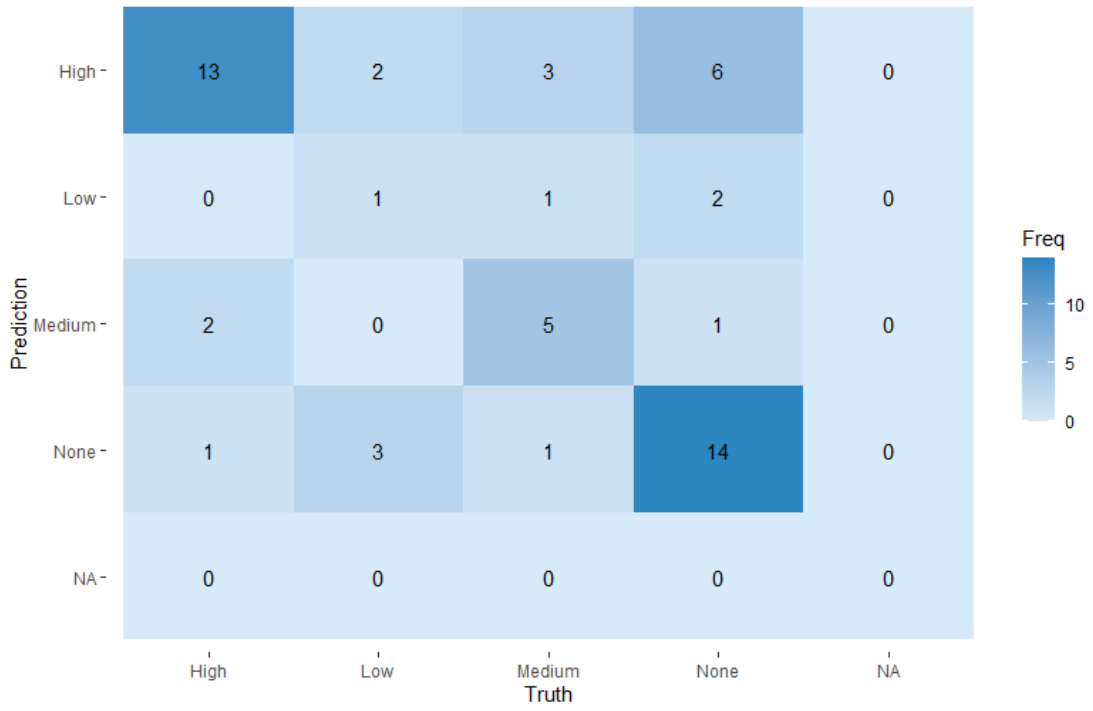
# Παράρτημα

*Δομές δικτύων και πίνακες σύγχυσης*

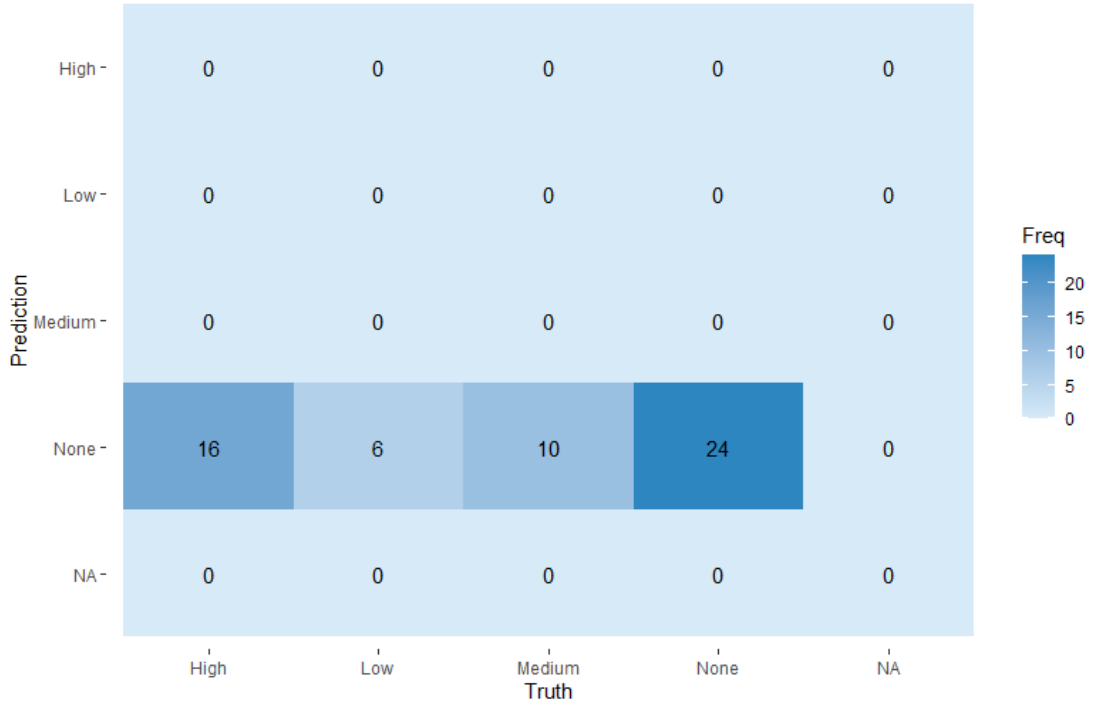
*1η Περίπτωση*: Γνωστή δομή δικτύου



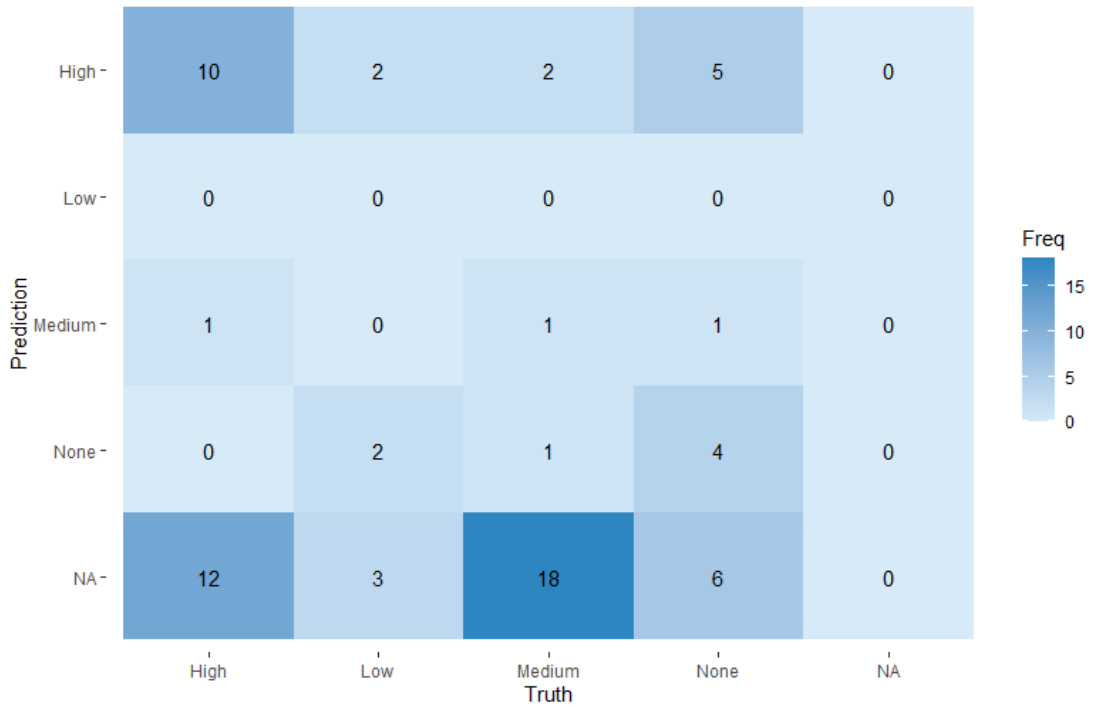

**Εικόνα Παραρτήματος 1. Η δομή του μπεϋζιανού δικτύου από τους Marvin *et al*. (2017).**



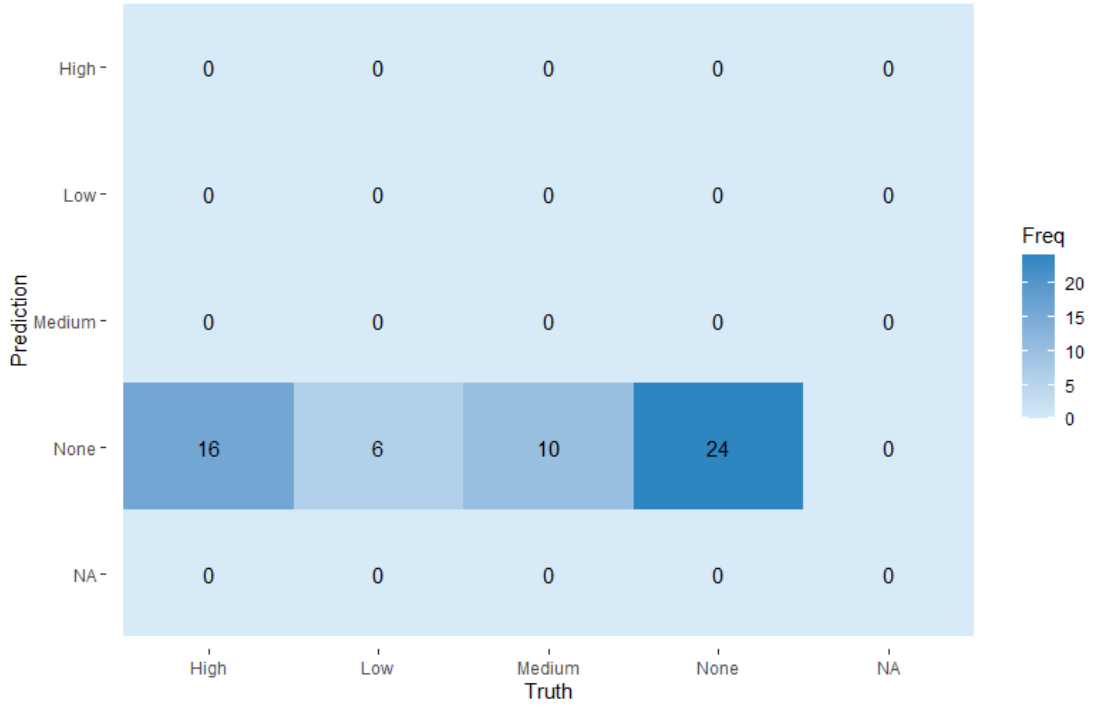

**Εικόνα Παραρτήματος 2. Πίνακας σύγχυσης για εύρεση παραμέτρων με τη μέθοδο bayes και πρόβλεψη με τη μέθοδο bayes lw.**




**Εικόνα Παραρτήματος 3. Πίνακας σύγχυσης για εύρεση παραμέτρων με τη μέθοδο bayes και πρόβλεψη με τη μέθοδο parents.**

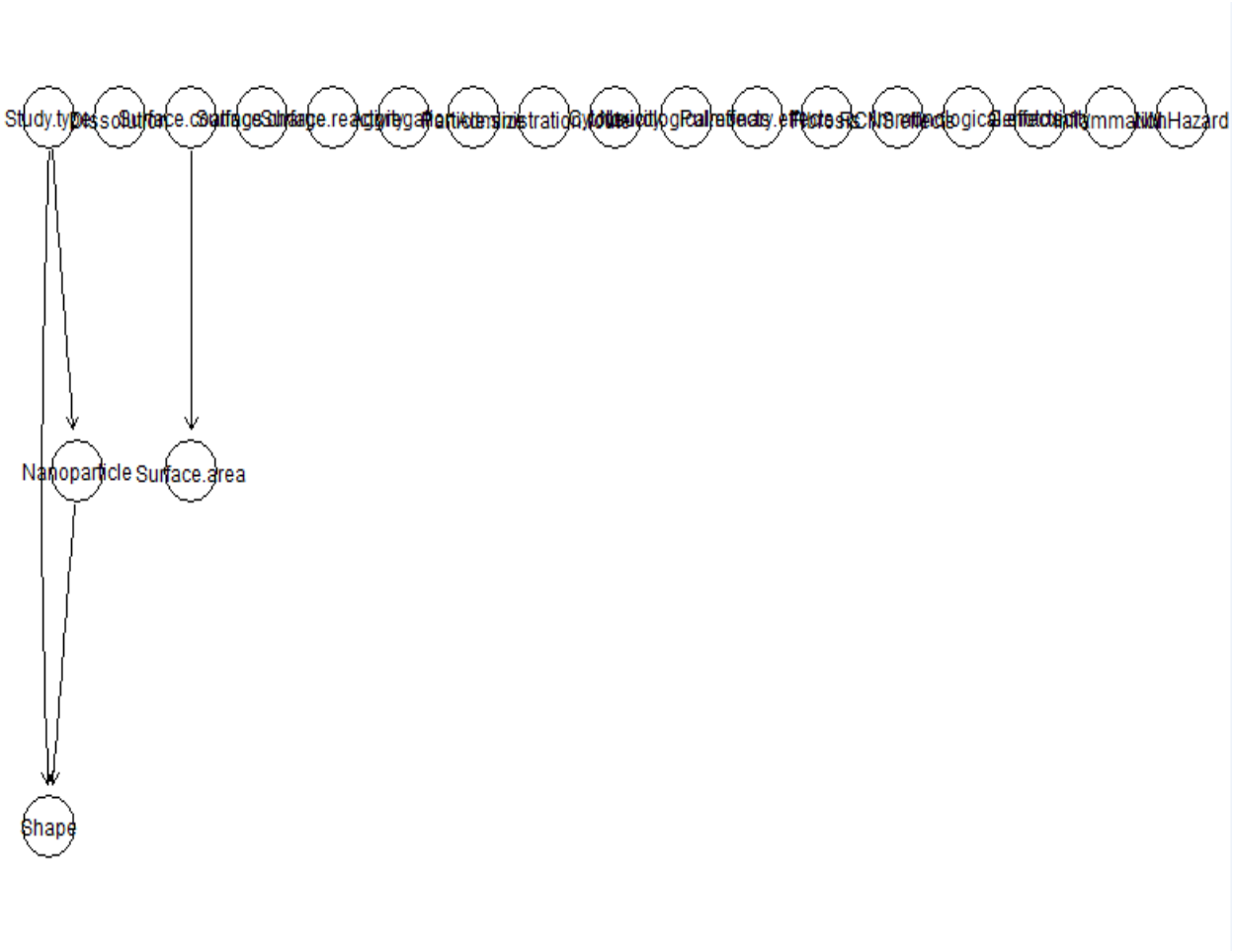



**Εικόνα Παραρτήματος 4. Πίνακας σύγχυσης για εύρεση παραμέτρων με τη μέθοδο mle και πρόβλεψη με τη μέθοδο bayes lw.**

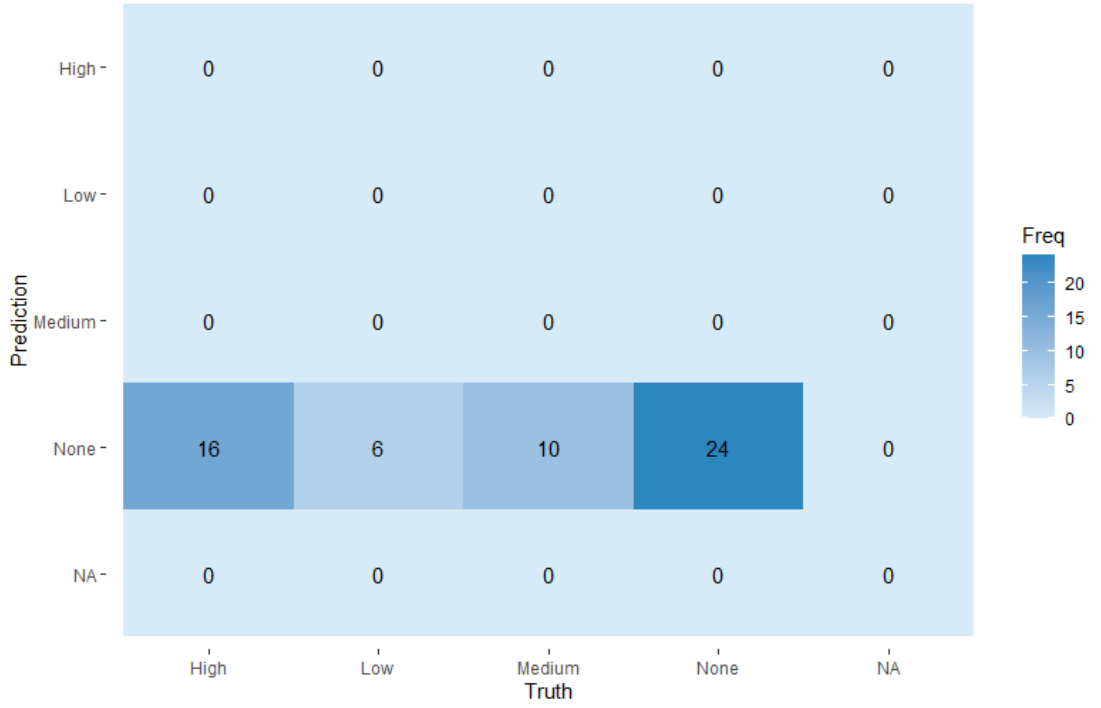



**Εικόνα Παραρτήματος 5. Πίνακας σύγχυσης για εύρεση παραμέτρων με τη μέθοδο mle και πρόβλεψη με τη μέθοδο parents.**

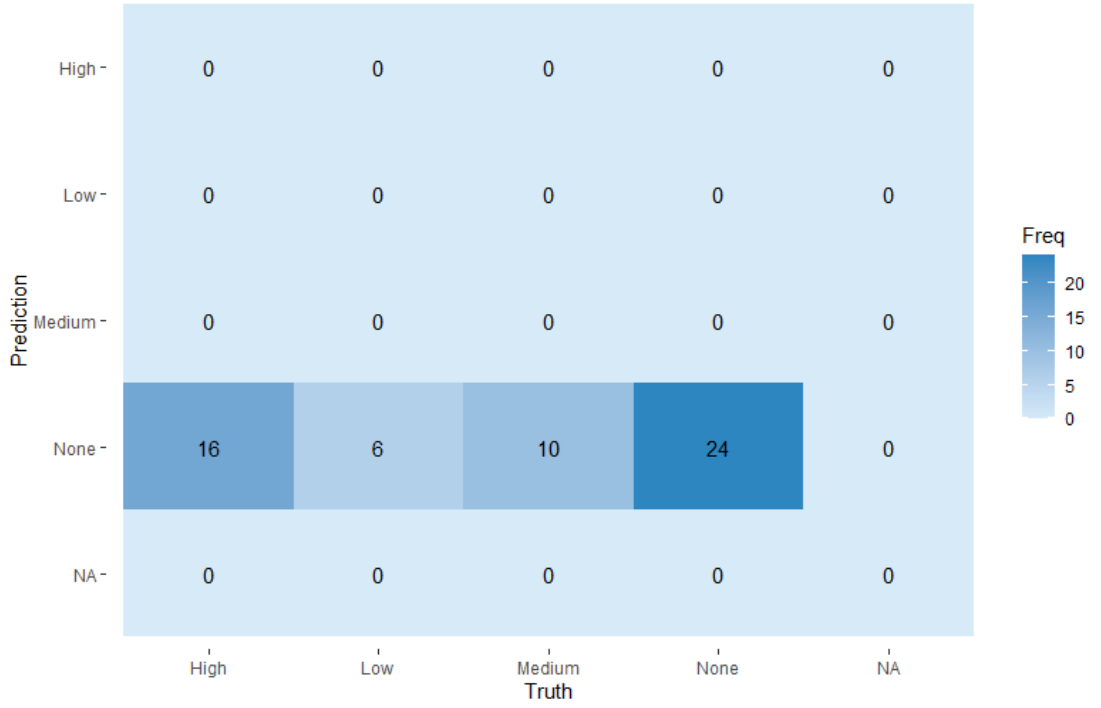
*2η Περίπτωση*: Κατασκευή δομής από τα δεδομένα με constrained based αλγόριθμους



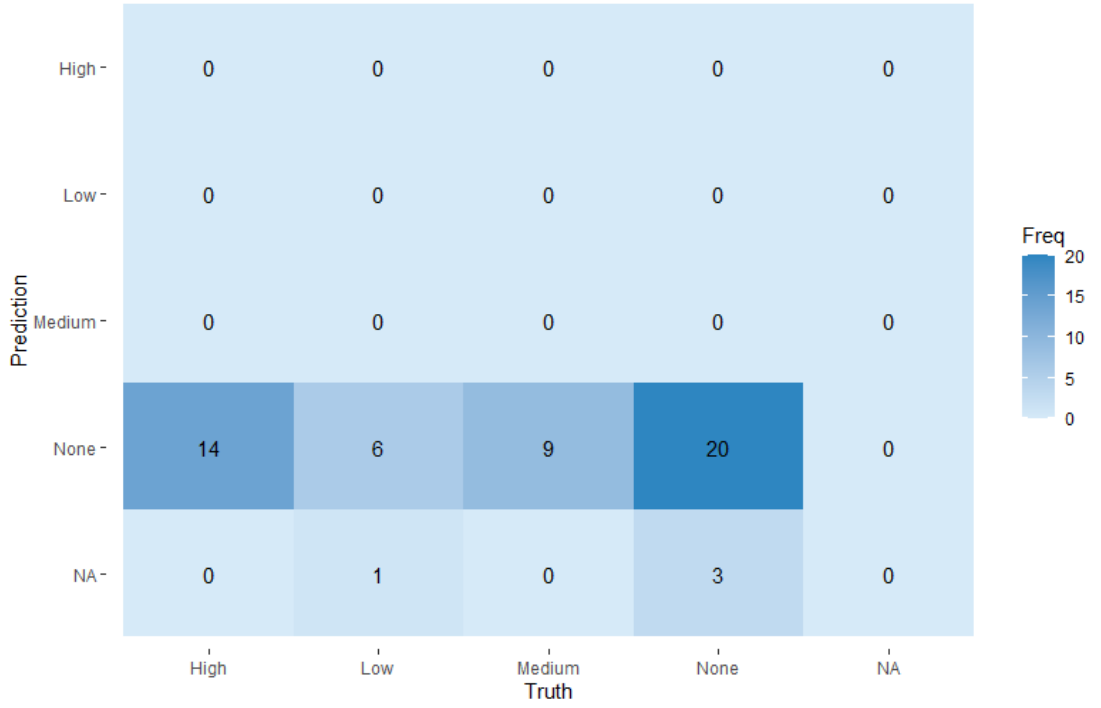

**Εικόνα Παραρτήματος 6. Η δομή του δικτύου ως αποτέλεσμα του constrained based αλγόριθμου gs.**



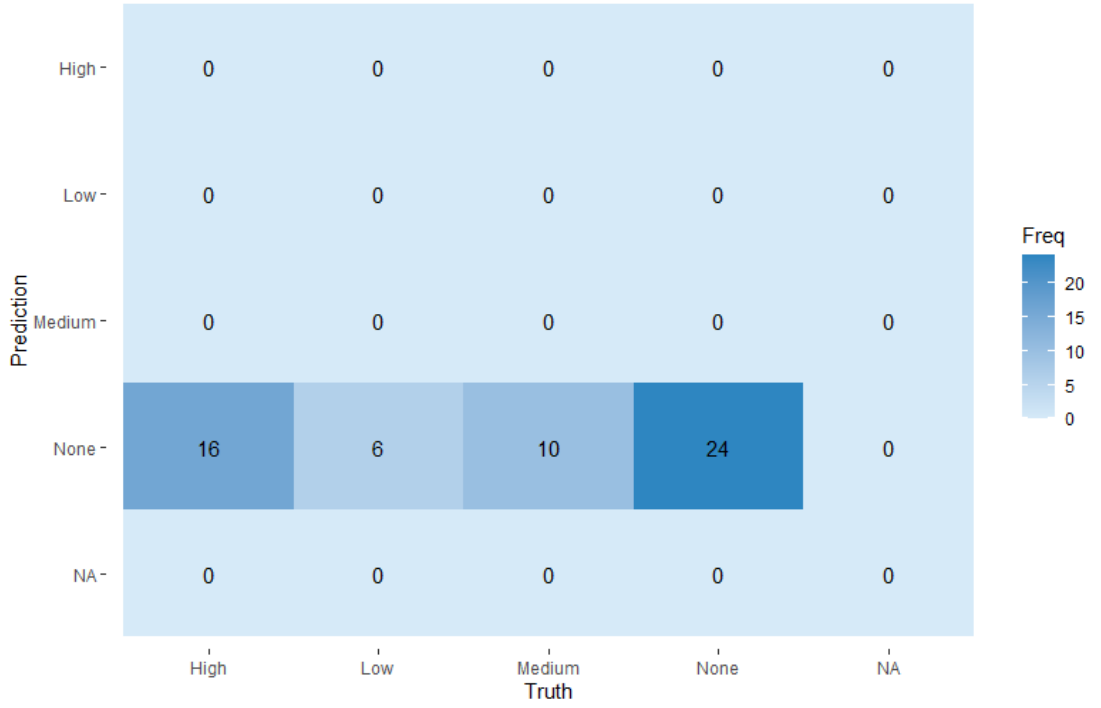

**Εικόνα Παραρτήματος 7. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο gs, εύρεση παραμέτρων με τη μέθοδο bayes και πρόβλεψη με τη μέθοδο bayes lw.**



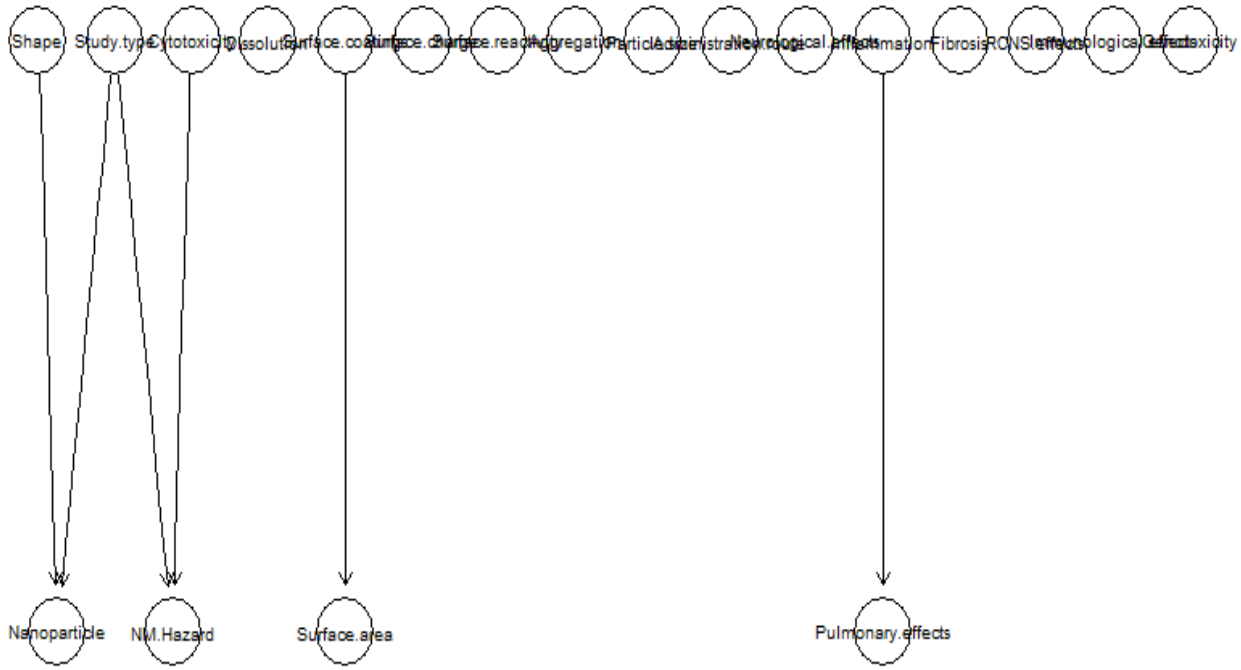

**Εικόνα Παραρτήματος 8. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο gs, εύρεση παραμέτρων με τη μέθοδο bayes και πρόβλεψη με τη μέθοδο parents.**



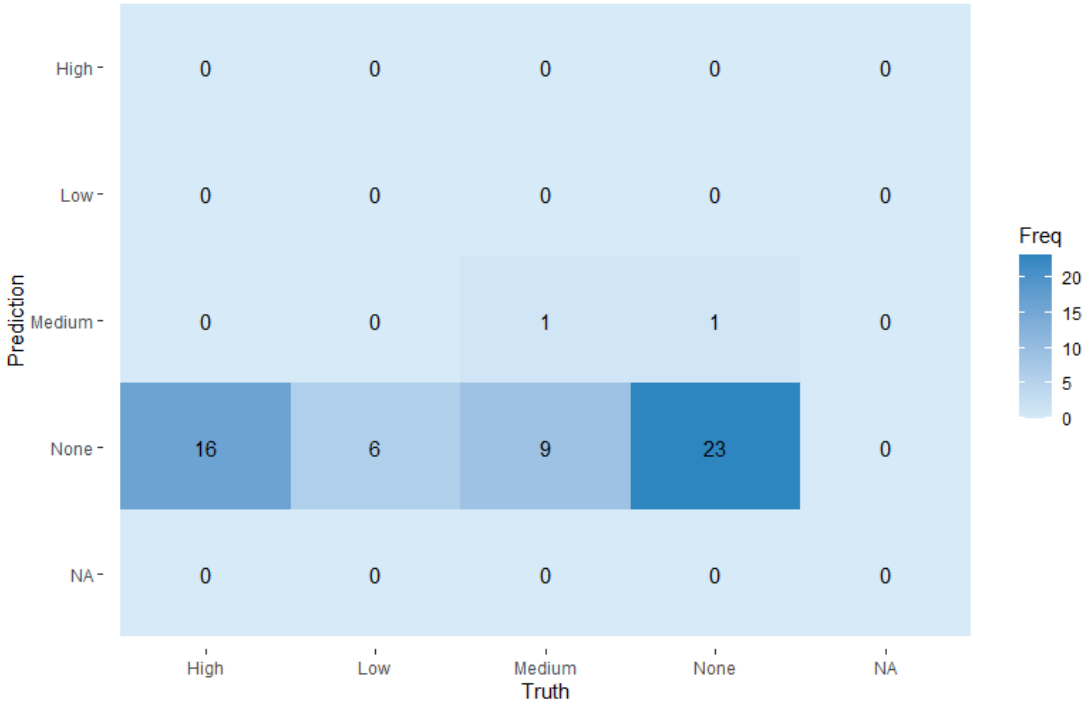

**Εικόνα Παραρτήματος 9. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο gs, εύρεση παραμέτρων με τη μέθοδο mle και πρόβλεψη με τη μέθοδο bayes lw.**



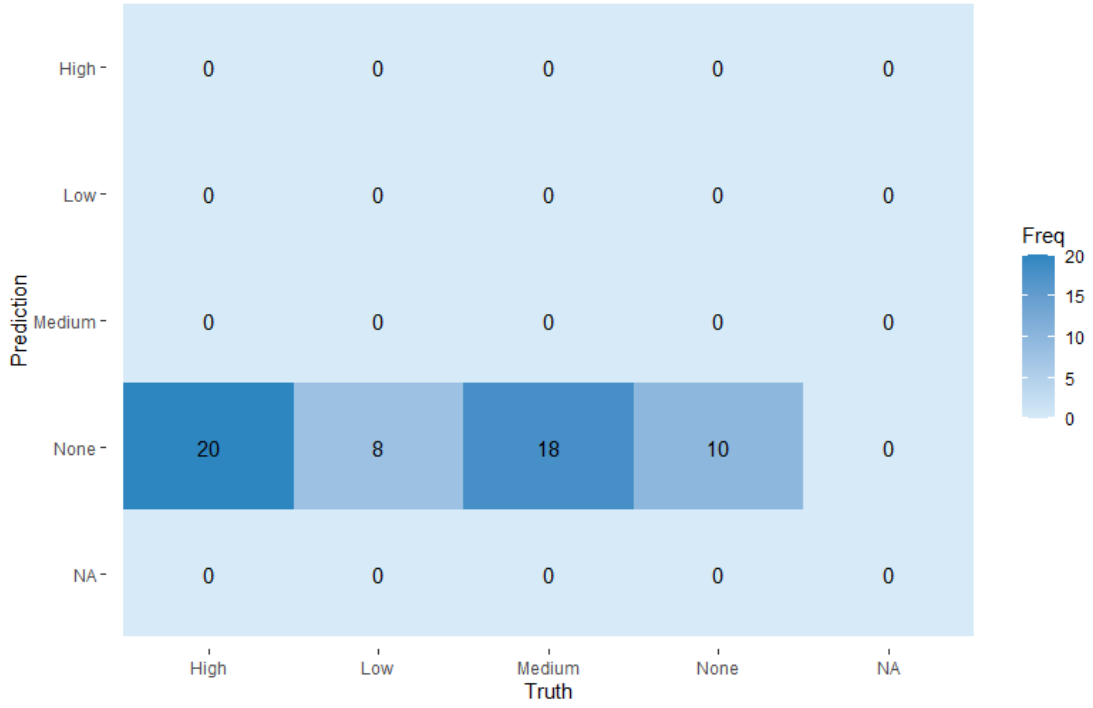

**Εικόνα Παραρτήματος 10. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο gs, εύρεση παραμέτρων με τη μέθοδο mle και πρόβλεψη με τη μέθοδο parents.**



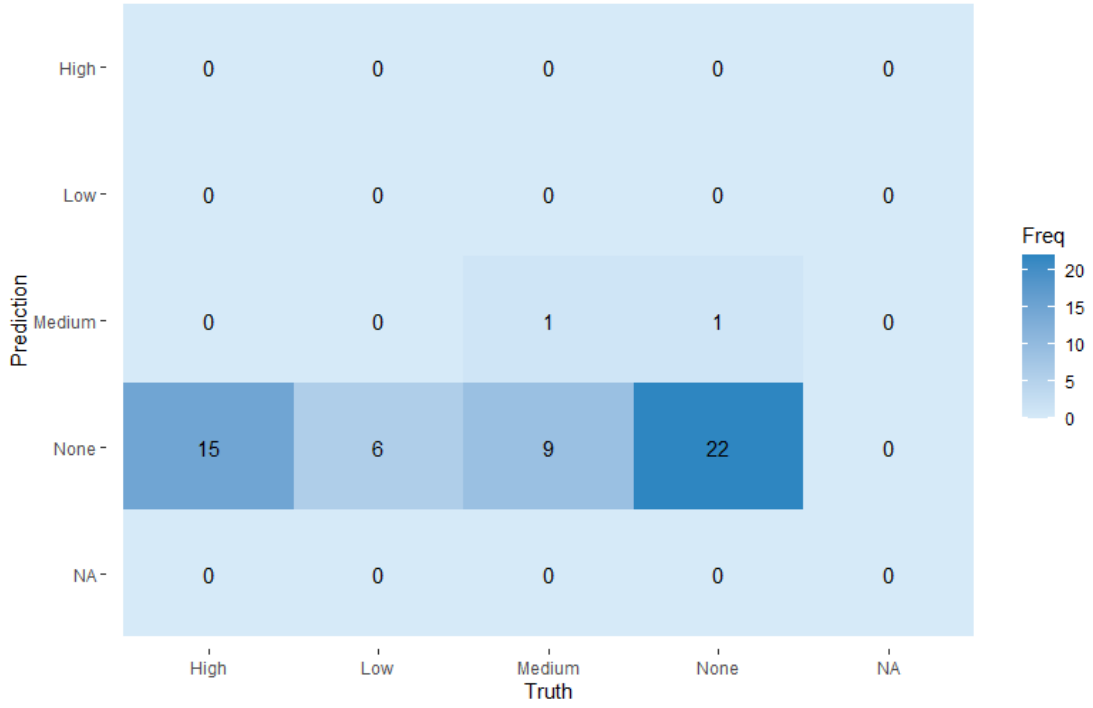

**Εικόνα Παραρτήματος 11. Η δομή του δικτύου ως αποτέλεσμα του constrained based αλγόριθμου iamb.**



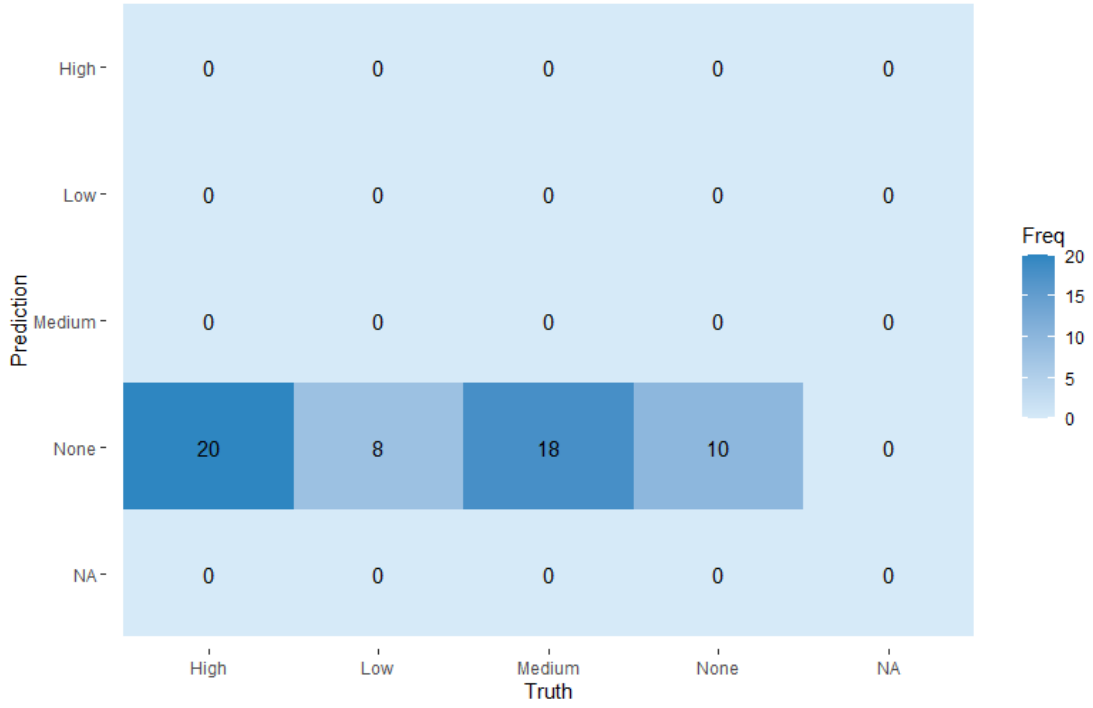

**Εικόνα Παραρτήματος 12. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο iamb, εύρεση παραμέτρων με τη μέθοδο bayes και πρόβλεψη με τη μέθοδο bayes lw.**




**Εικόνα Παραρτήματος 13. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο iamb, εύρεση παραμέτρων με τη μέθοδο bayes και πρόβλεψη με τη μέθοδο parents.**

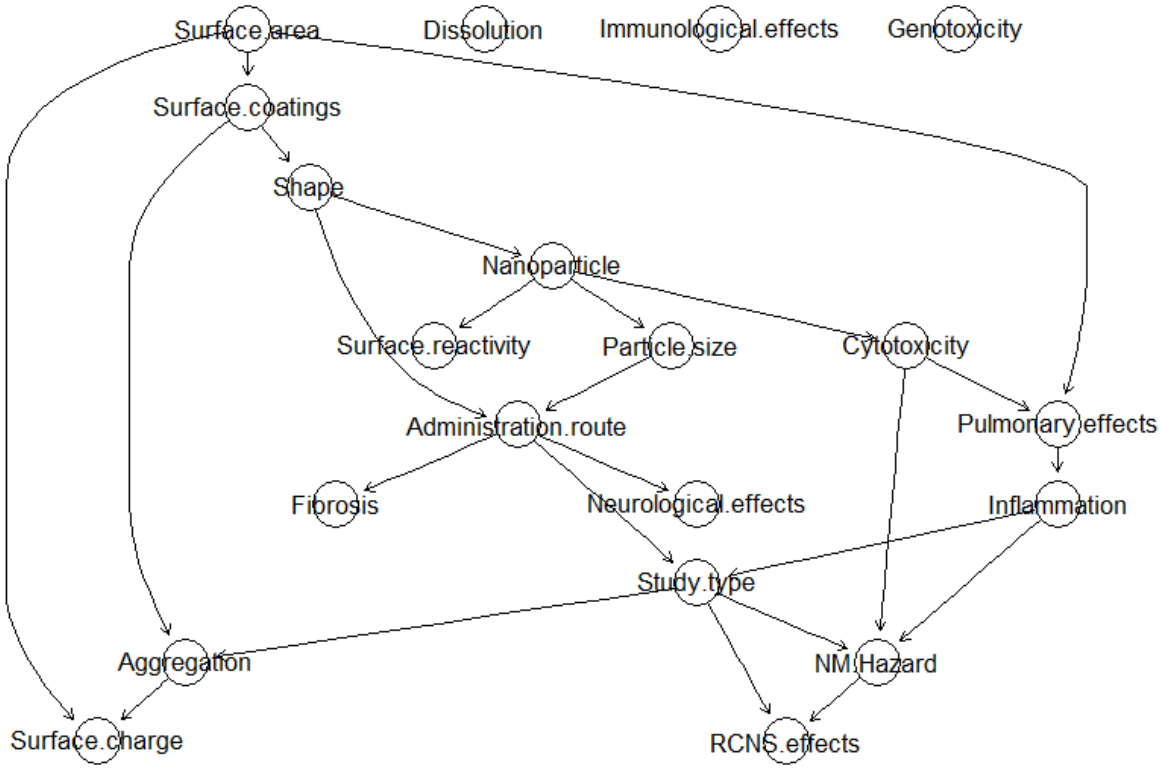



**Εικόνα Παραρτήματος 14. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο iamb, εύρεση παραμέτρων με τη μέθοδο mle και πρόβλεψη με τη μέθοδο bayes lw.**

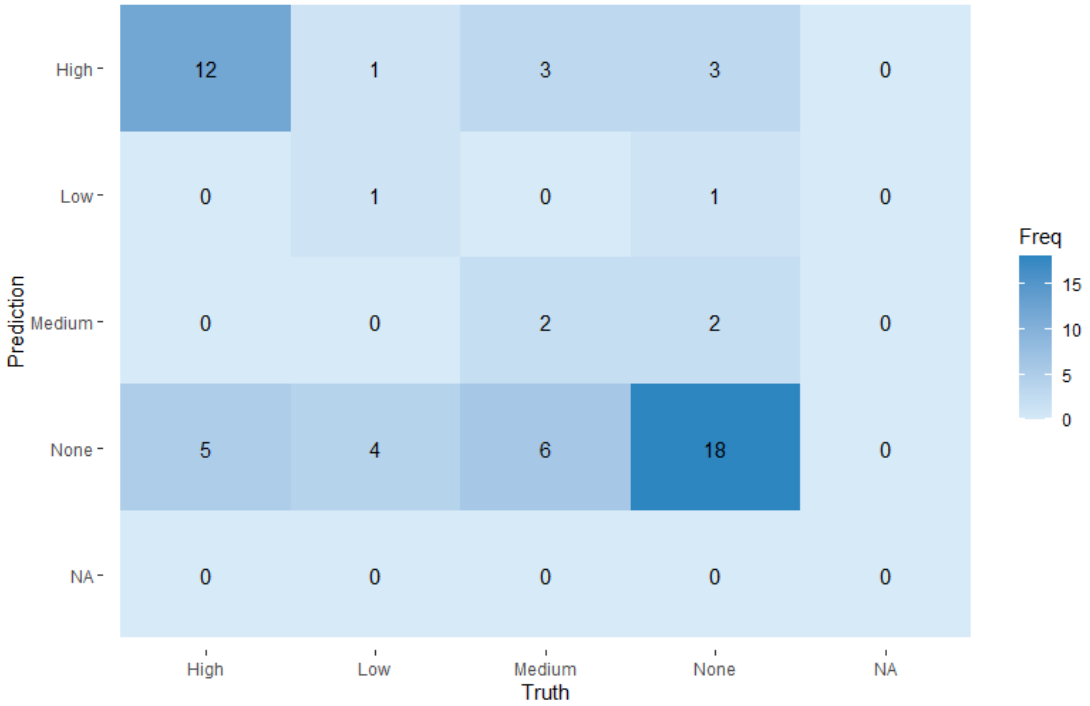



**Εικόνα Παραρτήματος 15. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο iamb, εύρεση παραμέτρων με τη μέθοδο mle και πρόβλεψη με τη μέθοδο parents.**

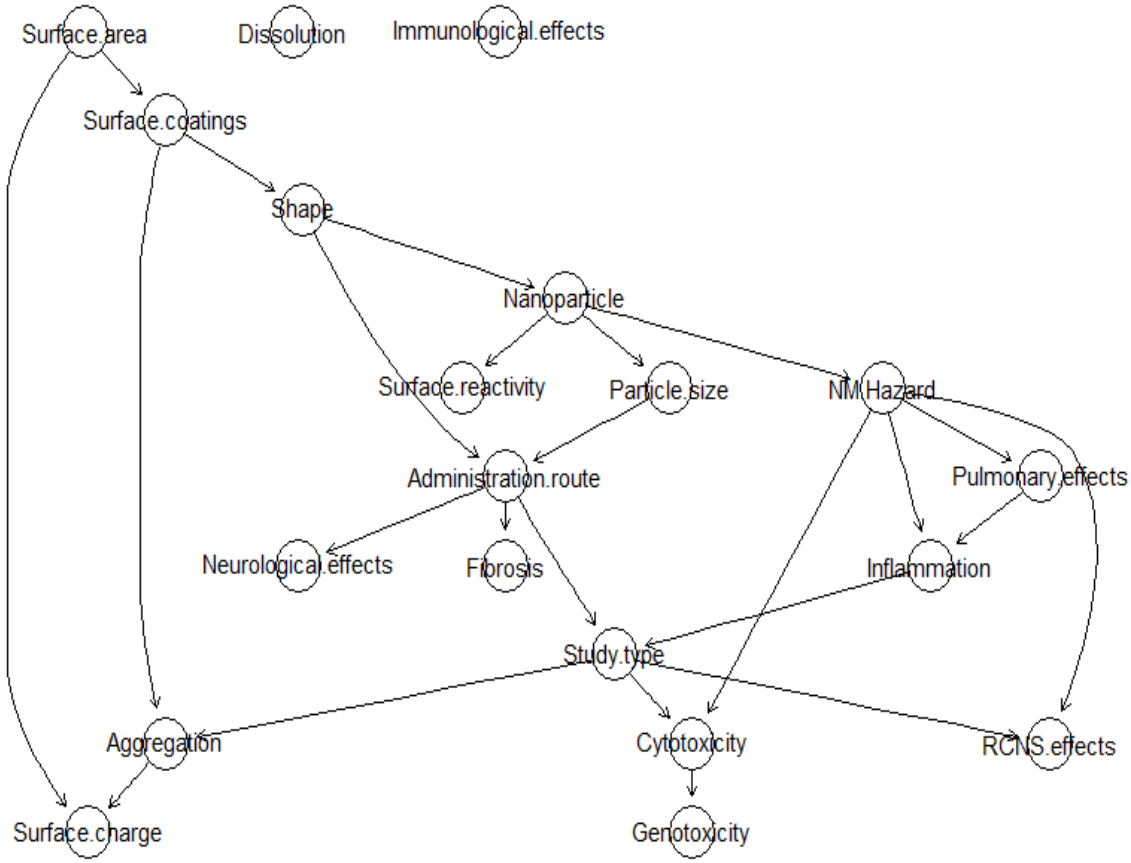
*3η Περίπτωση*: Κατασκευή δομής από τα δεδομένα χρησιμοποιώντας score based αλγόριθμους, τον δομικό αλγόριθμο EM και με εκκίνηση μια τυχαία δομή δικτύου.



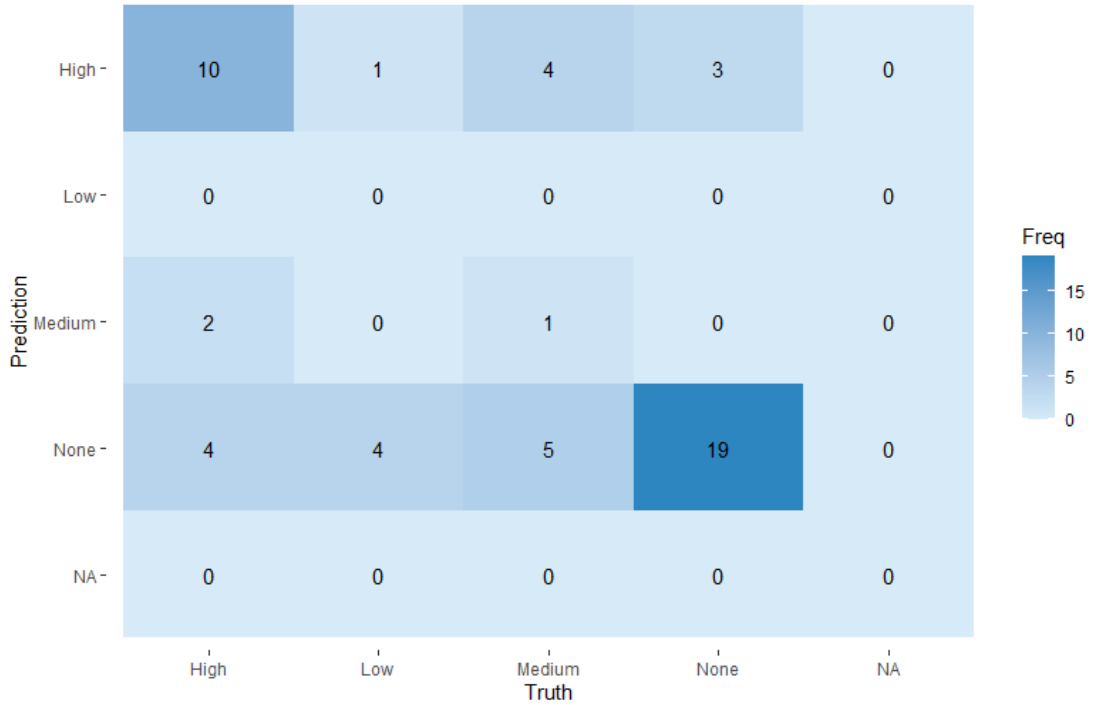

**Εικόνα Παραρτήματος 16. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης aic με πρόβλεψη χρησιμοποιώντας bayes lw.**



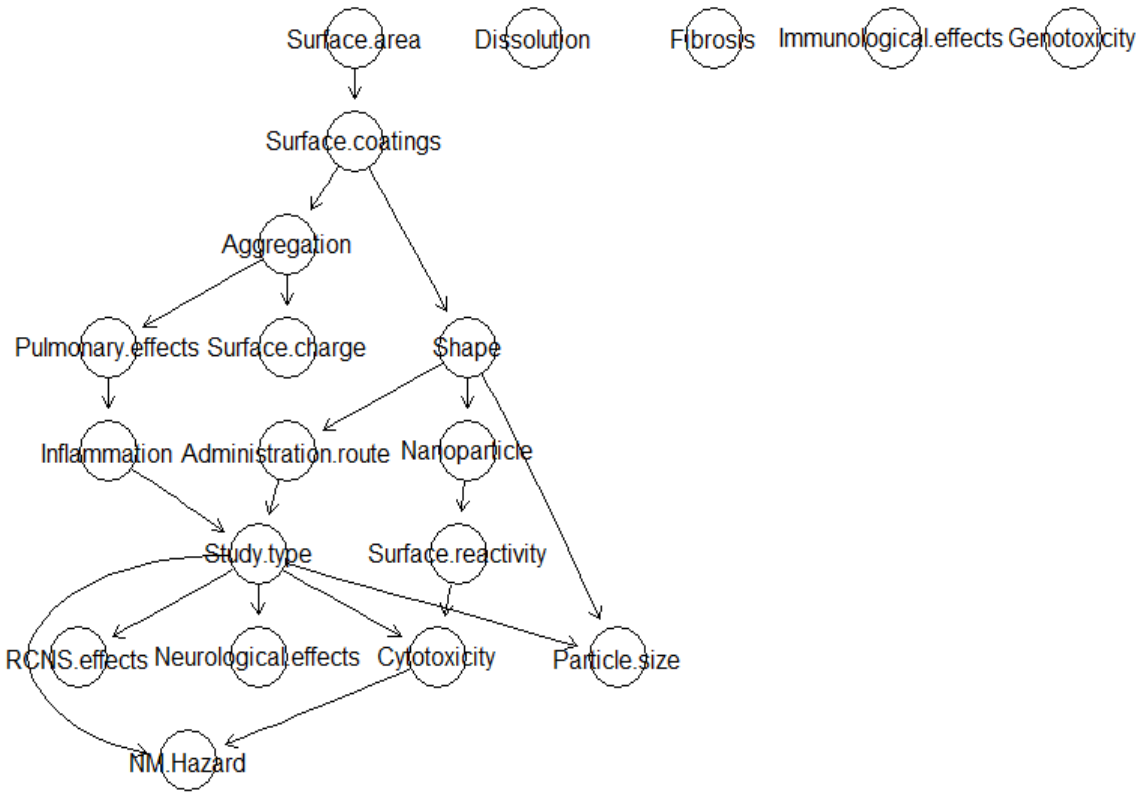

**Εικόνα Παραρτήματος 17. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης aic και πρόβλεψη με τη μέθοδο bayes lw.**



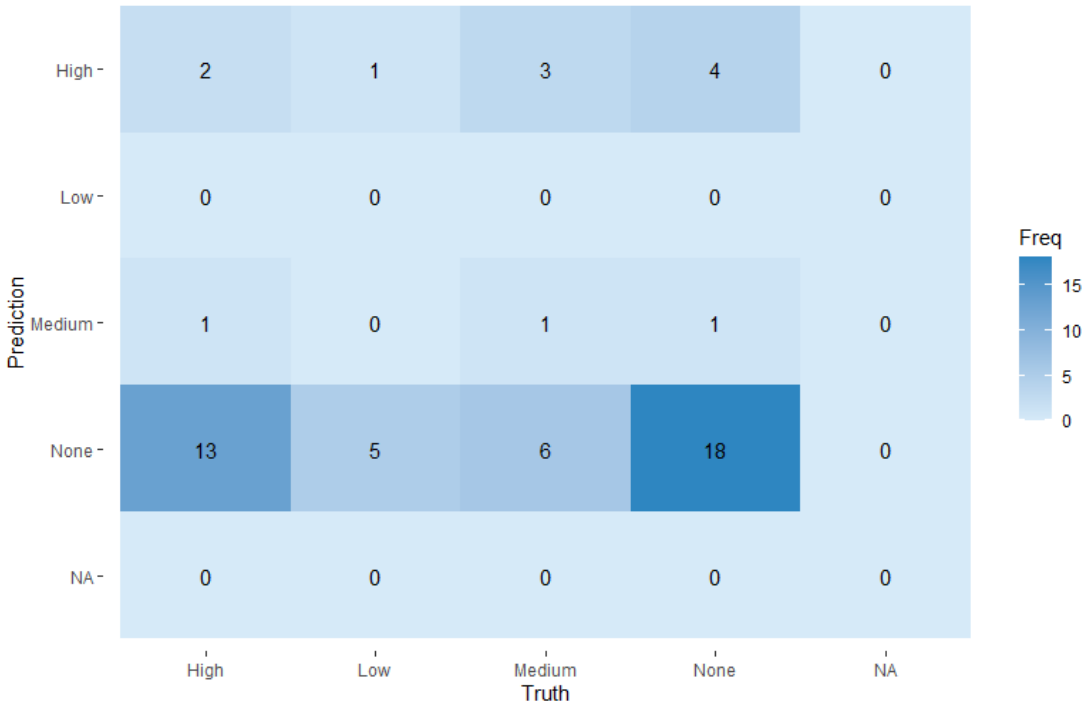

**Εικόνα Παραρτήματος 18. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης aic με πρόβλεψη χρησιμοποιώντας parents.**



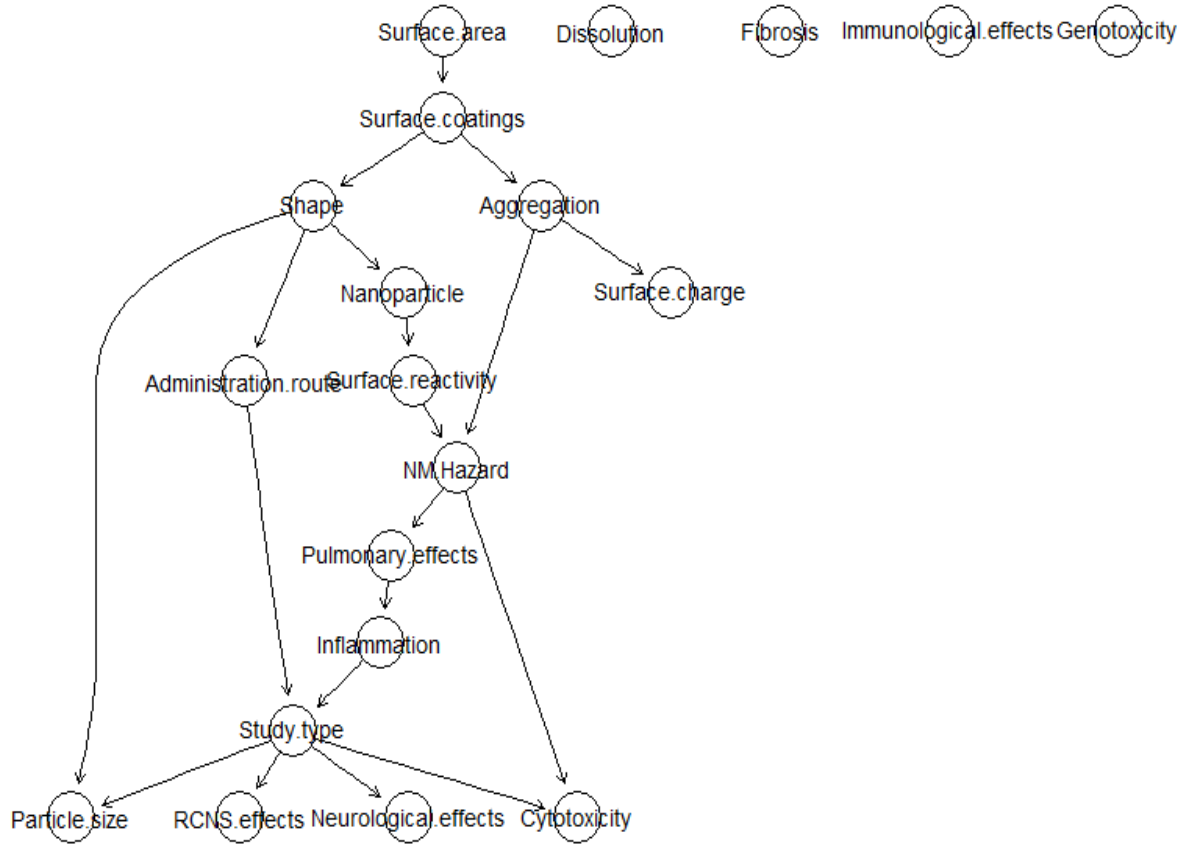

**Εικόνα Παραρτήματος 19. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης aic και πρόβλεψη με τη μέθοδο parents.**




**Εικόνα Παραρτήματος 20. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας bayes lw.**

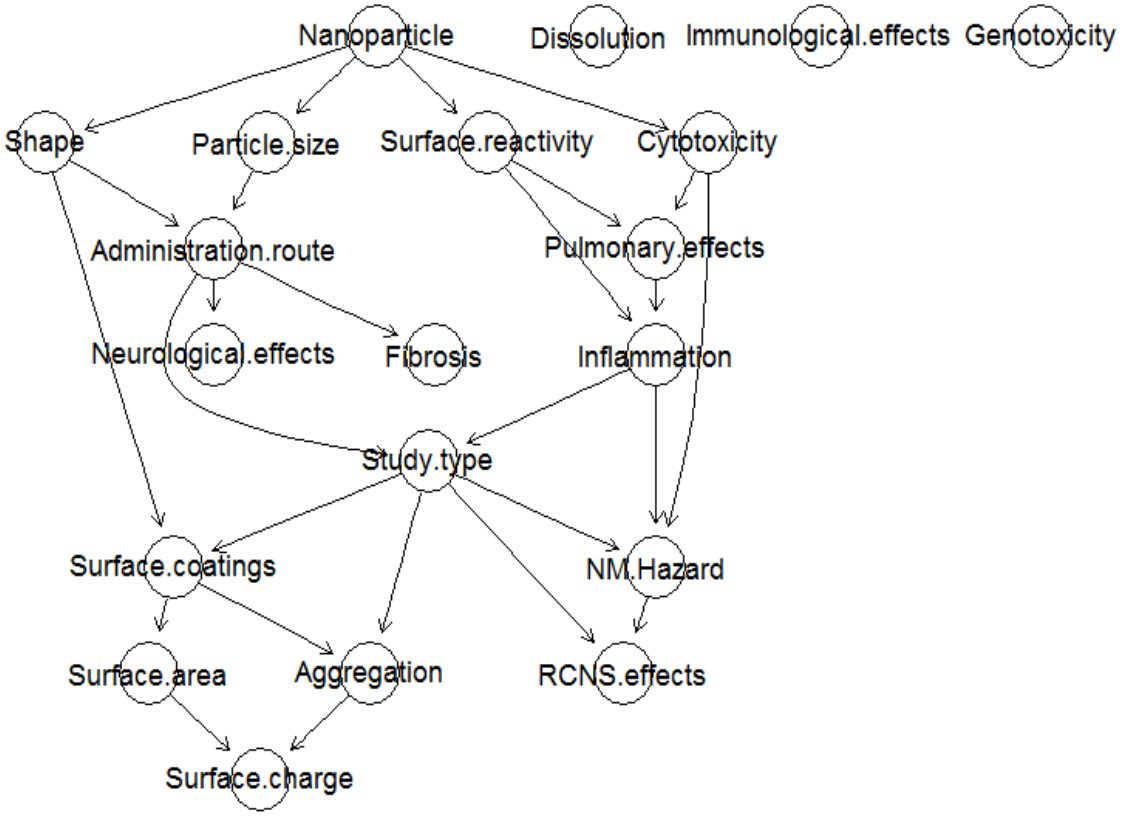



**Εικόνα Παραρτήματος 21. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο bayes lw.**

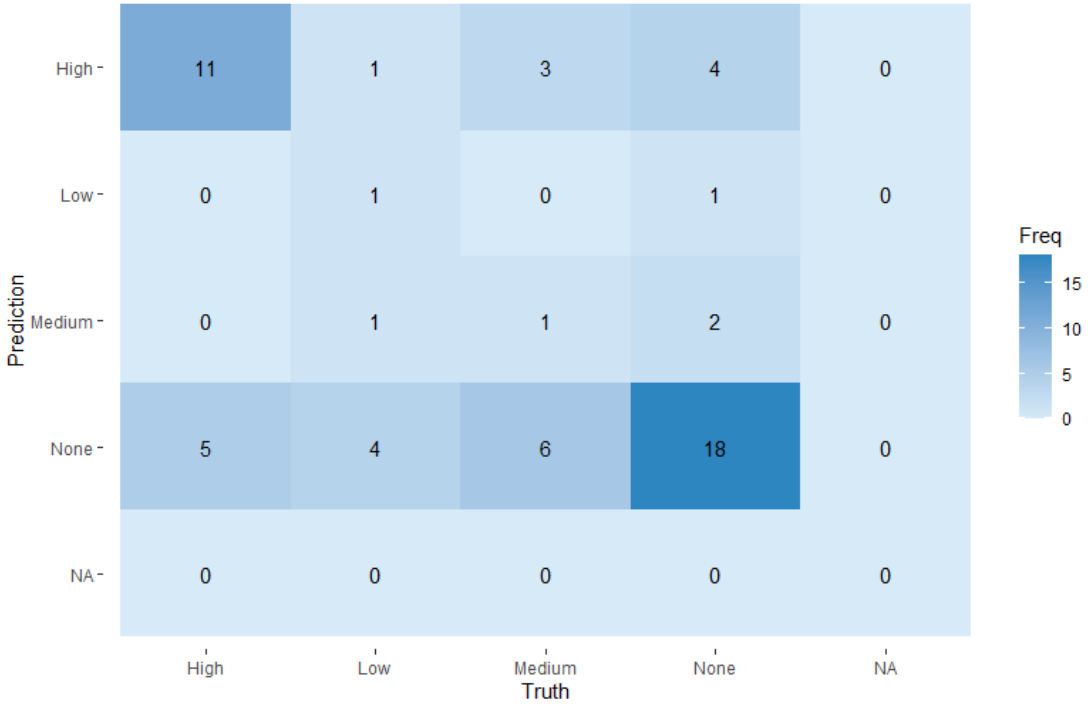



**Εικόνα Παραρτήματος 22. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας parents.**

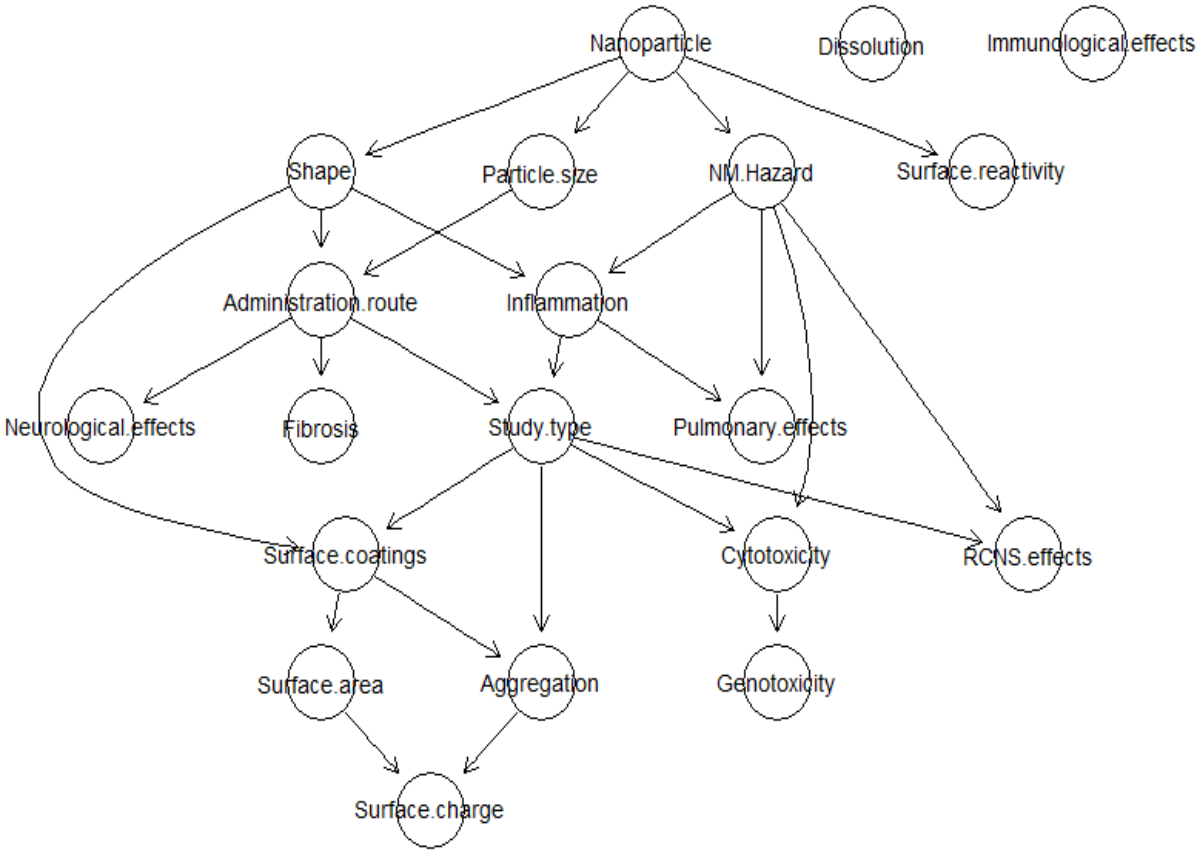

**Εικόνα Παραρτήματος 23. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο parents.**




**Εικόνα Παραρτήματος 24. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου tabu και της συνάρτησης βαθμολόγησης aic με πρόβλεψη χρησιμοποιώντας bayes lw.**

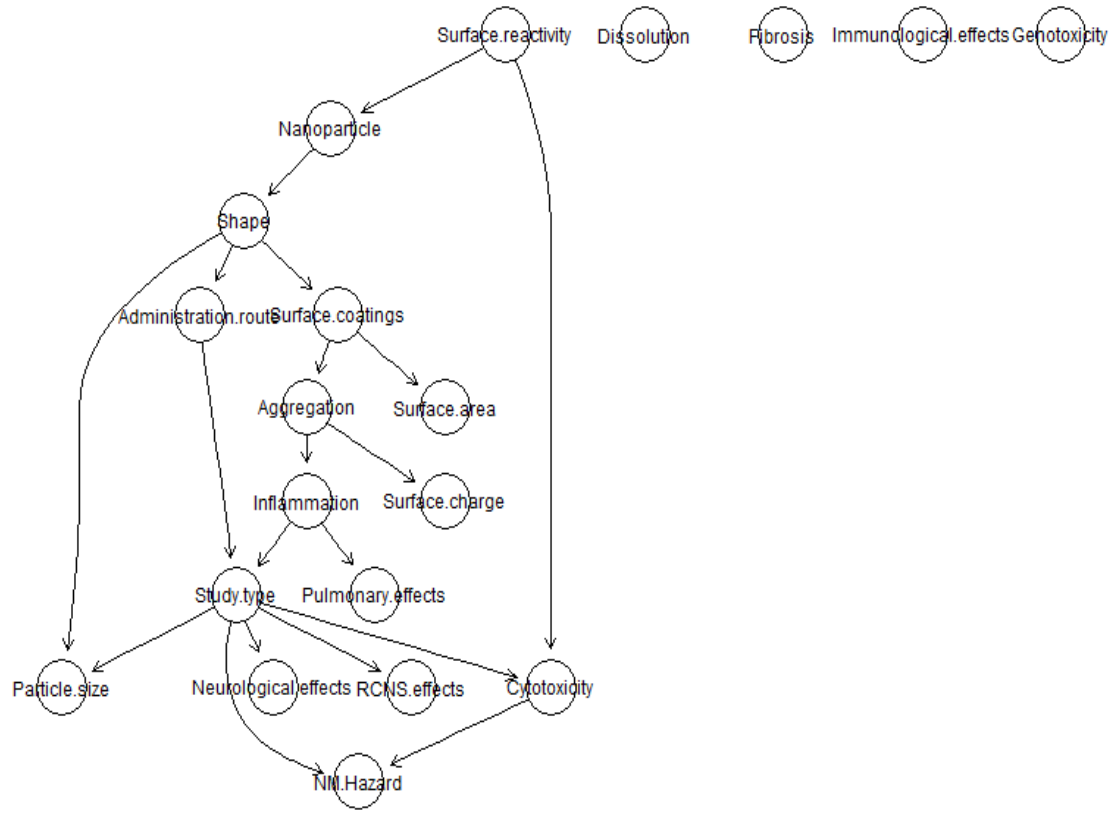



**Εικόνα Παραρτήματος 25. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο tabu, συνάρτηση βαθμολόγησης aic και πρόβλεψη με τη μέθοδο bayes-lw.**

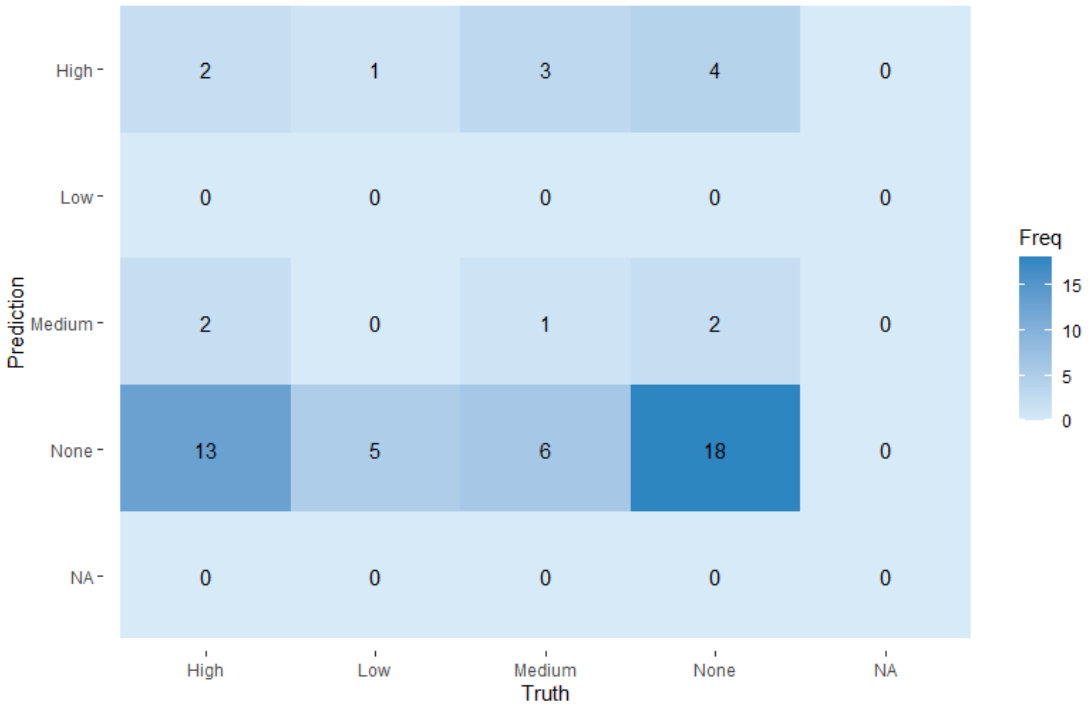



**Εικόνα Παραρτήματος 26. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου tabu και της συνάρτησης βαθμολόγησης aic με πρόβλεψη χρησιμοποιώντας parents.**

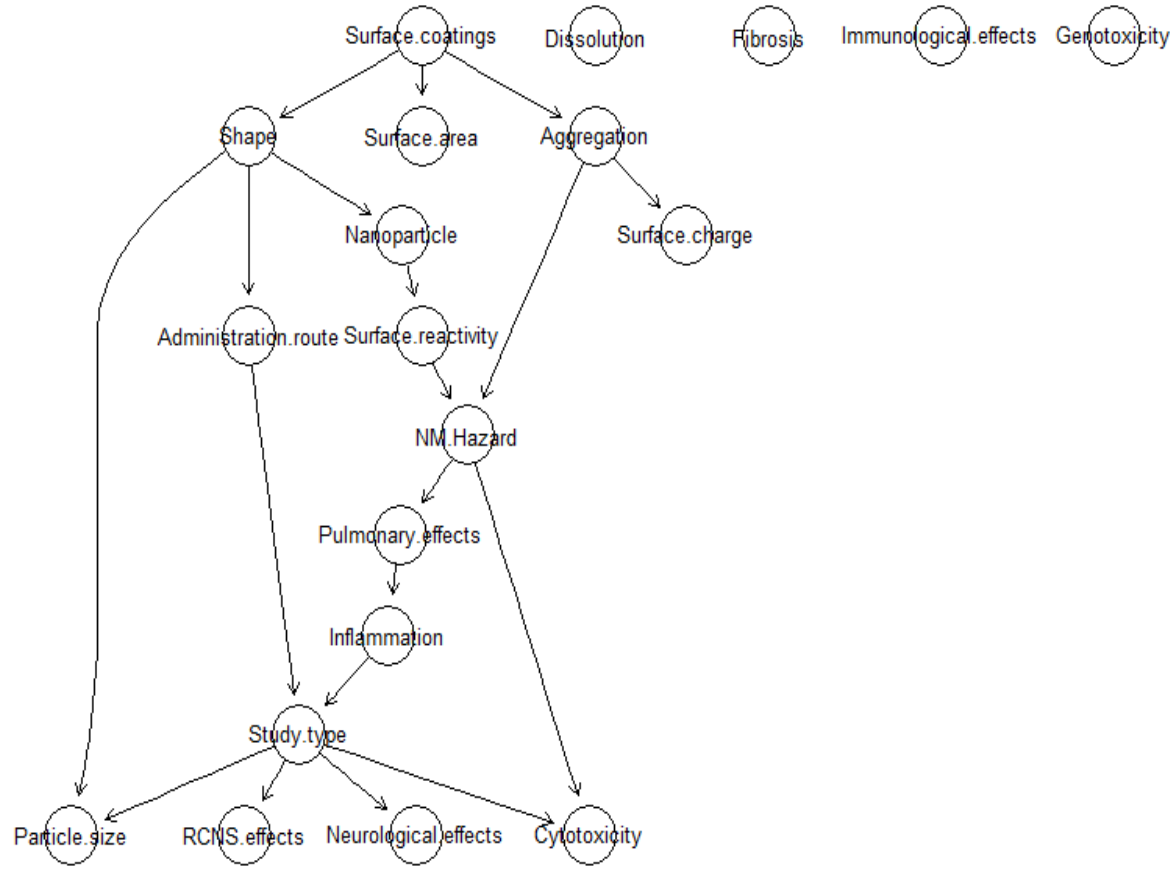

**Εικόνα Παραρτήματος 27. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο tabu, συνάρτηση βαθμολόγησης aic και πρόβλεψη με τη μέθοδο parents.**



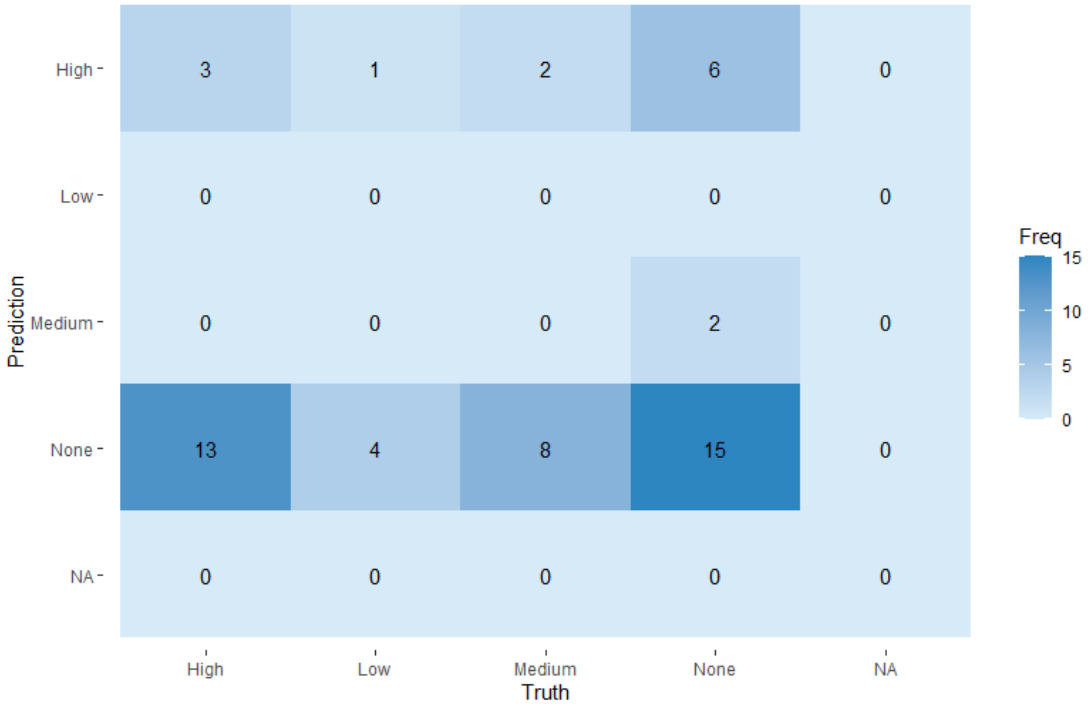

**Εικόνα Παραρτήματος 28. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου tabu και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας bayes-lw.**




**Εικόνα Παραρτήματος 29. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο tabu, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο bayes lw.**

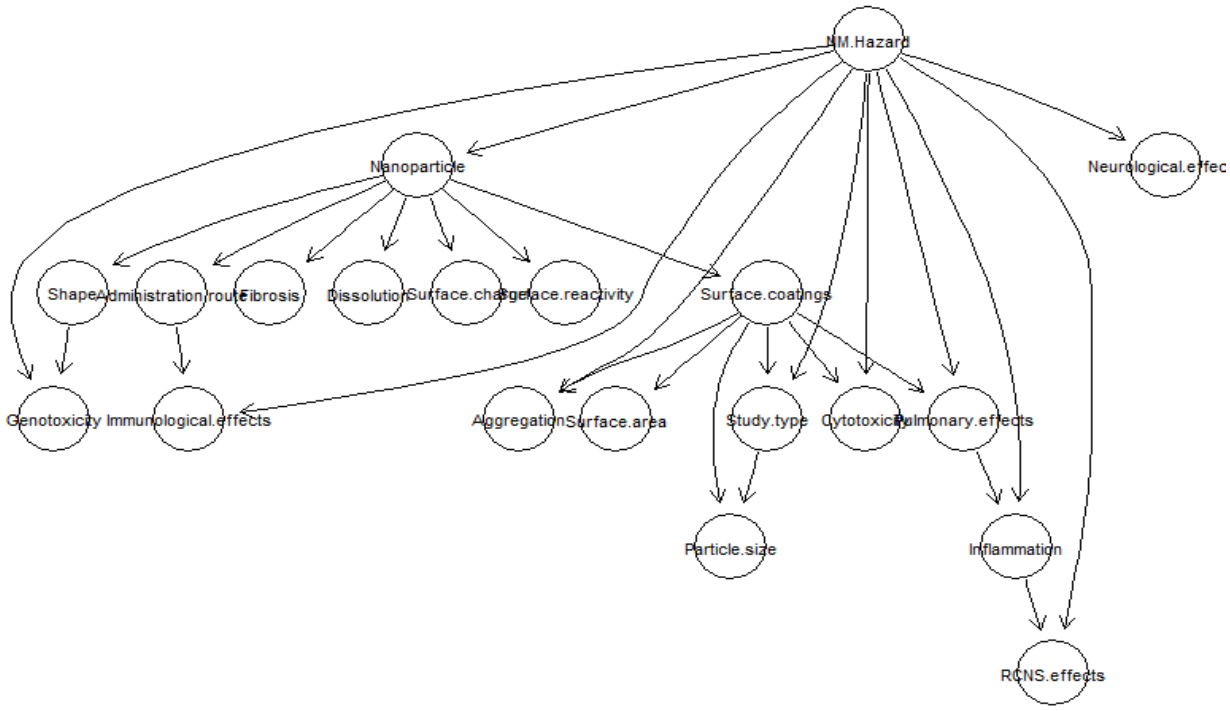



**Εικόνα Παραρτήματος 30. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου tabu και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας parents.**



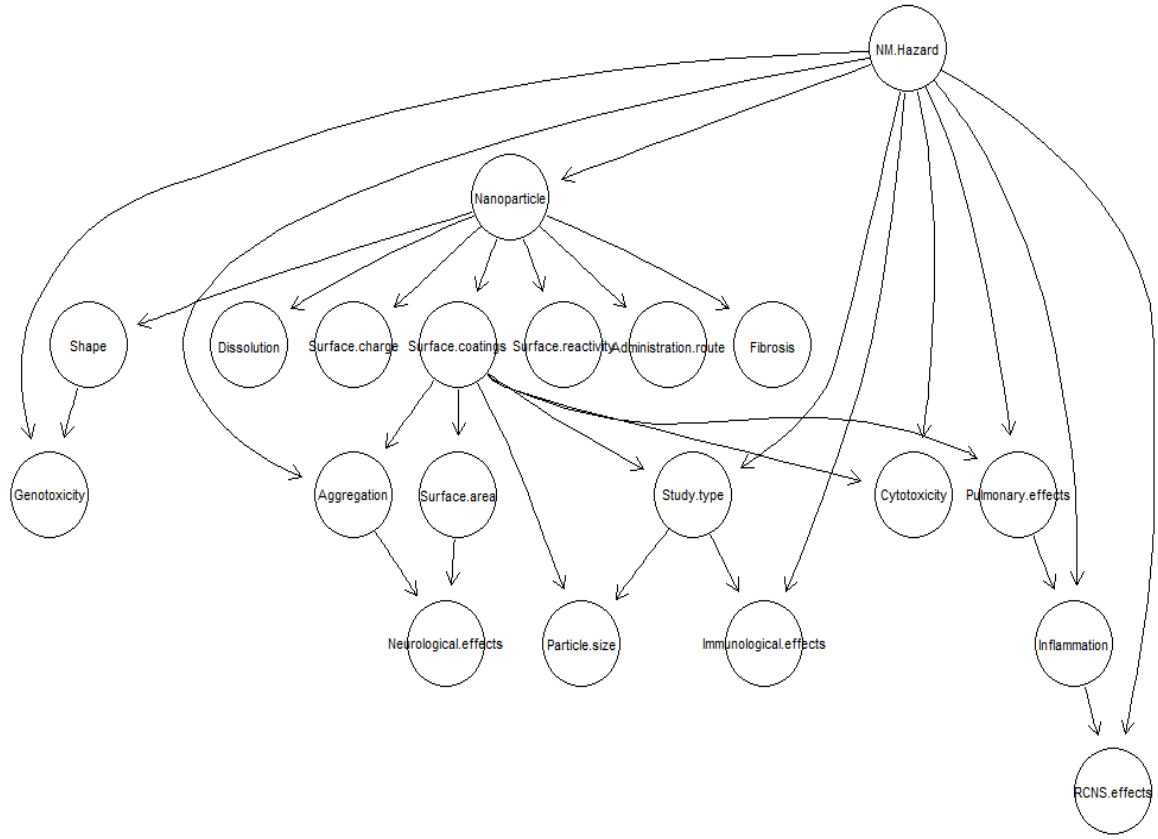

**Εικόνα Παραρτήματος 31. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο tabu, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο parents.**

*4η Περίπτωση*: Κατασκευή δομής από τα δεδομένα χρησιμοποιώντας score based αλγόριθμους, τον δομικό αλγόριθμο EM και με εκκίνηση εκ των προτέρων γνωστή δομή δικτύου εκείνη των Marvin et al. (2017).

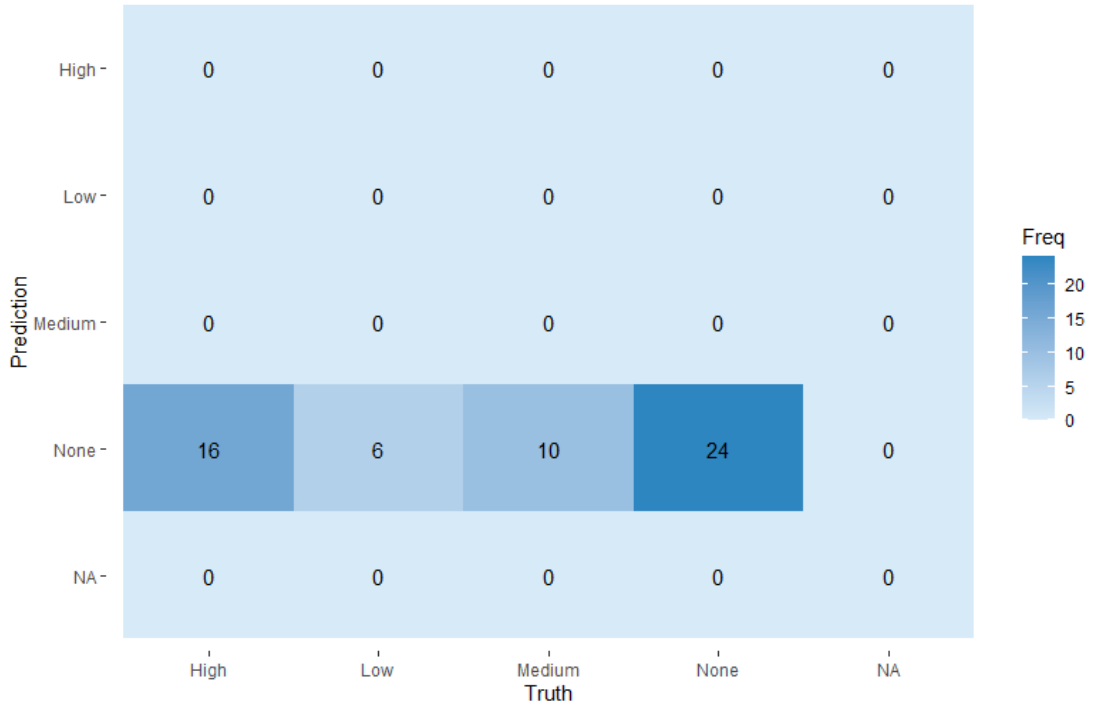



**Εικόνα Παραρτήματος 32. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης aic με πρόβλεψη χρησιμοποιώντας bayes-lw.**

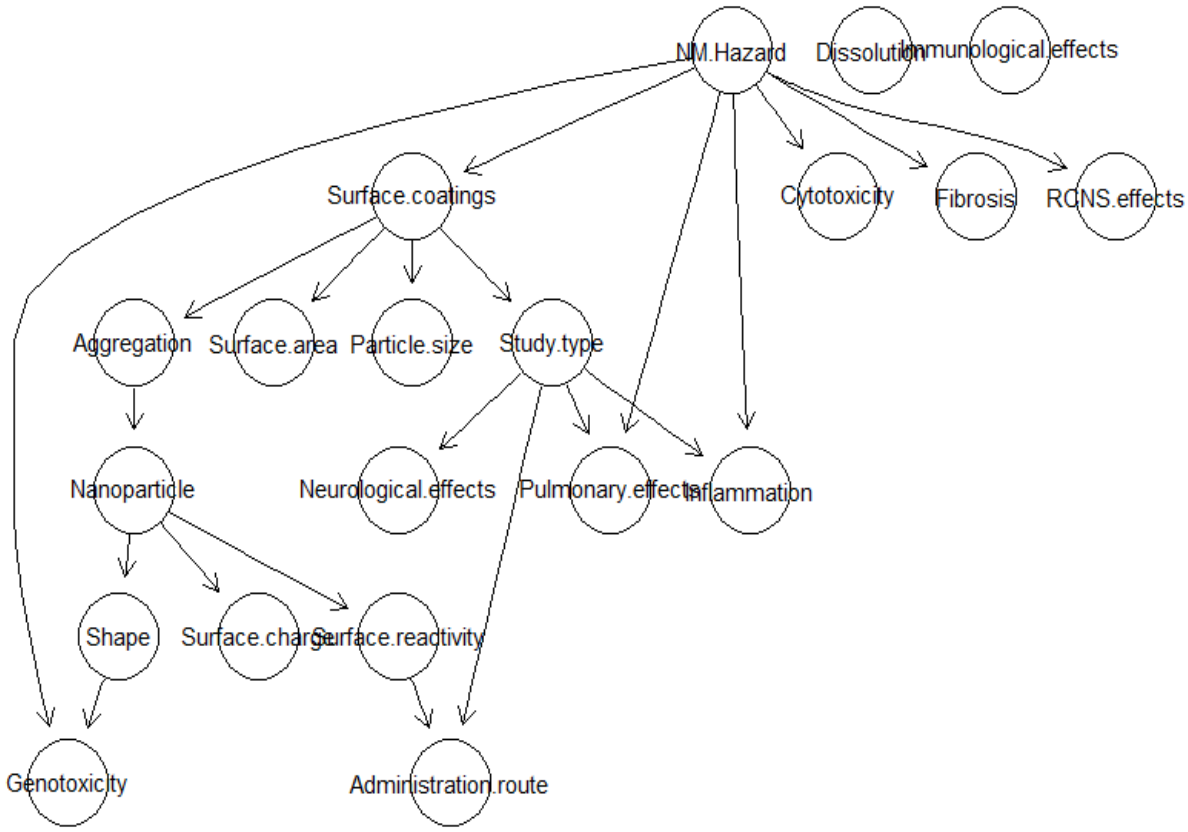

**Εικόνα Παραρτήματος 33. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης aic και πρόβλεψη με τη μέθοδο bayes-lw.**



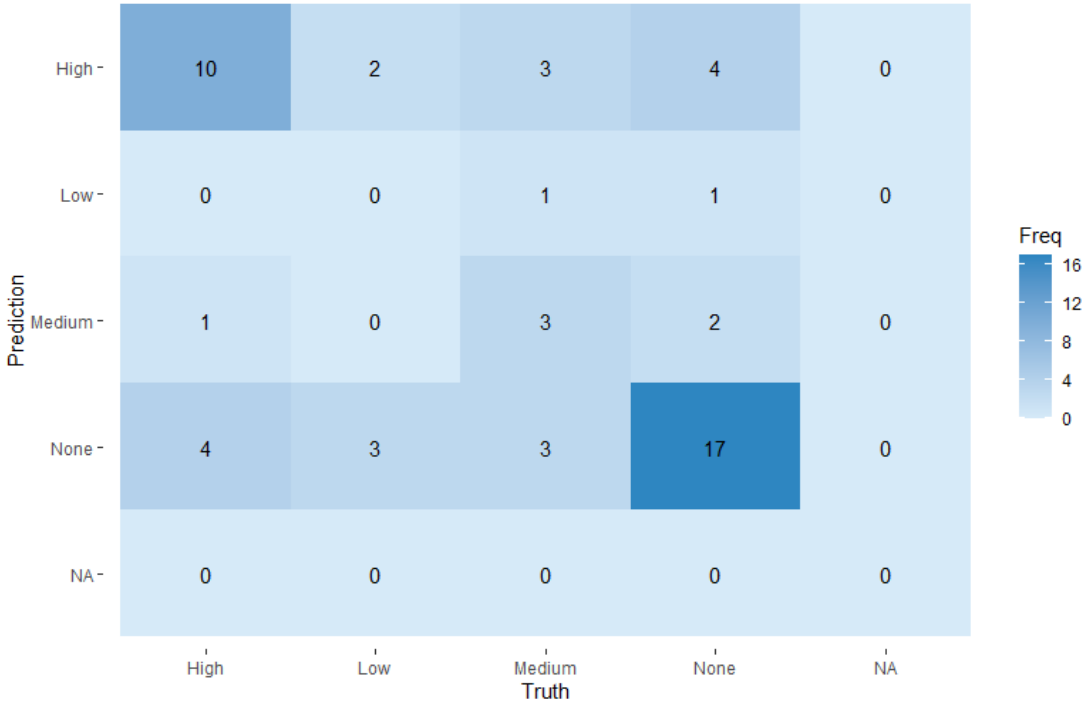

**Εικόνα Παραρτήματος 34. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης aic με πρόβλεψη χρησιμοποιώντας parents.**



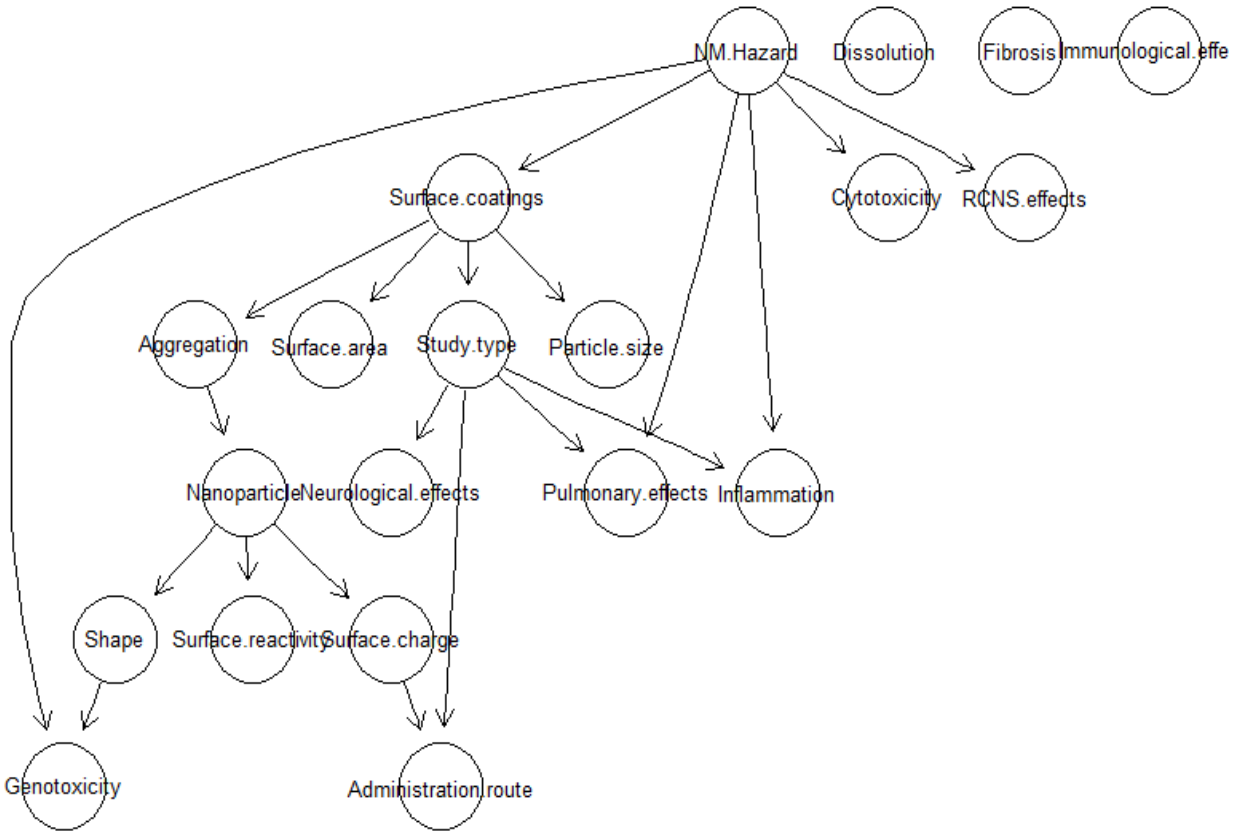

**Εικόνα Παραρτήματος 35. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης aic και πρόβλεψη με τη μέθοδο parents.**



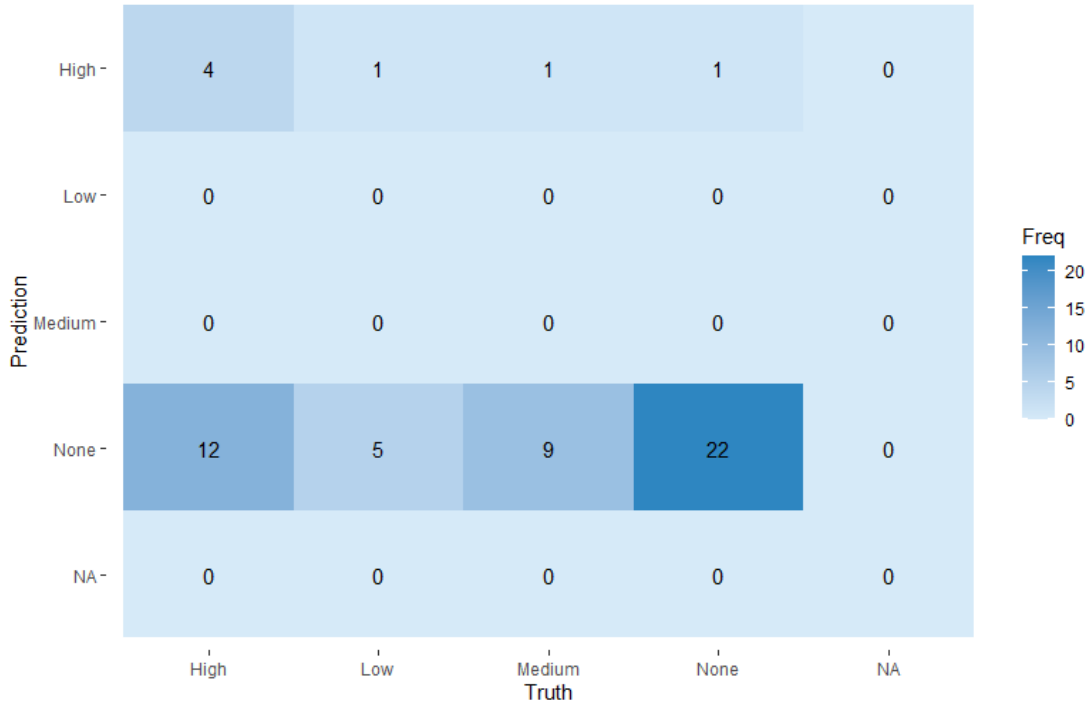

**Εικόνα Παραρτήματος 36. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας bayes-lw.**



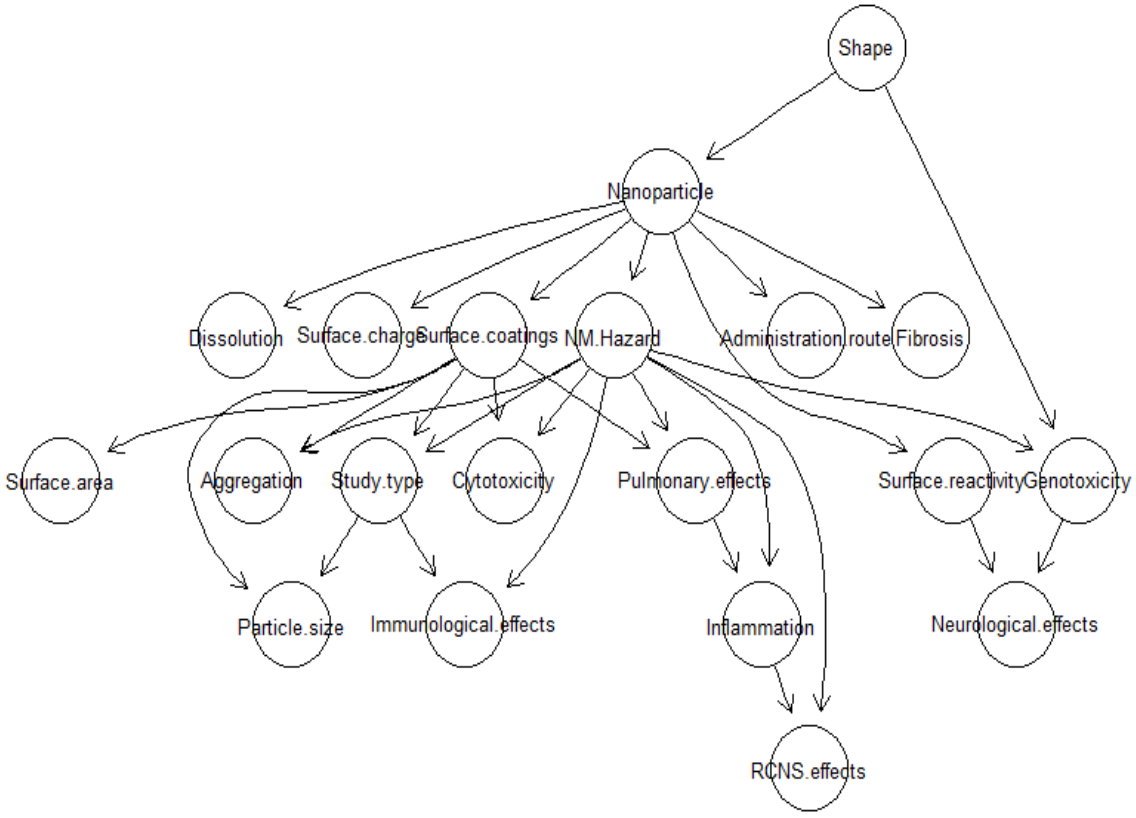

**Εικόνα Παραρτήματος 37. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο bayes-lw.**



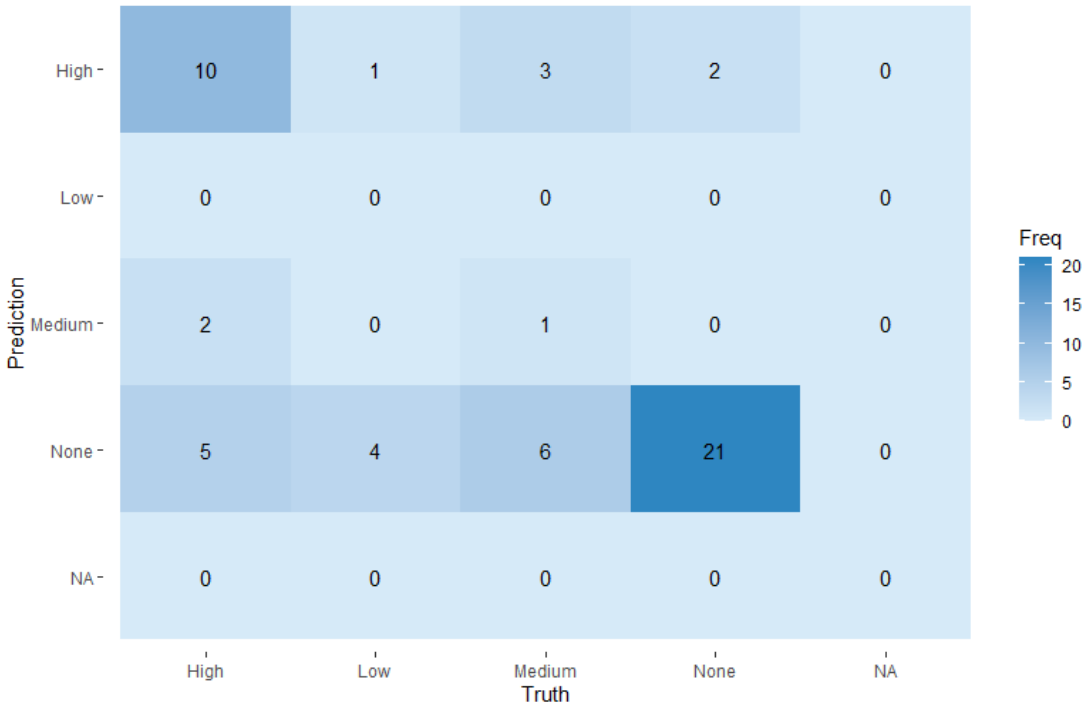

**Εικόνα Παραρτήματος 38. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου hc και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας parents.**



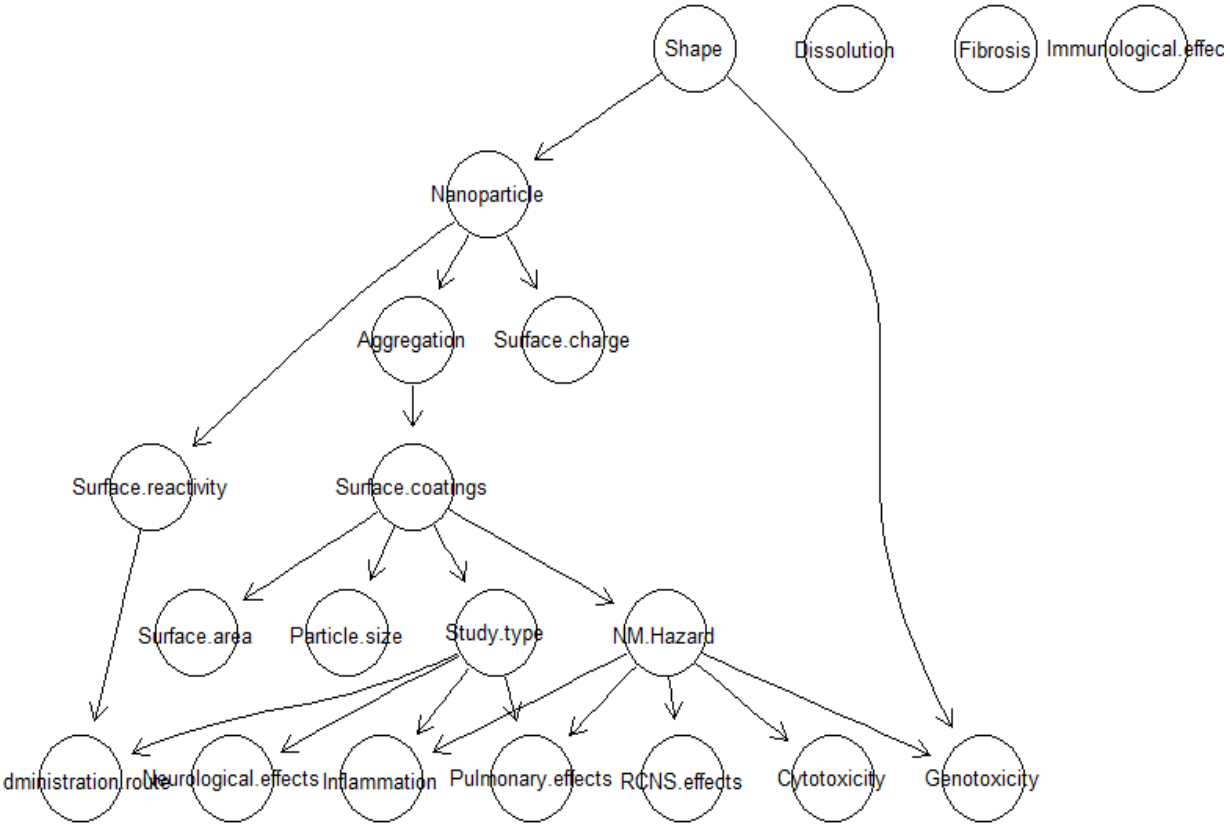

**Εικόνα Παραρτήματος 39. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο hc, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο parents.**



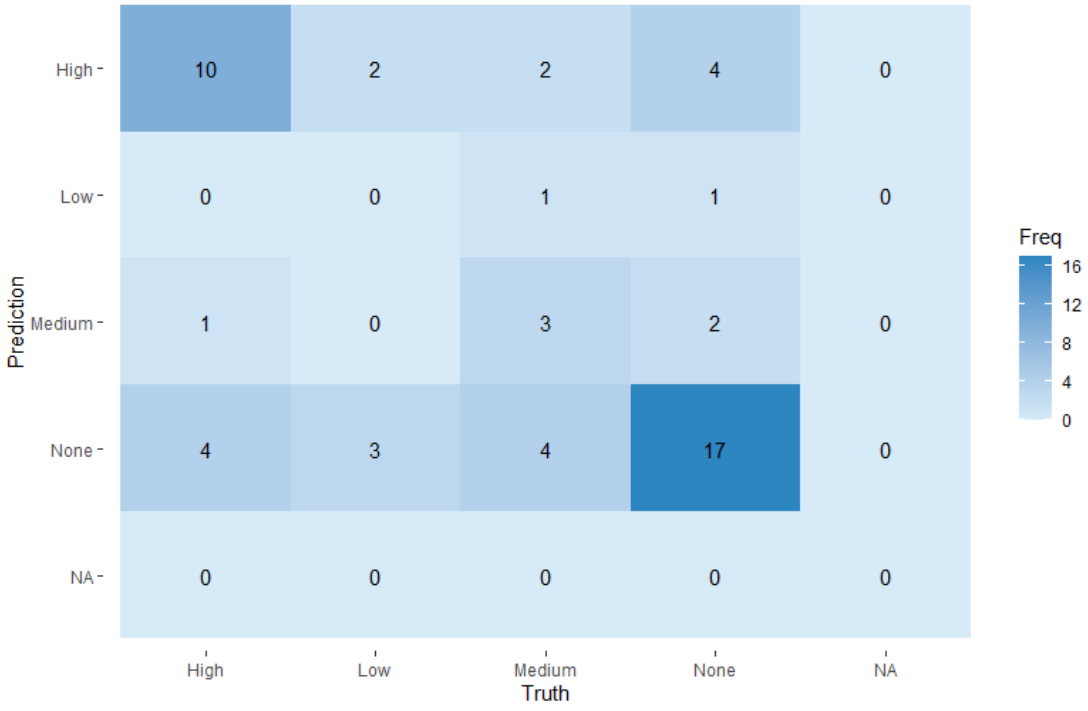

**Εικόνα Παραρτήματος 40. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου tabu και της συνάρτησης βαθμολόγησης aic με πρόβλεψη χρησιμοποιώντας parents.**



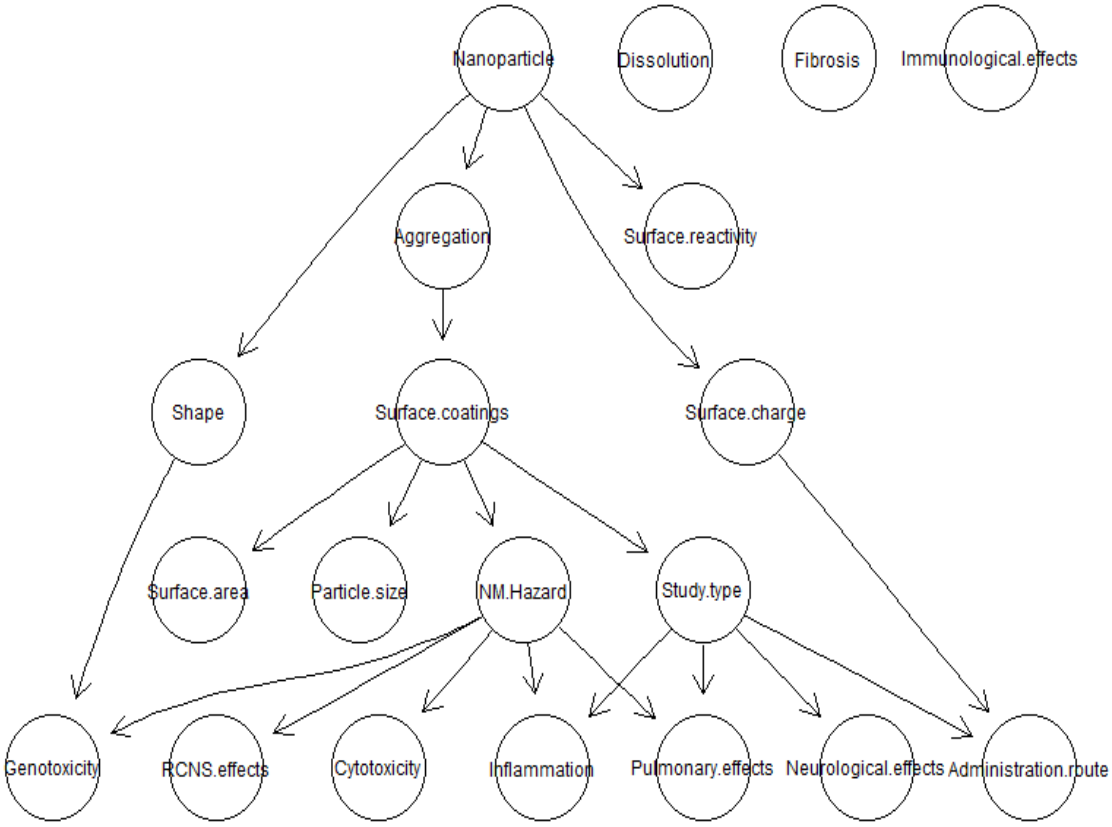

**Εικόνα Παραρτήματος 41. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο tabu, συνάρτηση βαθμολόγησης aic και πρόβλεψη με τη μέθοδο parents.**



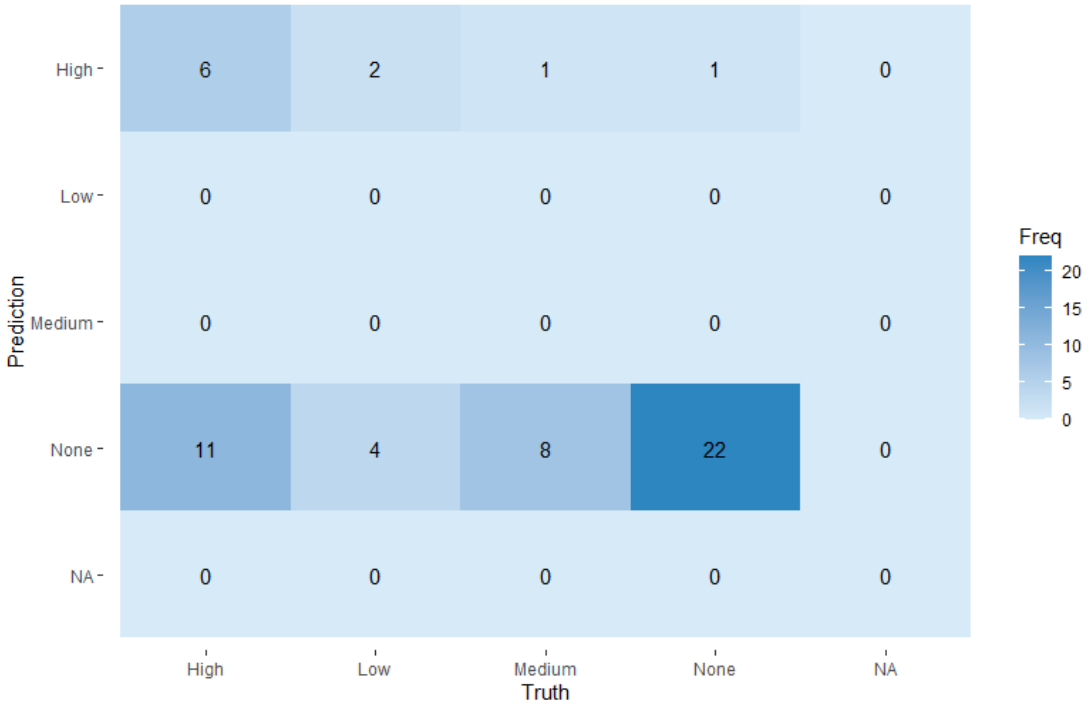

**Εικόνα Παραρτήματος 42. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου tabu και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας bayes lw.**




**Εικόνα Παραρτήματος 43. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο tabu, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο bayes lw.**




**Εικόνα Παραρτήματος 44. Η δομή του δικτύου ως αποτέλεσμα της ευρετικής μεθόδου tabu και της συνάρτησης βαθμολόγησης bic με πρόβλεψη χρησιμοποιώντας parents.**




**Εικόνα Παραρτήματος 45. Πίνακας σύγχυσης για εύρεση δομής δικτύου με τη μέθοδο tabu, συνάρτηση βαθμολόγησης bic και πρόβλεψη με τη μέθοδο parents.**