Νέες Μπεϋσιανές τεχνικές ομαδοποίησης με εφαρμογή στην αυτόματη δεικτοδότηση ομιλητών σε αρχεία ήχου

Θέμος Σταφυλάκης

Εθνικό Μετσόβιο Πολυτεχνείο

Σχολή Ηλεκτρολόγων Μηχανικών και Μηχανικών Υπολογιστών

Τομέας Σημάτων, Ελέγχου και Ρομποτικής

Κείμενο διδακτορικής διατριβής

Επιβλέπων καθηγητής:

καθ. Γ. Καραγιάννης

Μέλη επταμελούς επιτροπής:

καθ. Π. Μαραγκός

καθ. Γ. Α. Σταφυλοπάτης

καθ. Γ. Καμπουράκης

καθ. Τ. Σελλής

καθ. Β. Μέρτζιος

ερευν. Β. Κατσούρος

ΠΡΟΛΟΓΟΣ ΚΑΙ ΣΥΝΕΙΣΦΟΡΕΣ ΤΗΣ ΔΙΑΤΡΙΒΗΣ

Η παρούσα διατριβή αφορά στο πρόβλημα της κατάτμησης και ομαδοποίησης αρχείων ομιλίας σε ομιλητές, πρόβλημα το οποίο απαντάται στη διεθνή βιβλιογραφία με τον όρο *speaker diarization*. Είναι ένα πρόβλημα κομβικό, καθώς πολλές εφαρμογές επεξεργασίας ανθρώπινης φωνής απαιτούν μία τέτοια βαθμίδα ως στάδιο προεπεξεργασίας. Θέλουν δηλαδή έναν μηχανισμό ο οποίος να είναι σε θέση αξιόπιστα και μέσα σε ένα εύλογο χρονικό διάστημα να εκτιμήσει α) πόσοι είναι οι συμμετέχοντες ομιλητές και β) σε ποιές χρονικές περιόδους μιλάει ο καθένας.

Το πρόβλημα αυτό έχει ένα ιδιαίτερο χαρακτηριστικό που το καθιστά ταυτόχρονα δυσεπίλυτο και ελκυστικό, καθώς δεν υπάρχει καμία εκ των προτέρων πληροφορία όσον αφορά στον αριθμό και την ταυτότητα των ομιλητών. Επαφίεται έτσι στον αλγόριθμο να εκτιμήσει τα μοντέλα των ομιλητών και τον αριθμό τους, ομαδοποιώντας κατάλληλα τα διανύσματα χαρακτηριστικών που αποτελούν το αρχείο. Ο τομέας της μάθησης μηχανών με στατιστικές τεχνικές (statistical machine learning) έχει αναπτύξει πλήθος αλγόριθμων μη-επιβλεπόμενης ομαδοποίησης. Ωστόσο, οι περισσότεροι εξ αυτών απαιτούν a priori γνώση του αριθμού των ομάδων. Εδώ ακριβώς έγκειται και η ελκυστικότητα του προβλήματος. Ο αλγόριθμος πρέπει να λειτουργήσει τυφλά.

Στην παρούσα διατριβή αρχικά παρουσιάζουμε υπάρχουσες μεθόδους και τεχνικές που έχουν κατά καιρούς προταθεί και χρησιμοποιούνται στην πράξη. Κατόπιν, διατυπώνουμε τις δικές μας προσεγγίσεις και βελτιώσεις. Κατά τη διάρκεια αυτής της διατριβής ασχοληθήκαμε με μία σειρά από θέματα που κείνται είτε στην καρδιά είτε στην περιφέρεια του προβλήματος. Εξερευνήσαμε εναλλακτικές μοντελοποιήσεις πολλές από τις οποίες δανειστήκαμε από άλλους τομείς της αναγνώρισης προτύπων, όπως της επεξεργασίας εικόνας, βίντεο και φυσικής γλώσσας. Τρεις είναι κατά βάση οι προτάσεις και εξελίξεις που προκύπτουν από αυτή τη διατριβή.

α) Η πρώτη σχετίζεται με την ανάπτυξη μίας πιθανοτικής απόστασης μεταξύ τμημάτων ομιλίας, η οποία συνδυάζει δυαδικούς ταξινομητές και ροές πληροφορίας. Ένα ιδιαίτερο χαρακτηριστικό της μεθόδου είναι η κατάτμηση του χώρου εισόδου και η εκπαίδευση ενός μοντέλου για κάθε κατηγορία, έτσι ώστε η απόφαση να προκύπτει ως πιθανοτικός συνδυασμός των αποκρίσεων καθε μοντέλου. Τέλος, εξετάζουμε τη χρήση Μπεϋσιανών τεχνικών για να ενισχύσουμε τη στιβαρότητα

στην εκτίμηση των παραμέτρων των μοντέλων. Από το κεφάλαιο αυτό προέκυψε η εργασία μας [1].

β) Η δεύτερη συνεισφορά της διατριβής αφορά στην αναδιατύπωση ενός από τα πλέον θεμελιώδη και πολυχρησιμοποιούμενα κριτήρια ομαδοποίησης ομιλητών, το Μπεϋσιανό Κριτήριο Πληροφορίας (BIC). Εμβαθύνοντας στο μαθηματικό υπόβαθρο του κριτηρίου, αποδεικνύουμε ότι οι και δύο μορφές του (ολική και τοπική) που χρησιμοποιούνται είναι υποβέλτιστες για το πρόβλημα ομαδοποίησης ομιλητών. Χρησιμοποιώντας ως μέσο ανάλυσης τις εκ των προτέρων κατανομές των παραμέτρων τις οποίες το BIC υπονοεί, προτείνουμε μία νέα μορφή του, την τμηματική, η οποία προσφέρει σημαντικότατη αύξηση στην ακρίβεια ομαδοποίησης. Δείχνουμε τέλος ότι η χρήση εκ των προτέρων κατανομών Dirichlet στις πιθανότητες μετάβασης μεταξύ καταστάσεων είναι ικανή να εμπλουτίσει το κριτήριο με χρονική πληροφορία, η οποία μένει ανεκμετάλλευτη με τις δύο τωρινές του μορφές. Από το κεφαλαίο αυτό προέκυψαν οι εργασίες [2], [3], [4] (IEEE Journal), ενώ άμεση συσχέτιση με αυτό έχουν και οι εργασίες μας [5] και [6].

γ) Η τελευταία συνεισφορά της διατριβής είναι η εξερεύνηση του δυνατοτήτων που παρέχει ο αλγορίθμος μετατόπισης του μέσου (mean-shift) στην ομαδοποίηση ομιλητών. Ο συγκεκριμένος αλγόριθμος έχει ήδη επιδείξει σημαντικά αποτελέσματα στον τομέα της επεξεργασίας εικόνας και έχει καθιερωθεί ως μια από τις δημοφιλέστερες μεθόδους μη-παραμετρικής επεξεργασίας. Δείχνουμε ότι ο συγκεκριμένος αλγόριθμος μπορεί να εφαρμοσθεί σε ευρύτερα προβλήματα ομαδοποίησης, όπου οι προς ομαδοποίηση οντότητες ανήκουν σε μη-Ευκλείδιους χώρους, όπως αυτοί των παραμέτρων στατιστικών μοντέλων και συγκεκριμένα Εκθετικών κατανομών. Κάνοντας εκτενή χρήση της Γεωμετρίας της Πληροφορίας (Information Geometry) προσαρμόζουμε κατάλληλα τον αλγόριθμο και αποδεικνύουμε ότι είναι σε θέση να υπερβεί σε ακρίβεια ομαδοποίησης την καθιερωμένη προσέγγιση της ιεραρχικής ομαδοποίησης. Από το κεφάλαιο αυτό προέκυψε η εργασία μας [7], ενώ η πλήρης μέθοδος που παρουσιάζουμε έχει αποσταλεί για δημοσίευση σε διεθνές περιοδικό της μάθησης μηχανών.

Επιπλέον, για τις ανάγκες της διατριβής έγινε υλοποίηση του πλήρους σεναρίου προεπεξεργασίας - κατάτμησης - ομαδοποίησης με την οποία δομικάστηκαν οι διάφορες τεχνικές, σε Matlab και C++.

Introduction and contributions

This thesis focuses on the problem of segmentation and clustering of audio files to speakers, of problem termed in literature as *speaker diarization*. It is considered as a central problem, since many applications that are related to speech technologies require it as a preprocessing step. They require an algorithm that is capable of estimating in a computationally efficient way (a) the number of speakers and (b) the time segments that each of the speakers is active. Compared to other clustering and classification tasks, speaker diarization exhibits a pair of special characteristics that makes it both attractive and hard-to-tackle; the lack of knowledge of both the number of speakers and their identity. A proper algorithm should therefore estimate both their number and their density function, by grouping those utterances that belong to the same speaker. The statistical machine learning community has developed several clustering algorithms. However most of them require the number of clusters to be known beforehand. In speaker diarization though, the number of clusters should be estimated from the data, as well. In the thesis we first present some of the main approaches to the problem that have been proposed. We then focus on our proposals, which are divided into the three following contributions.

(a) Our first contribution is the development of a probabilistic measure of discrepancy between two speech segments. This discrepancy aims to estimate the posterior probability of the segments to belong to different speakers. The proposed model is capable of combining an unlimited number of binary weak classifiers, each of which should be considered as a combination of a feature space, a statistical model, a statistical divergence and a threshold. Several such models are trained, one for every partition of the input space (i.e. a sensible combination of recording conditions, gender, segment duration, a.o.) and are combined into a single probabilistic mixture-of-experts model. The use of Maximum A Posteriori (MAP) training is also compared to the classical Maximum Likelihood estimation. Our proposal has been presented in [1].

(b) The second contribution is a redefinition of one of the most frequently used approaches

to speaker diarization, namely the Bayesian Information Criterion (BIC). By examining the Bayesian rationale for BIC, we show that both of its current versions (the global and the local) are suboptimal for speaker diarization. Using the implied priors of BIC, we we propose a new version, the segmental-BIC, that leads to a significant increase in diarization accuracy. We further show how the use of Dirichlet distribution over the transition matrix can enrich the BIC with temporal information, which is usually ignored when using the BIC. This approach is analyzed in [2], [3], [4] (IEEE Journal) while several of its ideas in [5] and [6].

(c) Our third and final contribution is the examination of the potentials that the mean shift algorithm offers to the problem of speaker diarization. This algorithm is highly used in the image processing and computer vision and has been established a milestone in nonparametric segmentation. We show that it can be used to tackle more general clustering tasks, where the entities lie on non-Euclidean spaces, like those of statistical parametric models of exponential families. Using elements of Information Geometry and a Bayesian framework, we adapt the original algorithm and show that is capable of increasing the diarization accuracy when compared to the standard hierarchical clustering approach. A primary version of this approach is presented in [7], while its full version has been sent for publication is an international machine learning journal.

All algorithms have been developed in Matlab and C++ programming languages.

Περιεχόμενα

Κατάλογος Σχημάτων

Κατάλογος Πινάκων

1 ΕΙΣΑΓΩΓΗ

1.1 Γενική περιγραφή του προβλήματος

Ένα από τα κεντρικά προβλήματα στην επεξεργασία αρχείων ήχου με ομιλία αφορά στη διαδικασία κατάτμησης και ομαδοποίησης του αρχείου στους διακριτούς ομιλητές. Η ιδιαιτερότητα του προβλήματος έγκειται στην έλλειψη εκ των προτέρων (a priori) γνώσης τόσο ως προς τον συνολικό αριθμό των διακριτών ομιλητών K, όσο και ως προς την ταυτότητά τους. Σκοπός των αλγορίθμων είναι να επισημάνουν τα χρονικά διαστήματα κατά τα οποία ο κάθε διακριτός ομιλητής είναι ενεργός. Το πρόβλημα εντάσσεται στον χώρο της εκμάθησης χωρίς επίβλεψη (unsupervised learning), αφού οι κρυφές μεταβλητές (latent variables) θα πρέπει να εκτιμηθούν από τις φανερές (τα διανύσματα παρατήρησης στον επιλεγμένο παραμετρικό χώρο). Ώς κρυφές μεταβλητές ορίζουμε τους δείκτες που αντιστοιχούν κάθε διάνυσμα παρατήρησης με έναν από τους ομιλητές. Εξαιτίας της μη-γνώσης του K, η εν λόγω διαδικασία χαρακτηρίζεται και ως τυφλή ομαδοποίηση (blind clustering), έτσι ώστε να διαχωρισθεί από συγγενείς τεχνικές εκμάθησης χωρίς επίβλεψη, στις οποίες το K θεωρείται εκ των προτέρων γνωστό (π.χ. οικογένεια αλγορίθμων Expectation-Maximization, EM, [8], [9]).

Ο όρος που έχει επικρατήσει διεθνώς για την εξεταζόμενη διαδικασία καλείται Speaker Diarization. Επειδή θεωρούμε αδόκιμη τη μεταφρασή του όρου ως "Ημερολογιοποίηση", θα χρησιμοποιήσουμε τον περιφραστικό όρο *Κατάτμηση και Ομαδοποίηση Ομιλητών* (ΚΟΟ).

1.2 Πεδία εφαρμογής

Τα βασικότερα πεδία εφαρμογής (domains) της ΚΟΟ είναι τα ακόλουθα

- Πεδίο εκπομπών ειδησεογραφικού χαρακτήρα (Broadcast News) τηλεόρασης και ραδιοφώνου.

Το πεδίο αυτό μπορεί να υποδιαιρεθεί σε αμιγώς θεματολογικά δελτία, σε εκπομπές λόγου (γνωστά και ως talk-show) και σε συνδυασμό των δύο κατηγοριών, όπου η ροή της θεματολογίας διακόπτεται, προκειμένου να αναλυθεί ένα ή παραπάνω θέματα από σχολιαστές και προσκεκλημένους ειδήμονες (πολιτικούς, επιστήμονες, καλλιτέχνες, κ.ο.κ.).

- Πεδίο συνεδριάσεων (meetings). Το ιδιαίτερο χαρακτηριστικό του πεδίου αυτού είναι η ενδεχόμενη ύπαρξη συστοιχίας μικροφώνων (microphone array), η οποία επιτρέπει την εφαρμογή τεχνικών τυφλού διαχωρισμού πηγών (Blind Source Separation, BSS) όπως η Ανάλυση σε Ανεξάρτητες Συνιστώσες (Independent Component Analysis, ICA) και εκτίμησης του αριθμού των συμμετεχόντων βάσει της τοποθεσίας (location-based segmentation & clustering, [10], [11], [12], [13]). Στην παρούσα διατριβή, θα θεωρήσουμε ότι το προς επεξεργασία σήμα είναι μονοφωνικό, είτε λόγω ύπαρξης ενός μόνο μικροφώνου κατά την ηχογράφηση, είτε λόγω μετεπεξεργασίας και μίξης πολλαπλών σημάτων στο στούντιο παραγωγής.

- Πεδίο τηλεφωνικών διαλόγων. Παρότι σε έναν τυπικό τηλεφωνικό διάλογο συμμετέχουν δύο ομιλητές, υπάρχουν αρκετές εξαιρέσεις (π.χ. τηλεδιασκέψεις) οι οποίες επιβάλλουν την εκτίμηση του αριθμού των ομιλητών και σε αυτό το πεδίο.

- Πεδίο κινηματογραφικών ταινιών. Ένα ιδιαίτερα ενδιαφέρον πεδίο, με πολλές δυσκολίες λόγω επικάλυψης ομιλίας με μουσική και ηχητικά εφέ, αλλά και δυνατότητα σύμμιξης ηχητικής και οπτικής ροής πληροφορίας.

- Πεδίο ομαδοποίησης μεμονωμένων ηχογραφήσεων ομιλητών. Συνήθως τα τμήματα ομιλίας δεν προέρχονται από την ίδια ηχογράφηση (recording session). Έτσι, η διαφορά του πεδίου αυτού με τα προαναφερθέντα είναι διπλή. Πρώτον, η κατάτμηση είναι δοσμένη οπότε το πρόβλημα εξαντλείται στην ομαδοποίηση των τμημάτων. Η δεύτερη διαφορά έγκειται στην ύπαρξη ηχογραφήσεων του ίδιου ομιλητή υπό διαφορετικές συνθήκες ή και εξοπλισμό. Για να επιτευχθεί ομαδοποίηση τέτοιων τμημάτων απαιτούνται ειδικές κανονικοποιήσεις των διανυσμάτων παρατήρησης, ούτως ώστε να ελαττωθεί η επίδραση των παραμέτρων ηχογράφησης. Όπως θα αναλυθεί στα επόμενα κεφάλαια, οι τεχνικές αυτές έχουν ως κόστος την απώλεια χρήσιμης πληροφορίας σχετικά με τη χροιά του ομιλητή. Έτσι, οι τεχνικές αυτές εφαρμόζον-

ται μόνο σε περιπτώσεις ύπαρξης πολλαπλών συνθηκών ηχογράφησης του ιδίου ομιλητή. Το πρόβλημα αυτό εμφανίζεται και στην περίπτωση των ειδησεογραφικών εκπομπών, αλλά σε μικρότερο βαθμό. Μάλιστα, όπως θα δούμε στη συνέχεια, υπάρχουν περιπτώσεις στις οποίες η ομαδοποίηση παρεμβάσεων ομιλητή προερχόμενες από πολλαπλές συνθήκες ηχογράφησης σε μία ενιαία ομάδα δεν είναι κατ΄ ανάγκη επιθυμητή.

1.3 Χρήσεις της διαδικασίας

1.3.1 Αναγνώριση Φωνής προσαρμοζόμενη στον ομιλητή

Ένας από τους θεμελιώδεις στόχους της ΚΟΟ συσχετίζεται με την αναγνώριση φωνής. Η επιτυχία συστημάτων αναγνώρισης φωνής τα οποία προσαρμόζουν τις παραμέτρους των φωνηματικών μοντέλων στον ομιλητή και το περιβάλλον της ηχογράφησης οδήγησε στην ανάγκη κατάτμησης των αρχείων σε διακριτούς ομιλητές. Τέτοιες τεχνικές προσαρμογής παραμέτρων είναι η Γραμμική Παλινδρόμηση Μέγιστης Πιθανοφάνειας (Maximum Likelihood Linear Regression, MLLR, [14], [15]), η Μέγιστη *εκ των υστέρων* (Maximum A Posteriori, MAP) προσαρμογή, η πρόβλεψη μοντέλων βασισμένη στην Παλινδρόμηση (Regression-based Model Prediction), καθώς και υβριδικές τεχνικές. Τόσο η τεχνική MLLR και κυρίως η MAP απαιτούν τη συγκέντρωση αρκετού υλικού ωστε να επιτευχθει η προσαρμογή αυτή με τρόπο στιβαρό (robust). Για να επιτευχθεί αυτό σε περιβάλλον στο οποίο περισσότεροι του ενός ομιλητές είναι ενεργοί, προέκυψε η ανάγκη κατάτμησης και ομαδοποίησης το οφέλιμου υλικού σε διακριτούς ομιλητές. Στα συστήματα αυτά θα πρέπει να προστεθούν και αυτά που δεν προσαρμόζουν τις παραμέτρους σύμφωνα με τον ομιλητή, αλλά μετασχηματίζουν τα διανύσματα παρατήρησης. Στις τεχνικές αυτές περιλαμβάνονται η Κανονικοποίηση με βάση το μήκος της φωνητικής οδού Vocal Tract Length Normalization [16], [17], η οποία επιφέρει μείωση του ποσοστού λαθών επιπέδου λέξης (Word Error Rate, WER) της τάξης του 10% σε καθαρή ομιλία [18].

1.3.2 Εμπλουτισμός κειμένου αυτόματης απομαγνητοφώνησης

Η κατάτμηση και ομαδοποίηση σε διακριτούς ομιλητές καθιστά δυνατή την εξαγωγή εμπλουτισμένου κειμένου αυτόματης απομαγνητοφώνησης (rich transcription), στο οποίο κάθε πρόταση αποδίδεται και σε έναν συγκεκριμένο ομιλητή. Η πληροφορία αυτή, παρότι δεν παρέχει αναγνώριση της ταυτότητας του ομιλητή, βοηθάει στην παρακολούθηση και ανάγνωση του κειμένου απομαγνητοφώνησης, ιδιαίτερα σε περιβάλλον διαλόγου μεταξύ ομιλητών.

1.3.3 Συστήματα αναγνώρισης ομιλητών

Μία σημαντικότατη χρήση της ΚΟΟ είναι η αναγνώριση ομιλητών [19]. Στόχος εδώ είναι να εξετασθεί η παρουσία ενός ή παραπάνω ομιλητή σε συγκεκριμένη εκπομπή. Για να επιτευχθεί αυτό, προαπαιτείται η συλλογή επαρκούς ακουστικού υλικού για τον κάθε ομιλητή-στόχο και η εκπαίδευση των μοντέλων κάθε ομιλητή βάσει του διαθέσιμου υλικού. Έτσι, η υπάρξη της βαθμίδας ΚΟΟ έχει διττό ρόλο. Αρχικά, η τυφλή ομαδοποίηση σε διακριτούς ομιλητές επιταχύνει και διευκολύνει τη συλλογή του υλικού από εκπομπές αρχείου, μέσα από μια ημι-επιβλεπόμενη διαδικασία. Αρκεί κάποιος στο να ακούσει ένα μόνο μέρος του υλικού από κάθε μία από τις K ομάδες (clusters) και να αποφασίσει εάν όντως η συγκεκριμένη ομάδα αντιστοιχεί σε κάποιον εκ των ομιλητών-στόχων. Επιπλέον, κατά την πραγματική λειτουργία συστήματος αναγνώρισης ομιλητών, η βαθμίδα της ΚΟΟ επιτρέπει στην αναγνώριση να γίνεται με βάση μόνο τις K ομάδες. Το κέρδος εν προκειμένω είναι τόσο χρονικό - υπολογιστικό, αφού ένα τμήμα ομιλίας διάρκειας 10 sec είναι συνήθως αρκετό για να ταυτοποιηθεί η εξεταζόμενη ομάδα, όσο και ακρίβειας της αναγνώρισης. Λόγω της ομαδοποίησης, η εφαρμογή των απαιτούμενων τεχνικών κανονικοποίησης (channel compensation) γίνεται σε επίπεδο ομάδας, αυξάνοντας τη στιβαρότητα στην εκτίμηση των παραμέτρων.

1.3.4 Συστήματα κατάτμησης βάσει θεματολογίας

Μία ενδιαφέρουσα εφαρμογή της ΚΟΟ είναι η υποβοήθηση που μπορεί να προσφέρει σε συστήματα αυτόματης κατάτμησης ειδησεογραφικών εκπομπών κατά θέμα ή ιστορία. Ο όρος που έχει επικρατήσει διεθνώς για το πρόβλημα αυτό είναι Story Segmentation. Το σύστημα αυτό αποσκοπεί στην εύρεση των χρονικών σημείων όπου αλλάζει το θέμα (ή η ιστορία) κατά τη ροή της εκπομπής. Βάσει της κατάτμησης αυτής, η επόμενη βαθμίδα του συστήματος μπορεί να προβεί στην αυτόματη περίληψη των θεμάτων και στην κατάταξή του σε ευρείες θεματικές ενότητες. Έτσι, παρέχεται η δυνατότητα αυτόματης προσπέλασης στα αρχεία εκπομπών μέσω συστημάτων ανάκτησης πληροφορίας (Information Retrieval, [20]).
Ο ρόλος της βαθμίδας ΚΟΟ είναι η εξαγωγή χρήσιμης πληροφορίας από την ηχητική συνιστώσα, καθώς και η υποβοήθηση της διαδικασίας αναγνώρισης φωνής. Αρχικά, ο εντοπισμός του βασικού παρουσιαστή (Anchor Tracking) αποτελεί μία ισχυρή συνιστώσα, καθώς ένα θέμα συνήθως εισάγεται από τον βασικό παρουσιαστή. Επιπλέον, η ΚΟΟ μπορεί να εντοπίσει και να κατατάξει περιοχές του αρχείου σε κατηγορίες όπως διάλογος προσκεκλημένων, τηλεφωνική παρέμβαση, εξωτερικό ρεπορτάζ, καθώς και να εντοπίσει τα χαρακτηριστικά μουσικά σήματα του σταθμού ή της εκπομπής - γνωστά και ως jingles. Το σύνολο των πληροφοριών που εξάγει η ΚΟΟ, μαζί με άλλες τροπικές συνιστώσες (modes) όπως πληροφορία από επεξεργασία εικόνας και επεξεργασία φυσικής γλώσσας (Natural Language Processing, NLP) συνδυάζονται σε ένα ενιαίο πιθανοτικό μοντέλο, το οποίο προτείνει σημεία αλλαγής θεματολογίας, [21] [22].

1.4 Συμβολισμοί μεταβλητών

Ορίζουμε το διάνυσμα παρατήρησης ως $\mathbf{y}^{(i)} \in \mathbb{R}^d, i = 1, \ldots, n$, όπου n είναι το πλήθος των διανυσμάτων στο αρχείο. Τα διανύσματα τα θεωρούμε πάντοτε διανύσματα-στήλες. Ο πίνακας παρατηρήσεων που αντιστοιχεί στο αρχείο και τον επιλεγμένο d-διάστατο παραμετρικό χώρο θα συμβολίζεται ως $\mathbf{y} = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}]^T$, και θα είναι διάστασης $n \times d$. Δεδομένου ότι το πρόβλημα είναι η

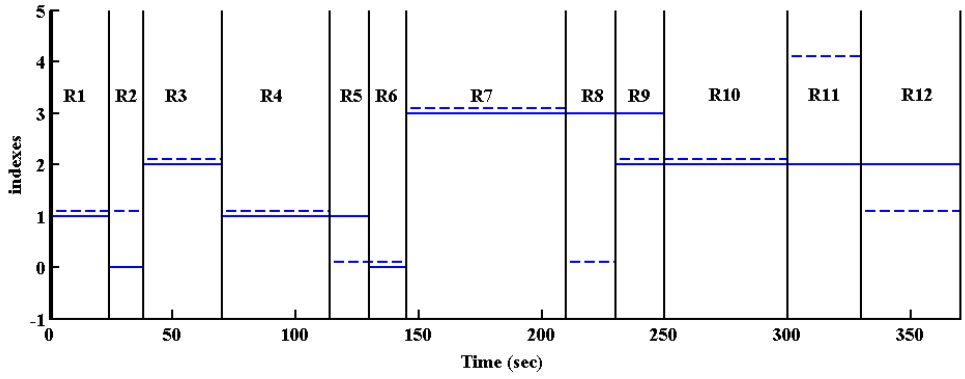
ομαδοποίηση του $\mathbf{y}$, ορίζουμε τα ολοκληρωμένα δεδομένα (complete data), $\mathbf{z}^{(i)} = (\mathbf{y}^{(i)}, x^{(i)})$, όπου το $\mathbf{x}^{(i)} \in \{1, \ldots, K\}$ είναι ο δείκτης που ορίζει σε ποιά από τις K ομάδες αντιστοιχεί *πραγματικά* το $\mathbf{y}^{(i)}$. Ο στόχος της βαθμίδας είναι να εκτιμήσει την πραγματική αντιστοίχιση $\mathbf{x} = [x^{(1)}, \ldots, x^{(n)}]^T$ από το $\mathbf{y}$, χωρίς εκ των προτέρων γνώση του K. Συμβολίζουμε την εκτίμηση των $\mathbf{x}$ και K ως $\hat{x}$ και $\hat{K}$ αντίστοιχα.

1.5 Μετρικές αξιολόγησης βαθμίδας

1.5.1 Γενικές παρατηρήσεις πάνω στις μετρικές αξιολόγησης

Εξετάζουμε τώρα τις μετρικές με τις οποίες αξιολογείται η απόκριση της ΚΟΟ. Σαν γενική παρατήρηση, πιστεύουμε ότι μετρική η οποία να αναδεικνύει όλα τα είδη σφαλμάτων δεν μπορεί να υπάρξει. Η αξιολόγηση εξαρτάται πάντοτε από το πεδίο εφαρμογής της βαθμίδας. Όπως εξηγήθηκε παραπάνω, η βαθμίδα ΚΟΟ έχει πολλαπλό ρόλο και εντάσσεται σε πληθώρα συστημάτων, κάθε ένα εκ των οποίων έχει τις δικές του απαιτήσεις, προδιαγραφές και ανοχές στους διαφόρους τύπους σφαλμάτων.

Η κατανόηση των μετρικών αξιολόγησης θα γίνει πιο εύκολη με το ακόλουθο παράδειγμα. Στο Σχ.1 παρουσιάζεται η πραγματική ομαδοποίηση (κανονική οριζόντια γραμμή) ενός αρχείου ήχου διάρκειας 370 sec μαζί με την εκτιμώμενη (διακεκομμένη οριζόντια γραμμή). Ο συνολικός αριθμός των ομιλητών αναφοράς είναι 3, καθώς ο μηδενικός δείκτης συμβολίζει τμήμα μη-ομιλίας. Οι κάθετες γραμμές απεικονίζονται ώστε να ορίσουν τα μέγιστα διαστήματα κατά τα οποία τόσο το πραγματικό όσο και το εκτιμώμενο διάστημα αποδίδονται σε έναν διακριτό ομιλητή, ή σε τμήμα μη-ομιλίας.



Σχῆμα 1: *Παράδειγμα κατάτμησης και ομαδοποίησης*

1.5.2 Το ολικό σφάλμα βαθμίδας ΚΟΟ

Ένας αρχικός τρόπος αξιολόγησης της βαθμίδας είναι το ολικό σφάλμα βαθμίδας ΚΟΟ (total Diarization Error Rate, DER %), που ορίζεται ως

$$DER = \frac{\sum_{seg=1}^{N_s} dur(seg) \cdot (H_{miss} + H_{fa} + H_{spkr})}{\sum_{seg=1}^{N_s} dur(seg) \cdot H_{ref}} \tag{1}$$

όπου ως seg ορίζουμε το μέγιστο συνεχόμενο διάστημα κατά το οποίο τόσο η απόκριση του συστήματος όσο και η πραγματική κατάτμηση αποδίδουν έναν και μόνο έναν δείκτη και

1. $H_{miss} = 1$, αν και μόνο αν το τμήμα δεν συμπεριελήφθη στους ομιλητές, παρότι ήταν τμήμα ομιλίας, αλλιώς 0,

2. $H_{fa} = 1$, αν και μόνο αν το τμήμα συμπεριελήφθη στους ομιλητές, παρότι δεν ήταν τμήμα ομιλίας, αλλιώς 0,

3. $H_{spkr} = 1$, αν και μόνο αν είναι τμήμα ομιλίας αλλά αποδόθηκε σε άλλον ομιλητή, αλλιώς 0,

4. $H_{ref} = 1$, αν και μόνο αν είναι τμήμα που αντιστοιχεί σε ομιλία, αλλιώς 0, ανεξαρτήτως της απόφασης της βαθμίδας.

Βάσει αυτών των ορισμών θα κατατάξουμε τα διαστήματα του παραδείγματος που παρουσιάζεται στο Σχ.1. Τα διαστήματα R1, R3, R4, R7, R10 έχουν εκτιμηθεί σωστά από τη βαθμίδα και αποτελούν τμήμα ομιλίας. Επομένως, για τα διαστήματα αυτά θα ισχύει $H_{ref} = 1$ και οι υπόλοιπες δυαδικές ενδείξεις θα είναι μηδενικές. Το διάστημα R2 αντιστοιχεί σε τμήμα μη-ομιλίας, το οποίο εκτιμήθηκε λαθεμένα σαν τμήμα ομιλίας, επομένως $H_{ref} = 0, H_{fa} = 1$. Το αντίστροφο συμβαίνει με τα τμήματα R5, R8, τα οποία ενώ είναι τμήματα ομιλίας θεωρήθηκαν μη-ομιλίας, επομένως $H_{ref} = 1, H_{miss} = 1$. Τα διαστήματα R11, R12 καλούνται λάθη ομιλητή (speaker errors) καθώς αποδόθηκαν σε λάθος ομιλητή. Για τα διαστήματα αυτά θα ισχύει $H_{ref} = 1, H_{spkr} = 1$. Τέλος, στο διάστημα R9 παρατηρείται ταυτόχρονη ομιλία των ομιλητών αναφοράς 2 και 3. Το διάστημα αυτό είτε εξαιρείται από την αξιολόγηση, επομένως όλες οι δυαδικές ενδείξεις λαθών είναι μηδενικές, είτε απαιτείται ο εντοπισμός και των δύο ομιλητών.

Λόγω της σχετικότητας των δεικτών, για να υπολογισθεί η παραπάνω μετρική σφάλματος, έχει προηγηθεί αλγόριθμος ο οποίος αναδιατάσσει τους δείκτες απόκρισης βαθμίδας, σύμφωνα με τους δείκτες αναφοράς. Στόχος είναι να βρεθεί η αντιστοίχιση εκείνη η οποία οδηγεί στο ελάχιστο DER. Για να επιτευχθεί αυτή η αναδιάταξη των δεικτών, ο αλγόριθμος εντοπίζει την ομάδα εκείνη που παρουσιάζει τη μέγιστη επικάλυψη με τον κάθε ομιλητή αναφοράς.

1.5.3 Τα διαγράμματα καθαρότητας ομάδων και ομιλητών

Ένας εναλλακτικός τρόπος αξιολόγησης της βαθμίδας και κατά τη γνώμη μας πιο ολοκληρωμένος σε σχέση με το DER, είναι τα διαγράμματα *μέσης καθαρότητας ομάδας* (average cluster purity, *acp*) και *μέσης καθαρότητας ομιλητή* (average speaker purity, *asp*) [23]. Με τον όρο *ομάδα* αναφερόμαστε στην *εκτίμηση* της βαθμίδας, ενώ με τον όρο *ομιλητή* στον πραγματικό ομιλητή ή ομιλητή *αναφοράς* (ground truth or reference speaker). Αρχικά ορίζουμε την καθαρότητα της k-οστής ομάδας $p_{k.}$ και την καθαρότητα του r-οστού ομιλητή $p_{.r}$ ως

$$p_{k.} = \sum_{r=1}^{K} \frac{n_{kr}^2}{n_{k.}^2}, \quad p_{.r} = \sum_{k=1}^{\hat{K}} \frac{n_{kr}^2}{n_{.r}^2} \tag{2}$$

όπου με n_{kr} συμβολίζουμε τον σύνολο των $\mathbf{y}^{(i)}$ που η βαθμίδα συμπεριέλαβε στην k-οστή ομάδα ενώ στην πραγματικότητα ανήκουν στον r-οστο ομιλητή.

Οι μέσες τιμές των παραπάνω ποσοτήτων δίνουν τις ζητούμενες μετρικές acp και asp ως

$$acp = \frac{1}{n}\sum_{k=1}^{K} n_{k.}p_{k.}, \quad asp = \frac{1}{n}\sum_{r=1}^{\hat{K}} n_{.r}p_{.r} \tag{3}$$

Από τον ορισμό των παραπάνω ποσοτήτων είναι γίνεται η συμπληρωματικότητα των δύο μετρικών. Η μέση καθαρότητα ομάδας (acp) αυξάνεται όσο η κάθε ομάδα περιέχει παρατηρήσεις που προέρχονται από έναν και μόνο έναν ομιλητή, ανεξαρτήτως του βαθμού στον οποίο οι ομιλητές αναφοράς έχουν διασπασθεί στις K ομάδες. Συμμετρικά, η μέση καθαρότητα ομιλητή (asp) αξιολογεί τον μέσο βαθμό διασποράς του κάθε ομιλητή αναφοράς σε περισσότερες από μια ομάδες, ανεξαρτήτως της καθαρότητας της προτεινόμενης ομαδοποίησης. Έτσι, μία εναλλακτική ονομασία της μέσης καθαρότητας ομιλητή είναι η μέση κάλυψη ομάδας (average cluster coverage).

Αν ενδιαφερόμαστε να ενοποιήσουμε τις δύο μετρικές σε μία, δύο επιλογές είναι ο γεωμετρικός μέσος $\sqrt{acp \times asp}$ καθώς και ο δείκτης Rand (Rand Index)

$$RI = \frac{\sum_{k=1}^{K} n_{k.}^2 + \sum_{r=1}^{\hat{K}} n_{.r}^2 - 2\sum_{k=1}^{K}\sum_{r=1}^{\hat{K}} n_{kr}^2}{\sum_{k=1}^{K} n_{k.}^2 + \sum_{r=1}^{\hat{K}} n_{.r}^2} \tag{4}$$

Για μία ανάλυση των μετρικών απόκλισης μεταξύ εναλλακτικών ομαδοποιήσεων προτείνουμε το [24].

1.5.4 Η κατάτμηση και ομαδοποίηση με όρους εκτίμησης παραμέτρων και φειδωλής τάξης μοντέλου

Πριν προχωρήσουμε στην αναδρομή των βασικών τεχνικών και αλγορίθμων της βαθμίδας, χρήσιμο είναι να συνδέσουμε τις δύο μετρικές acp και asp με την ορολογία της μάθησης μηχανών και στατιστικής εκτίμησης παραμέτρων και τάξης μοντέλου. Οι ομαδοποιήσεις που δίνουν $acp > asp$ αντιστοιχούν σε στατιστική υπερ-εκπαίδευση ή υπερ-ταίριασμα του μοντέλου στα δεδομένα (overfit-

ting the data) και συνήθως αντιστοιχούν σε λύσεις όπου $\hat{K} > K$. Στην ορολογία της στατιστικής μοντελοποίησης, το K εκφράζει τη φειδωλή τάξη του μοντέλου (parsimonious order of the model) και οι ομαδοποιήσεις με $\hat{K} > K$ υπερβαίνουν την τάξη αυτή. Όπως θα αναλυθεί παρακάτω, η στρατηγική που συνήθως ακολουθείται είναι η πόλωση των εκτιμήσεων υπέρ της acp, τουλάχιστον κατά τα αρχικά στάδια της βαθμίδας.

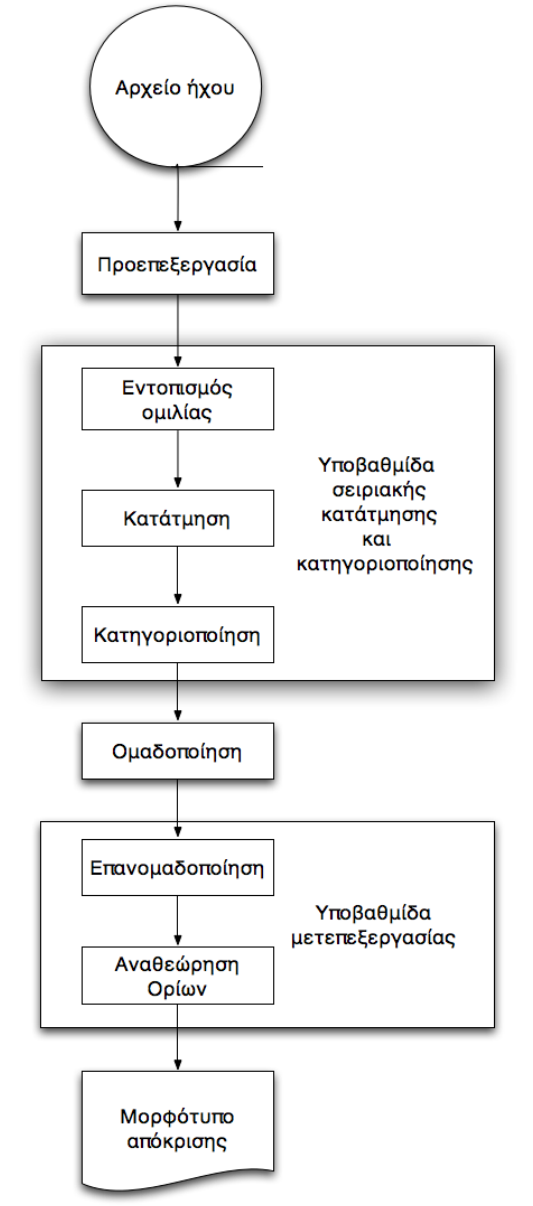
Η επιλογή της *πόλωσης* θα πρέπει επίσης να συσχετισθεί με το πεδίο εφαρμογής της βαθμίδας. Εάν η βαθμίδα επιτελεί τον ρόλο προσταδίου συστήματος αναγνώρισης ομιλητή ή συλλογής δεδομένων ομιλητή, η συγκεκριμένη πόλωση είναι δικαιολογημένη, καθώς το κόστος από ενδεχόμενη ομαδοποίηση άνω του ενός ομιλητή σε μία ομάδα επιφέρει σημαντική μείωση της αξιοπιστίας του συστήματος αναγνώρισης. Αντίθετα, το σύστημα είναι ικανό να λειτουργήσει σε χαμηλότερα επίπεδα μέσης καθαρότητας ομιλητή, με κόστος την ανάγκη ταυτοποίηση περισσότερων ομάδων με τους ομιλητές-στόχους.

2 ΑΝΑΔΡΟΜΗ ΒΑΣΙΚΩΝ ΤΕΧΝΙΚΩΝ ΚΑΤΑΤΜΗΣΗΣ ΚΑΙ ΟΜΑΔΟΠΟΙΗΣΗΣ

2.1 Γενικές έννοιες και κατηγοριοποιήσεις

Στο παρόν κεφάλαιο εξετάζονται οι διάφορες προσεγγίσεις του προβλήματος που έχουν προταθεί στη βιβλιογραφία, με έμφαση στο πεδίο των εκπομπών ειδησεογραφικού χαρακτήρα (Broadcast News, BN). Το πεδίο αυτό είναι πιο γενικό σε σχέση με το πεδίο συσκέψεων (meetings), αφού έχουμε να αντιμετωπίσουμε πληθώρα ηχητικών οντοτήτων, διάλογο μεταξύ ομιλητών, μίξη σημάτων ομιλίας με ήχους περιβάλλοντος και μουσικής στο στούντιο παραγωγής και πολλά άλλα απρόβλεπτα φαινόμενα. Θα θεωρήσουμε ότι έχουμε στη διάθεσή μας ένα μονοφωνικό αρχείο ήχου, το οποίο αποτελείται από τμήματα ομιλίας (αγνώστου αριθμού ομιλητών) καθώς και από άλλες ηχητικές οντότητες, όπως διαστήματα σιωπής, ήχους περιβάλλοντος, μουσική, κ.ά.
Μία βασική κατηγοριοποίηση των αλγορίθμων βαθμίδας ΚΟΟ έγκειται στη σύζευξη ή μη των διαδικασιών κατάτμησης και ομαδοποίησης. Οι *αποσυζευγμένοι* (uncoupled ή disjoint ή και βήμα-προς-βήμα) αλγόριθμοι ([25], [26], [27], [28]) χαρακτηρίζονται από ένα αρθρωτό (modular) διάγραμμα ροής χωρίς ανατροφοδότηση (feedback) μεταξύ των αρθρώσεων, όπου κάθε άρθρωση (module) επιτελεί μια καθορισμένη λειτουργία. Αντίθετα, οι *συζευγμένοι* αλγόριθμοι ([29],[30]) δεν προβαίνουν σε αποκλειστική (explicit) κατάτμηση των δεδομένων, αλλά αντιμετωπίζουν την κατάτμηση και ομαδοποίηση σε ένα ενιαίο στάδιο, με βάση αναδρομικούς (iterative) αλγορίθμους μη επιβλεπόμενης εκμάθησης για δεδομένα τα οποία εμπεριέχουν Μαρκοβιανές ιδιότητες. Η έμφαση θα δοθεί στους αποσυζευγμένους αλγορίθμους, ενώ στο τέλος του κεφαλαίου θα παρουσιασθούν και μερικές υλοποιήσεις συζευγμένων αλγορίθμων.

Στο Σχήμα 2 παρουσιάζεται ένα τυπικό διάγραμμα ροής αποσυζευγμένων αλγορίθμων ΚΟΟ. Θα πρέπει να τονισθεί ότι το διάγραμμα είναι αρκετά γενικό. Ένα βασικό χαρακτηριστικό που διέπει όλα τα στάδια είναι η σταδιακή μετάβαση της μοντελοποίησης από μοντέλα χαμηλής στατιστικής πολυπλοκότητας προς πιο σύνθετης. Η ίδια η φύση του προβλήματος επιβάλλει μια τέτοια



Σχῆμα 2: *Διάγραμμα ροής βαθμίδας αποσυζευγμένης κατάτμησης και ομαδοποίησης*

στρατηγική. Ελλείψει επαρκών στατιστικών δεδομένων, η στιβαρή μοντελοποίηση της ανθρώπινης ομιλίας με μοντέλα σύνθετα όπως τα μίγματα κανονικών κατανομών (Gaussian Mixture Models,

GMMs) δεν είναι δυνατή, τουλάχιστον στα αρχικά στάδια του αλγορίθμου. Έτσι, τόσο στην υποβαθμίδα σειριακής κατάτμησης και κατηγοριοποίησης, όσο και στην υποβαθμίδα ομαδοποίησης, η μοντελοποίηση των τμημάτων ομιλίας γίνεται μέσω απλών κανονικών κατανομών και όχι μιγμάτων. Η υποβαθμίδα μετεπεξεργασίας, έχοντας στη διάθεσή της μία ομαδοποίηση ομιλητών με επαρκή στατιστικά στοιχεία, είναι ικανή να προβεί σε πιο σύνθετες μοντελοποιήσεις μέσω GMMs, ώστε να περιγραφούν πιο αναλυτικά οι περιοχές ομιλίας γύρω από τα φωνήματα, [27], [31].
Ένα δεύτερο χαρακτηριστικό της ροής των αποσυζευγμένων αλγορίθμων είναι η συσσωρευτική (agglomerative) ή αλλιώς *από κάτω προς τα πάνω* στρατηγική (bottom-up). Αν και ο ορισμός αυτός είναι περισσότερο ταυτισμένος με την υποβαθμίδα της ομαδοποίησης και συγκεκριμένα με την ιεραρχική ομαδοποίηση (Hierarchical Clustering), θεωρούμε ότι είναι μια γενικότερη φιλοσοφία που διέπει τους αποσυζευγμένους αλγορίθμους. Όπως αναφέρθηκε παραπάνω, η στρατηγική είναι να πολώνονται οι εκτιμήσεις υπέρ της καθαρότητας των ομάδων σε σχέση με την κάλυψη, μεταθέτοντας την αύξηση της μέσης κάλυψης στις επόμενες αρθρώσεις και υποβαθμίδες. Έτσι, τόσο κατά την κατάτμηση όσο και κατά την ομαδοποίηση, η έμφαση δίδεται στο να μην ταξινομηθούν υπό την ίδια ομάδα διαφορετικές ηχητικές οντότητες.

2.2 Στάδιο προεπεξεργασίας

Το στάδιο αυτό εκτελεί τους μετασχηματισμούς στο σήμα ήχου από το πεδίο του χρόνου στον ή στους κατάλληλους παραμετρικούς χώρους (feature space). Για τον διαχωρισμό μεταξύ ομιλητών, οι πλέον καθιερωμένοι παραμετρικοί χώροι είναι αυτοί των Mel-Frequency Cepstral Coefficients (MFCC) ([32]) και των Perceptually-based Linear Predictive analysis (PLP) [33]. Εναλλακτικές επιλογές αποτελούν τα Linear-Frequency Cepstral Coefficients (LFCC) καθώς και των Line-Spectrum Pairs (LSP). Για την ταξινόμηση σε ηχητικές κατηγορίες, όπως ομιλία, μουσική, σιωπές και ήχους περιβάλλοντος, ιδιαίτερη διαχωριστική (discriminative) ικανότητα παρουσιάζουν τα LSP. Στο [34] παρουσιάζονται μέθοδοι κατάταξης των ειδών θορύβου, βασισμένη στα LSP, με απώτερο στόχο τη γνώση περιβάλλοντος (context-awareness). Η γνώση του περιβάλλοντος είναι

ιδιαίτερα χρήσιμη για την ανάπτυξη προσαρμοστικών (adaptive) συστημάτων. Ικανότητα διαχωρισμού παρουσιάζουν και μονοδιάστατοι χώροι, όπως ο ρυθμός των zero-crossings (Zero-Crossing Rate, ZCR), η μέση φασματική μεταβολή (Spectrum Flux) και το ποσοστό των θορυβωδών διανυσμάτων (Noise Frame Ratio, NFR) [35]. Ένα ενδιαφέρον χαρακτηριστικό της ανθρώπινης ομιλίας το οποίο μπορεί να βοηθήσει σημαντικά τον διαχωρισμό ομιλίας και μουσικής είναι και η διαμόρφωση της ενέργειας που παρουσιάζουν τα σήματα ομιλίας, λόγω του ρυθμού συλλαβισμού, με κεντρική συχνότητα διαμόρφωσης περί τα 4Hz. Η εξαγωγή του εν λόγω χαρακτηριστικού επιτυγχάνεται με πιο στιβαρό τρόπο ανά κανάλι μέσω της Mel-Filter Bank. Η απόκριση κάθε ζώνης (δηλαδή η λογαριθμική ενέργεια ζώνης) φιλτράρεται εκ νέου με χρήση ζωνοπερατού φίλτρου κεντρικής συχνότητας 4Hz και κατόπιν χρονικής εξομάλυνσης, οι ενέργειες των ζωνών προστίθενται, δίνοντας ένα βαθμωτό χαρακτηριστικό [36]. Τα χαρακτηριστικά αυτά χρησιμοποιούνται από συστήματα τα οποία προβαίνουν σε αναλυτική κατάταξη των τμημάτων σε ηχητικές κατηγορίες (audio indexing). Αν η εφαρμογή απαιτεί μόνο τον εντοπισμό των τμημάτων ομιλίας και τον διαχωρισμό μεταξύ ομιλητών, η βαθμίδα εξάγει χαρακτηριστικά μόνο σε έναν παραμετρικό χώρο, συνήθως τα MFCC.
Αναφέρουμε τέλος συστήματα που χρησιμοποιούν πολλαπλούς παραμετρικούς χώρους στη βαθμίδα ΚΟΟ. Οι μέθοδοι αυτοί επιχειρούν μέσω της σύμμιξης (fusion) των χώρων να επιτύχουν πιο ακριβή αποτελέσματα. Ως παράδειγμα, αναφέρουμε το [37], όπου ο αλγόριθμος επιλέγει μεταξύ των πολλαπλών διαθέσιμων ροών (streams) χαρακτηριστικών (συγκεκριμένα μεταξύ των MFCC, PLP και το pitch) βάσει ενός αλγορίθμου δυναμικού προγραμματισμού. Το πεδίο εφαρμογής είναι οι συσκέψεις (meetings domain) και αναφέρεται σημαντική αύξηση της ακρίβειας.

2.3 Στάδιο σειριακής κατάτμησης και κατηγοριοποίησης

Εξετάζουμε τώρα ένα καθοριστικό στάδιο της βαθμίδας ΚΟΟ αποσυζευγμένων αλγορίθμων. Το στάδιο σειριακής κατάτμησης και κατηγοριοποίησης λαμβάνει ως είσοδο n το πλήθος d-διάστατες παρατηρήσεις $\mathbf{y} \in \mathcal{Y}^n$, $\mathcal{Y} \subseteq \mathbb{R}^d$ που προκύπτουν από το στάδιο της προεπεξεργασίας και εκτιμά τις

χρονικές στιγμές όπου λαμβάνει χώρα μία εναλλαγή. Τα σημεία εναλλαγής μπορεί να αντιστοιχούν σε εναλλαγή μεταξύ ομιλητών, καθώς και σε οποιαδήποτε άλλη εναλλαγή μεταξύ ηχητικών οντοτήτων. Μαζί με τα όρια κάθε ηχητικής οντότητας, κάθε τμήμα κατατάσσεται και σε μία προκαθορισμένη κατηγορία (γνωστές και ως *μακροσκοπικές κλάσεις*, [28]). Εστιάζοντας στο πεδίο των ειδησεογραφικών εκπομπών και στα τμήματα ομιλίας, οι συνηθέστερες κατηγορίες αφορούν στο φύλο του ομιλητή και στο εύρος φάσματος του τμήματος (ευρύ ή τηλεφωνικό φάσμα). Χρήσιμες επεκτάσεις αποτελούν το περιβάλλον ηχογράφησης, όπως ομιλία από στούντιο ή εξωτερική-θορυβώδης ηχογράφηση, η ύπαρξη αντήχησης (reverberation) στον χώρο ηχογράφησης, καθώς και ο τύπος μικροφώνου (πυκνωτικό ή δυναμικό) που χρησιμοποιήθηκε.

Η αιτία που οδηγεί στην κατάταξη των τμημάτων σε αναλυτική περιγραφή είναι διττή. Πρώτον, αποτελεί μία χρήσιμη μετα-πληροφορία (metadata) στο πλαίσιο του εμπλουτισμού του εξαγώμενου κειμένου απομαγνητοφώνησης (transcript). Ο δεύτερος λόγος αφορά στην ίδια την ακρίβεια της βαθμίδας ΚΟΟ. Η σε βάθος κατηγοριοποίηση επιτρέπει την κατάλληλη προσαρμογή των παραμέτρων του συστήματος, σε επίπεδο διανυσμάτων χαρακτηριστικών, στατιστικών μοντέλων και αποστάσεων καθώς και κατωφλίων (thresholds) απόφασης. Όπως εξετάζεται στο [38], όπου παρατηρείται σημαντικό κέρδος κυρίως σε συστήματα συζευγμένης κατάτμησης και ομαδοποίησης. Επί παραδείγματι, η στατιστική απόσταση δύο τμημάτων ομιλίας που προέρχονται από δύο συγκεκριμένους ομιλητές μειώνεται όσο αυξάνεται ο θόρυβος ηχογράφησης (εφόσον ο θόρυβος θεωρηθεί ταυτόσημων στατιστικών). Το ίδιο ισχύει και για χώρους αυξημένης αντήχησης, όπως οι αίθουσες Κοινοβουλίων. Στις περιπτώσεις αυτές, αν δεν προηγηθεί κατηγοριοποίηση των τμημάτων στην κατάλληλη κατηγορία, ενυπάρχει ο κίνδυνος ομαδοποίησης τμημάτων ομιλίας διαφορετικών ομιλητών σε μία και μόνο ομάδα ή παρόμοια μη εντοπισμού σημείου εναλλαγής μεταξύ ομιλητών, γεγονός που οδηγεί σε υποβάθμιση της ακρίβειάς της.

2.4 Στάδιο εντοπισμού ομιλίας

Η πιο γενική αντιπετώπιση του προβλήματος εντοπισμού ομιλίας είναι αυτή της κατάταξης μέγιστης πιθανοφάνειας (Maximum Likelihood Classification) μέσω Μειγμάτων Κανονικών Κατανομών (Gaussian Mixture Models, GMMs). Τα μοντέλα αυτά είναι διάστασης συνήθως 128 συντελεστών, έχουν εκπαιδευθεί με αντιπροσωπευτικό υλικό διάρκειας τουλάχιστον δύο ωρών. Η εκπαίδευση των μοντέλων γίνεται μέσω του αλγορίθμου Expectation - Maximization (EM) [8] είτε με εκτίμηση μέγιστης πιθανοφάνειας και κατάλληλη αρχικοποίηση [39], είτε με Μπεϋσιανή εκτίμηση. Πολλές υλοποιήσεις εκπαιδεύουν διαφορετικά μοντέλα ομιλίας βάσει των συνδυασμών φύλου/εύρους ζώνης/θορύβου κ.λπ. Για πεδία εφαρμογής που υφίσταται μόνο διάλογος μεταξύ ομιλητών, ο μόνος διαχωρισμός που απαιτείται είναι αυτός μεταξύ ομιλίας και παρατεταμένης σιωπής (συνήθως άνω των 0.5 sec). Οι τεχνικές αυτές διαφέρουν καθώς οι αλγόριθμοι ανήκουν στην κατηγορία εκμάθησης χωρίς επίβλεψη (unsupervised classification). Πιο συγκεκριμένα, ένα μίγμα κανονικών κατανομών δύο συντελεστών μοντελοποιεί τη λογαριθμική ενέργεια μέσω του EM αλγορίθμου. Η σύγκλιση του αλγορίθμου είναι ταχεία καθώς τα μεγέθη είναι βαθμωτά. Κατόπιν, τα διανύσματα που κατατάχθηκαν στην κατανομή με τη χαμηλότερη μέση τιμή αντιστοιχίζονται σε σιωπές.

Οι παραπάνω τεχνικές οδηγούν σε απότομες διακυμάνσεις στην κατάταξη των διανυσμάτων παρατήρησης, καθώς δεν λαμβάνουν υπ' όψη τις Μαρκοβιανές ιδιότητες στο πεδίου του χρόνου. Έτσι, απαιτούνται μέθοδοι εξομάλυνσης στο πεδίο του χρόνου. Τέτοιες τεχνικές βασίζονται στη χρήση Υπονοούμενων Μαρκοβιανών Μοντέλων (Hidden-Markov Models, HMMs) καθώς και στα μορφολογικά φίλτρά (morphological filters). Στην περίπτωση των HMMs, οι κατηγορίες αντιστοιχούν στις καταστάσεις του μοντέλου και περιγράφονται με τα GMMs που προαναφέρθηκαν.

Οι μαρκοβιανές ιδιότητες μοντελοποιούνται μέσω του (μη συμμετρικού) πίνακα μετάβασης (state transition matrix)

$$\alpha_{kk'} = p(s_{t+1} = k' | s_t = k) \tag{5}$$

$$\sum_{k'=1}^{N_s} \alpha_{kk'} = 1, \forall k \in [1, \ldots, N_s] \tag{6}$$

μεταξύ των N_s καταστάσεων. Ο πίνακας αυτός μπορεί να εκτιμηθεί μέσω τεχνικών εκμάθησης με επίβλεψη, όπου απαιτούνται προταξινομημένα σύνολα εκπαίδευσης (labeled training sets). Στην περίπτωση αυτή, η εκτίμηση της πιθανότητας μετάβασης μπορεί να υπολογισθεί με τη μέθοδο της μέγιστης πιθανοφάνειας

$$\hat{\alpha}_{kk'} = \frac{n(k,k')}{n(k)} \tag{7}$$

όπου με $n(k,k')$ συμβολίζουμε το πλήθος των μεταβάσεων από την κατάσταση k στην κατάσταση k' και με $n(k)$ το πλήθος των εμφανίσεων (occurrences) της κατάστασης k.

Οι τεχνικές μέγιστης πιθανοφάνειας ενέχουν τον κίνδυνο απόδοσης μηδενικής πιθανότητας μετάβασης σε ορισμένα από τα ζεύγη - κίνδυνος που αυξάνεται λόγω του περιορισμένου διαθέσιμου προταξινομημένου υλικού. Έτσι, οι παραπάνω πιθανότητες απαιτούν εξομάλυνση, είτε μέσω ευρετικών κανόνων (heuristic rules) είτε μέσω Μπεϋσιανής μοντελοποίησης (χρήση Dirichlet εκ των προτέρων πιθανότητας μετάβασης στην κατάσταση k' δεδομένης της k). Ο εκτιμώμενος πίνακας μετάβασης χαρακτηρίζεται από μεγάλες πιθανότητες παραμονής στην ίδια κατάσταση (self-transition probability), της τάξης του $\alpha_{jj} = 1 - \gamma$, όπου $\gamma \approx 10^{-3}$ για κατάσταση ομιλίας και μια τάξη μεγέθους μικρότερη για τις άλλες πιθανότητες παραμονής, όπως σιωπές. Η αποκωδικοποίηση της ακολουθίας των καταστάσεων γίνεται μέσω του αλγορίθμου Viterbi. Τεχνικές μορφολογικού φιλτραλίσματος επιλέγονται συνήθως για τον διαχωρισμό μεταξύ δύο κλάσεων, όπως αυτόν μεταξύ των κλάσεως ομιλίας και σιωπής. Μία σημαντική διαφορά με τα HMMs είναι ότι τα μορφολογικά φίλτρα δέχονται ως είσοδο δυικές μεταβλητές (binary variables) δηλαδή τις ετικέτες κατάταξης των διανυσμάτων παρατήρησης, οι οποίες προέκυψαν από το GMM.

Προσθέτουμε, επίσης και μεθόδους οι οποίες εκμεταλλεύονται την έξοδο συστημάτων Αυτόματης Αναγνώρισης Φωνής ώστε να προβούν σε διαχωρισμό μεταξύ ομιλίας και λοιπών ηχητικών κατηγοριών, [40].

2.4.1 Μετρικές αξιολόγησης

Οι μετρικές με τις οποίες αξιολογείται η απόδοση ενός αλγορίθμου κατάτμησης είναι οι συνήθεις μετρικές εντοπισμού γεγονότων (event detection) και ανάκτησης πληροφορίας (information retrieval), δηλαδή με το ποσοστό ανάκτησης (recall) και την ακρίβεια ανάκτησης (precision). Οι μετρικές αυτές είναι ιδιαίτερα χρήσιμες και για τη ρύθμιση των παραμέτρων (tuning-calibration) του αλγορίθμου, κατά τη φάση εκμάθησης. Για μια συνολική αξιολόγηση του συστήματος, πολλές φορές χρησιμοποιείται και το F-measure, το οποίο ορίζεται ως ο γεωμετρικός μέσος του ποσοστού και της ακρίβειας ανάκτησης. Η κριτική στο F-measure είναι ότι αντιμετωπίζει τα δύο είδη λαθών ισοδύναμα, κάτι που δεν ισχύει στους αλγορίθμους βαθμίδας ΚΟΟ. Πιο συγκεκριμένα, ο εντοπισμός ενός μη υπαρκτού γεγονότος (False Alarm, FA) μπορεί να αναθεωρηθεί από τα επόμενα στάδια του αλγορίθμου, κάτι που δεν συμβαίνει με τον μη εντοπισμό ενός γεγονότος (Miss Detection, MD), το οποίο θεωρείται μη αναστρέψιμο σφάλμα για ένα πλήθος αλγορίθμων που χρησιμοποιούνται στην πράξη. Επομένως, η τάση που επικρατεί είναι η πόλωση του αλγορίθμου υπέρ των FA, έτσι ώστε η τελική σειριακή κατάτμηση στο τελευταίο να έχει όσο το δυνατόν λιγότερα σφάλματα MD. Με όρους ανάκτησης πληροφορίας, η στρατηγική αυτή ισοδυναμεί με πόλωση της εκτίμησης υπέρ του recall.

2.5 Αλγόριθμοι κατάτμησης

Όπως αναφέρθηκε και στην εισαγωγή, η βασική αλγοριθμική αντιμετώπιση είναι αυτή του ολισθαίνοντος παραθύρου το οποίο σαρώνει τα δεδομένα στον κατάλληλο παραμετρικό χώρο. Θεωρούμε αρχικά ένα παράθυρο $\mathcal{W}$ συνολικού μήκους $2L$. Τυπικές τιμές του L είναι 1 έως 4 sec. Το παράθυρο είναι τεμαχισμένο σε δύο υπο-παράθυρα $\mathcal{W}^s = (w_l^s, w_r^s)$ ούτως ώστε το καθένα από τα δύο τμήματα να μοντελοποιεί τα αντίστοιχα τμήματα που βλέπει κάθε ολίσθηση s. Έστω τ η χρονική ολίσθηση του παραθύρου, τυπικές τιμές της οποίας είναι 0.3 έως 0.7 sec. Κατά την s-οστή ολίσθηση, τα τμήματα που μοντελοποιεί καθένα από τα δύο υπο-παράθυρα μπορούν να εκφρασθούν

μέσω της συνάρτησης $\delta(\cdot,\cdot)$ του Kronecker

$$w_l^s = \{\mathbf{y}^{(i)} : \sum_{p=s\tau}^{s\tau+L-1} \delta(i,p) = 1\} \tag{8}$$

$$w_r^s = \{\mathbf{y}^{(i)} : \sum_{p=s\tau+L}^{s\tau+2L-1} \delta(i,p) = 1\} \tag{9}$$

Καθένα από τα υπο-παράθυρο μοντελοποιείται με μία κανονική κατανομή $\mathcal{N}(\mu_l^s, \Sigma_l^s)$ τις οποίες θα συμβολίζουμε ως $\theta_l^s = (\mu_l^s, \Sigma_l^s)$ για το w_l^s και αντίστοιχα για το w_r^s. Θεωρούμε εκτίμηση μέγιστης πιθανοφάνειας των παραμέτρων, οπότε οι παράμετροι είναι οι δειγματικοί μέσοι των υπο-παραθύρων. Η παραπάνω διαδικασία παράγει μια νέα ακολουθία $\mathbf{q} \equiv \{q_s\}_{s=0}^{S-1} \in \mathcal{Q}^S$, $\mathcal{Q} \subseteq \mathbb{R}$ η οποία μπορεί να αναπαραστεί ως $\mathcal{S} : \mathcal{Y}^n \mapsto \mathcal{Q}^S$. Το q_s είναι η απόκλιση παραμέτρων του στατιστικού μοντέλου των δύο υπο-παραθύρων $q_s = \mathcal{D}(\theta_l^s || \theta_r^s)$ κατά την s-οστή ολίσθηση. Η βασική διελκυστίνδα (trade-off) που προκύπτει για την επιλογή του μήκους του παραθύρου είναι μεταξύ της ομαλότητας της ακολουθίας των αποστάσεων (μεγάλο L) και της διακριτικής ικανότητας (resolution) που επιθυμούμε να έχει ο αλγόριθμος εντοπισμού και η οποία σχετίζεται με τις ελάχιστες αναμενώμενες διάρκειες των τμημάτων ομιλίας.

Πολλές από τις μετρικές που χρησιμοποιούνται στην πράξη παρουσιάστηκαν παραπάνω. Για έρευνα πάνω στην απόδοση των μετρικών αυτών καθώς και για εναλλακτικές μετρικές αποκλίσεις παραπέμπουμε στα [41], [42], [43], [44] και [45].

2.5.1 Αλγόριθμοι κατάτμησης διπλού περάσματος

Η τάση που επικρατεί στους αλγορίθμους κατάτμησης είναι αυτή του διπλού περάσματος (two-pass algorithms). Το πρώτο πέρασμα του αλγορίθμου - γνωστό και ως πέρασμα υπερ-κατάτμησης είναι υπεύθυνο για τον εντοπισμό του συνόλου των υποψηφίων σημείων εναλλαγής. Ο εντοπισμός των σημείων αυτών επιτυγχάνεται μέσω μετρικών χαμηλού υπολογιστικού φόρτου. Συνήθεις επιλογές είναι η συμμετρική KL απόσταση, αλλά και μετρικές βασισμένες στην απόκλιση των μέσων των

κατανομών, όπως η Hotelling T^2 στατιστική

$$T^2 = \frac{n_2 \times n_2}{n_1 + n_2}(\mu_1 - \mu_2)^T W^{-1}(\mu_1 - \mu_2) \tag{10}$$

όπου $W = \frac{n_1}{n_1+n_2}\Sigma_1 + \frac{n_2}{n_1+n_2}\Sigma_2$. Αυτό που απομένει είναι η εύρεση των τοπικών μεγίστων της $\mathbf{q}$, τα οποία θα αντιστοιχούν στα υποψήφια σημεία αλλαγής ηχητικής οντότητας. Ώς τοπικό μέγιστο, ορίζεται το σημείο εκείνο που πληροί τις παρακάτω ιδιότητες

- $q_s > q_{s+1}$
- $q_s > q_{s-1}$
- $q_s > th_i$

όπου το κατώφλι μπορεί να είναι και αυτό μεταβλητό, έτσι ώστε να ακολουθεί τις μεταβολές του ακουστικού περιβάλλοντος. Μία επιλογή μεταβλητού κατωφλίου είναι ο μέσος όρος των k προηγούμενων απόστασεων

$$th_s = \alpha \frac{1}{k} \sum_{s'=s-k}^{s-1} q_{s'} \tag{11}$$

για αλγορίθμους εντός γραμμής (φιλτράρισμα) και $\alpha = 1.2$. Αντίστοιχα, για επεξεργασία εκτός γραμμής, ο παραπάνω μέσος όρος μπορεί να είναι κεντραρισμένος γύρω από το υποψήφιο σημείο (εξομάλυνση).

Κατά το δεύτερο πέρασμα, ένα προς ένα τα υποψήφια σημεία επαληθεύονται, συνήθως μέσω του κριτηρίου ΔBIC. Για αύξηση της ακρίβειας εντοπισμού (η οποία περιορίζεται από την επιλογή της ολίσθησης παραθύρου τ), χρήσιμη είναι και η εύρεση του τοπικού μεγίστου σε μια περιοχή γύρω από αυτό το σημείο (boundary refinement). Η μέθοδος αυτή αυξάνει και πάλι την ακρίβεια του συνολικού αλγορίθμου κατάτμησης, από το μήκος ολίσθησης τ (τάξης 0.5 sec) στο διάνυσμα χαρακτηριστικών (0.01 sec).

Το υπολογιστικό κόστος σε σχέση με τον αλγόριθμο μονού περάσματος, μειώνεται σημαντικά, ιδιαίτερα με χρήση διαγώνιων πινάκων συμμεταβλητότητας κατά το πρώτο πέρασμα. Επιπλέον, η

ανάγκη για στιβαρή εκτίμηση των παραμέτρων οδηγεί σε αύξηση του μήκους των υπο-παραθύρων L που με τη σειρά της μειώνει την ανάλυση (resolution). Έτσι, σε περιβάλλον διαλόγου μεταξύ ομιλητών και γενικότερα συχνών εναλλαγών, όπου απαιτείται μεγάλη ανάλυση (L της τάξης του 0.5 έως 1.0 sec) καθίσταται αναγκαία μια τέτοια αλγοριθμική προσέγγιση. Η προσέγγιση αυτή καλείται DISTBIC, και προτάθηκε στο [46].

Συγγενής είναι και η τεχνική που παρουσιάζεται στο [47]. Η μετρική απόστασης είναι ο Γενικευμένος Λόγος πιθανοφάνειας, και η βασική ιδέα είναι ο έλεγχος της ολίσθησης τ του παραθύρου, έτσι ώστε να μικραίνει κοντά στα μέγιστα της μετρικής. Η τεχνική αυτή καλείται αλγόριθμος εστιασμένου εντοπισμού (Localized Search Algorithm) και χρησιμοποιείται κυρίως σε εφαρμογές εντός γραμμής (on-line).

Μία εναλλακτική προσέγγιση στην κατάτμηση είναι αυτή του παραθύρου μεταβλητού μήκους ([48]). Το παράθυρο εδώ είναι ενιαίο - επομένως το σημείο εναλλαγής μπορεί να είναι οποιοδήποτε εντός των παρατηρήσεων που σαρώνει (εκτός από κάποιες περιοχές κοντά στα άκρα όπου δεν μπορεί να υπάρξει στιβαρή εκτίμηση παραμέτρων). Όσο δεν παρουσιάζεται σημείο αλλαγής εντός του, το παράθυρο επεκτείνεται κατά ΔN_i παρατηρήσεις. Για να επιταχυνθεί η διαδικασία, η αύξηση του μήκους του παραθύρου ΔN_i βαίνει και αυτή αυξανόμενη, αφού $\Delta N_{i+1} = \Delta N_i + \delta_{i+1}$, όπου $\delta_{i+1} = 2\delta_i$. Παρόμοιες τεχνικές παραθύρου μεταβλητού μήκους παρουσιάζονται και στο [44] με ανάλυση στην απαιτούμενη πολυπλοκότητα των υπολογισμών, καθώς και στο [49], όπου εξετάζεται η χρήση δυναμικού προγραμματισμού για την εύρεση των σημείων εναλλαγής.

2.6 Ομαδοποίηση τμημάτων ομιλίας

2.6.1 Εισαγωγικά

Εξετάζουμε τώρα μία από τις πλέον βασικές βαθμίδες των αλγορίθμων ΚΟΟ. Ο ρόλος της βαθμίδας αυτής είναι να ενοποιήσει τα τμήματα που προέκυψαν από την παραπάνω βαθμίδα σε διακριτές (ή αλλιώς ομογενείς) ηχητικές οντότητες. Στην περίπτωση που ενδιαφερόμαστε μόνο για τμήματα

ομιλίας, οι οντότητες αυτές αντιστοιχούν σε διακριτούς ομιλητές. Στο στάδιο αυτό, είναι συνήθης η παραδοχή να θεωρείται ως διακριτός ομιλητής ο συνδυασμός ομιλητή και ηχητικού περιβάλλοντος (acoustic environment). Η παραδοχή αυτή επιτρέπει στη μοντελοποίηση τη χρήση μη κανονικοποιημένων διανυσμάτων παρατήρησης. Οι κανονικοποιήσεις των διανυσμάτων χαρακτηριστικών, παρότι ιδιαίτερα χρήσιμες σε εφαρμογές ταυτοποίησης ομιλητών υπό μη ταυτόσημες συνθήκες ηχογράφησης (miss-matched recording conditions) γενικά αφαιρούν και χαρακτηριστικά του ομιλητή. Η πλέον χρησιμοποιούμενη μοντελοποίηση για τη συγκεκριμένη βαθμίδα είναι αυτή της μιας κανονικής κατανομής ανά ομάδα (cluster). Με τον όρο ομάδα, αναφερόμαστε στο σύνολο των διανυσμάτων παρατήρησης τα οποία έχουν αποδοθεί σε κοινή ηχητική οντότητα ή ετικέτα (label). Οι ετικέτες αυτές δεν είναι τίποτε άλλο από συμβολικά ονόματα, τα οποία περιλαμβάνουν τον αύξοντα αριθμό - συνήθως με τη χρονική σειρά εμφάνισης - καθώς και τη μακροσκοπική κλάση στην οποία έχει καταταχθεί από το στάδιο κατηγοριοποίησης. Θα πρέπει ωστόσο να τονισθεί ότι στο στάδιο αυτό, η ελάχιστη αδιαίρετη οντότητα είναι τα τμήματα $\{\mathbf{y}_k\}_{k=1}^{N}$ και όχι το διάνυσμα παρατήρησης $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^{n}$.

2.6.2 Ο αλγόριθμος της ιεραρχικής ομαδοποίησης

Ο αλγόριθμος της συγκολλητικής ιεραρχικής ομαδοποίησης (Agglomerative Hierarchical Clustering, AHC) είναι ένας από τους πλέον διαδεδομένους αλγορίθμους ομαδοποίησης ενός συνόλου παρατηρήσεων $\mathcal{D} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\}$ σε K ομάδες, όταν το K είναι άγνωστο. Θεωρούμε αρχικά ότι το σύνολο παρατηρήσεων δεν έχει υποστεί κατάτμηση και στη συνέχεια θα επεκτείνουμε τον αλγόριθμο ώστε να αντιπετωπίζει το πρόβλημα της ομαδοποίησης τμημάτων ομιλίας. Ο αλγόριθμος της συγκολλητικής ιεραρχικής ομαδοποίησης είναι ο ακόλουθος

- **Είσοδος:** δεδομένα $\mathcal{D} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}\}$,
  μετρική απόστασης $d(\cdot, \cdot)$

- **Αρχικοποίηση:** αριθμός ομάδων $c = n$,
  $\mathcal{D}_i = \{x^{(i)}\}, \forall i = 1, \ldots, n$

- **Αναδρομή: while** $c > 1$ **do**

  - Εύρεση του ζεύγους με τη μικρότερη απόσταση $d(\mathcal{D}_i, \mathcal{D}_j)$

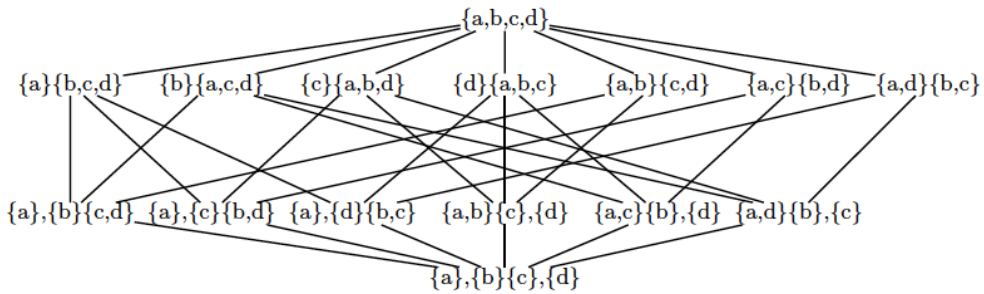
  - Ένωση $\mathcal{D}_k \leftarrow \mathcal{D}_i \cup \mathcal{D}_j, T_k \leftarrow (T_i, T_j)$

  - Διαγραφή των $\mathcal{D}_i$ και $\mathcal{D}_j$, $c \leftarrow c - 1$

- **end while**

- **Έξοδος:** το δεντρόγραμμα T

Το δεντρόγραμμα είναι ένα μονοπάτι πάνω στο πλέγμα των ομαδοποιήσεων (clustering lattice). Το πλέγμα αυτό απεικονίζεται στο Σχ. 3 για ένα σύνολο με τέσσερις παρατηρήσεις $\mathcal{D} = \{a, b, c, d\}$.



Σχῆμα 3: *Το πλέγμα των ομαδοποιήσεων για ένα σύνολο αποτελούμενο από τέσσερα στοιχεία.*

Μία ομαδοποίηση $\mathbf{x}$ θα καλείται *βασικής φόρμας* όταν ο δείκτης κάθε ομάδας αποδίδεται με σειρά εμφάνισης και με αύξηση κατά ένα. Έτσι, για $L = 5$, παραδείγματα ομαδοποίησης βασικής φόρμας είναι τα $\mathbf{x} = \{1, 1, 2, 3, 3\}$, $\mathbf{x} = \{1, 2, 2, 3, 3\}$, $\mathbf{x} = \{1, 2, 3, 4, 5\}$, κ.ο.κ. σε αντίθεση με το $\mathbf{x} = \{2, 2, 2, 3, 1\}$.

Η απόσταση μεταξύ $d(\mathcal{D}_i, \mathcal{D}_j)$ χρειάζεται περαιτέρω διερεύνηση. Αρχικά, πρέπει να ορισθεί η απόσταση μεταξύ μεμονωμένων παρατηρήσεων. Επιλογές είναι η Ευκλείδια, η Μανχάταν, η Minkowski, η Mahalanobis και άλλες. Ταυτόχρονα, ελλείψει παραμετρικού μοντέλου που περιγράφει τα στατιστικά της κάθε ομάδας, θα πρέπει να ορισθεί η απόσταση μεταξύ συνόλων σημείων, γνωστή ως σύνδεση (linkage). Συνήθεις επιλογές σύνδεσης είναι η μικρότερη σύνδεση μεταξύ δύο μελών (single linkage), η μεγαλύτερη (complete linkage), η μέση (average linkage), η ενδιάμεση (median linkage) και η σύνδεση Ward (ελαχιστοποίηση του αθροίσματων των τετραγωνικών αποστάσεων). Με βάση τους ορισμούς της απόστασης και σύνδεσης, το δεντρόγραμμα T μπορεί να κατατμηθεί έτσι ώστε η ομαδοποίηση που θα επιλεγεί να πληροί την ιδιότητα όλες οι ανά δύο αποστάσεις να είναι μεγαλύτερες από το κατώφλι. Υπό αυτή την αλγοριθμική στρατηγική, το σύνολο των αποστάσεων που πρέπει να υπολογισθούν είναι $n(n-1)/2$ (λόγω της συμμετρίας της απόστασης) και η όλη διαδικασία βασίζεται στον πίνακα αποστάσεων.

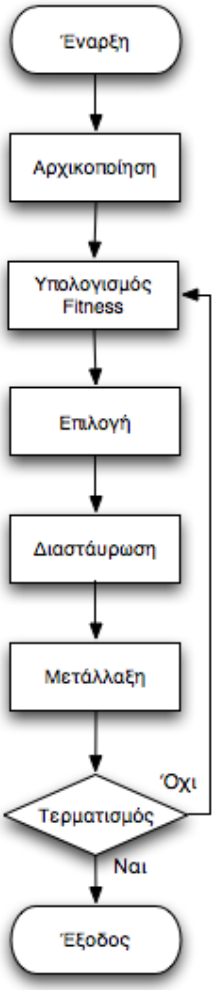
Στο πρόβλημα της ομαδοποίησης $\Xi = \{\mathbf{y}_k\}_{k=1}^{L}$ τμημάτων ομιλίας σε K ομιλητές, ο παραπάνω αλγόριθμος χρειάζεται μερικές τροποποιήσεις. Κατά την ενοποίηση δύο ομάδων, εκτός από την επιλογή των κλασσικών μεθόδων σύνδεσης που αναφέραμε, υπάρχει και η δυνατότητα επανυπολογισμού των επαρκών στατιστικών της νέας ομάδας και των αποστάσεών της από τις υπόλοιπες ομάδες. Αν και οι κλασσικές μέθοδοι σύνδεσης λειτουργούν και στο πρόβλημα της ομαδοποίησης τμημάτων ομιλίας, η τελευταία στρατηγική έχει επικρατήσει. Έτσι, ο ορισμός της απόστασης $d(\mathcal{D}_i, \mathcal{D}_j)$ εξαντλείται στην επιλογή μεταξύ των στατιστικών αποκλίσεων, χωρίς την ανάγκη ορισμού της μεθόδου σύνδεσης (linkage). Αρχικά, λοιπόν, θα πρέπει να υπολογισθούν οι επαρκείς στατιστικές της νέας ομάδας, και έπειτα η απόστασή της με τις κατανομές των υπολοίπων ομάδων. Το κόστος αυτό είναι αμελητέο για μοντελοποίηση με απλές κανονικές κατανομές (λόγω ύπαρξης κλειστής φόρμας υπολογισμού), αλλά είναι σημαντικό για μοντελοποίηση με μίγματα κανονικών

κατανομών. Τέλος, η εξαγωγή του πλήρους δενδρογράμματος πολλές φορές αποφεύγεται. Η διαδικασία σταματάει όταν όλες οι ανά δύο αποστάσεις υπερβαίνουν το κατώφλι απόφασης.
Η συνηθέστερη επιλογή απόστασης είναι το τοπικό ή το ολικό ΔBIC. Μετά από σειρά πειραμάτων, το τοπικό έχει αποδειχθεί ότι παρουσιάζει πιο ακριβή αποτελέσματα σε σχέση με το ολικό, [31], [50], [51].

2.7 Εναλλακτικές προσεγγίσεις της ομαδοποίησης

2.7.1 Τα μειονεκτήματα της ιεραρχικής ομαδοποίησης

Ένα από τα βασικά προβλήματα της ιεραρχικής ομαδοποίησης είναι η ευαισθησία της ως προς τα ενδεχόμενα σφάλματα, τα οποία μπορούν να εμφανισθούν σε κάποια από τις αναδρομές. Τέτοια σφάλματα διαχέονται ως στην τελική ομαδοποίηση αφού ο αλγόριθμος δεν έχει τρόπο να αναθεωρήσει μία εσφαλμένη ενοποίηση. Το πρόβλημα εμφανίζεται και στις δύο εκδοχές της ιεραρχικής ομαδοποίησης, αφού και στη διαιρετική (divisive, top-down) προσέγγιση η εσφαλμένη διάσπαση μίας ομάδας είναι μη αναστρέψιμη.
Μία εναλλακτική μέθοδος είναι η χρήση γενετικών αλγορίθμων. Οι γενετικοί αλγόριθμοι επιχειρούν να βρουν το βέλτιστο μίας αντικειμενικής συνάρτησης, όταν η συνάρτηση αυτή είναι μη διαφορίσιμη (επομένως οι κλασσικές μέθοδοι βελτιστοποίησης δεν μπορούν να εφαρμοσθούν) ενώ ταυτόχρονα το πεδίο τιμών είναι απαγορευτικά μεγάλο ώστε να υπολογισθεί η συνάρτηση σε κάθε μέλος του πεδίου (exhaustive search). Υλοποιήσεις γενετικών αλγορίθμων στην ομαδοποίηση τμημάτων ομιλίας εξετάζονται στο [52].
Το διάγραμμα ροής των γενετικών αλγορίθμων παρουσιάζεται στο Σχ. 4. Περιληπτικά, κάθε ομαδοποίηση θεωρείται και ένα χρωμόσωμα $\mathbf{g} \in \mathcal{G}$, όπου $\mathcal{G}$ ο χώρος των ομαδοποιήσεων βασικής φόρμας (Baseform space). Η συνάρτηση καταλληλότητας (fitness) παίζει τον ρόλο της αντικειμενικής συνάρτησης. Συνήθεις επιλογές είναι η ολική πιθανοφάνεια με έναν όρο που ποινικοποιεί την πολυπλοκότητα της μοντελοποίησης. Το ολικό BIC είναι μια τέτοια επιλογή, αλλά όχι η μόνη. Για



Σχῆμα 4: *Γενικό διάγραμμα ροής Γενετικών Αλγορίθμων*

τη σύγκριση της απόδοσης των γενετικών αλγορίθμων σε σχέση με την ιεραρχική ομαδοποίηση, καθώς και μεταξύ των εναλλακτικών μοντελοποιήσεων (επιλογή συνάρτησης καταλληλότητας, κατανομής πιθανότητας μετάλλαξης, κ.ο.κ.) παραπέμπουμε στο [52].

Αναφερόμαστε τέλος σε μία πολύ πετυχημένη προσέγγιση του προβλήματος της κατάτμησης με χρήση γενετικών αλγορίθμων που παρουσιάζεται στο [53]. Η μέθοδος αυτή χρησιμοποιεί ως συνάρτηση καταλληλότητας την από κοινού πληροφορία (Mutual Information) και εξετάζονται

οι προσεγγίσεις της που εμφανίστηκαν στην Ανάλυση Ανεξαρτήτων Συνιστωσών (Independent Component Analysis, ICA, [54]). Η προτεινόμενη μέθοδος παρουσιάζει αυξημένη ακρίβεια στον εντοπισμό σημείων εναλλαγής σε σχέση με τις καθιερωμένες μεθόδους κατάτμησης.
Στο [55] ίδια ομάδα ερευνητών εξετάζει τη χρήση της μεθόδου εξομοιούμενης ανόπτησης simulated annealing, μέθοδος η οποία παρουσιάζει αρκετά κοινά με τους γενετικούς αλγορίθμους. Τα αποτελέσματα δείχνουν περαιτέρω βελτίωση σε σχέση με τους γενετικούς αλγορίθμους στις ίδιες βάσεις αξιολόγησης. Η επιτυχία των μεθόδων αυτών τόσο στην κατάτμηση όσο και στην ομαδοποίηση αποδεικνύει τα πλεονεκτήματα των εξελικτικών (evolutionary) μεθόδων έναντι των συμβατικότερων μεθόδων αποσυζευγμένης κατάτμησης και ομαδοποίησης, ιδίως αυτών που έχουν ως βάση την ιεραρχική ομαδοποίηση.

2.8 Στάδιο μετεπεξεργασίας

Μετά το πέρας της ομαδοποίησης, πολλές υλοποιήσεις περιλαμβάνουν και ένα στάδιο μετεπεξεργασίας. Το στάδιο αυτό επιτελεί τρεις κατά βάση σκοπούς.

1. Ο πρώτος είναι να αναθεωρούν τα όρια των τμημάτων που προέκυψαν από τη βαθμίδα κατάτμησης. Η αναθεώρηση αυτή γίνεται συνήθως με χρήση Υπονοούμενων Μαρκοβιανών Μοντέλων (HMMs) και του αλγορίθμου Viterbi για την εύρεση του βέλτιστου μονοπατιού. Ως καταστάσεις του HMMs θεωρούμε και πάλι τις διακριτές ηχητικές οντότητες, όπως αυτές προέκυψαν από το στάδιο της ομαδοποίησης, συμπεριλαμβανομένης και της σιωπής. Κάθε κατάσταση του HMM μοντελοποιείται με ένα GMM 8 συντελεστών με εκτίμηση μέγιστης πιθανοφάνειας (ML) βάσει των παρατηρησεων και της ομαδοποίησης. Σε αντίθεση με την εκτίμηση ML, ο πίνακας μεταβάσεων δεν χρησιμοποιεί την πληροφορία της ομαδοποίησης και το κόστος μετάβασης τίθεται σταθερό για όλες τις δυνατές μεταβάσεις και μηδενικό για παραμονή στην ίδια κατάσταση. Με βάση αυτήν τη μοντελοποίηση, το βέλτιστο μονοπάτι αποκωδικοποιείται με τον αλγόριθμο Viterbi.

2. Ο δεύτερος σκοπός είναι η περαιτέρω ενοποίηση των ομάδων εκείνων οι οποίες δεν συνενώθηκαν κατά την ομαδοποίηση. Η μέθοδος που χρησιμοποιείται για τον σκοπό αυτό είναι η Μπεϋσιανή προσαρμογή των παραμέτρων ενός γενικού μοντέλου ομιλίας στα δεδομένα των ομάδων και ο αλγόριθμος που συνήθως χρησιμοποιείται είναι και πάλι η ιεραρχική ομαδοποίηση.

3. Ένας τρίτος σκοπός του σταδίου είναι ο εντοπισμός των ταυτόχρονων ομιλιών. Οι ταυτόχρονες ομιλίες αποτελούν ένα συχνό φαινόμενο, τόσο στις εκπομπές ειδησεογραφικού περιεχομένου, όσο και στο πεδίο των συσκέψεων. Σε ολοκληρωμένα συστήματα ΚΟΟ και αναγνώρισης φωνής, η αντιμετώπιση των επικαλύψεων βασίζεται στο κείμενο αυτόματης απομαγνητοφώνησης (transcript) και στα ενδιάμεσα στάδιά του (χαμηλός βαθμός αξιοπιστίας μηχανής ASR στις συγκεκριμένες περιοχές, [56]. Η πρώτη ολοκληρωμένη απόπειρα εντοπισμού των επικαλύψεων για μονοφωνικά αρχεία ήχου χωρίς τη χρήση μηχανής ASR παρουσιάσθηκε το 2008 στο [57] για το πεδίο των συσκέψεων.

3 ΕΙΣΑΓΩΓΗ ΣΤΗ ΜΠΕΫΣΙΑΝΗ ΣΤΑΤΙΣΤΙΚΗ ΚΑΙ ΤΙΣ ΟΙΚΟΓΕΝΕΙΕΣ ΤΩΝ ΕΚΘΕΤΙΚΩΝ ΚΑΤΑΝΟΜΩΝ

3.1 Περίληψη κεφαλαίου

Στο κεφάλαιο αυτό θα παρουσιάσουμε ορισμένες από τις βασικές αρχές και θεωρήματα της Μπεϋσιανής Στατιστικής καθώς και των οικογενειών των εκθετικών κατανομών. Συγκεκριμένα, θα αναλύσουμε τη μεθοδολογία με την οποία η Μπεϋσιανή στατιστική αντιμετωπίζει τα δύο βασικά επίπεδα συμπερασματολογίας (inference), δηλαδή την εκτίμηση παραμέτρων και την επιλογή τάξης μοντέλου. Κατόπιν, θα εξετάσσουμε τις οικογένειες των εκθετικών κατανομών. Οι οικογένειες αυτές περικλείουν τις πλέον διαδεδομένες κατανομές και θα αποτελέσουν κομβικό άξονα στη διατριβή.

3.2 Εκτίμηση παραμέτρων και τάξης μοντέλου στη Μπεϋσιανή στατιστική

Αν και ο όρος Μπεϋσιανή στατιστική προέρχεται από τη συχνότατη χρήση του κανόνα του Thomas Bayes,

$$P(X = x|Y = y) = P(Y = y|X = x)\frac{P(X = x)}{P(Y = x)} \tag{12}$$

η χρήση του εν λόγω κανόνα είναι ανεξάρτητη από τη σχολή στατιστικής. Το πιο σωστό σημείο για την κατανόηση της διαφοροποίησης μεταξύ των δύο σχολών είναι ο τρόπος με τον οποίον αντιμετωπίζονται οι παράμετροι και τα μοντέλα. Εστιάζοντας αρχικά στο πρώτο επίπεδο συμπερασματολογίας (εκτίμηση παραμέτρων) η βασική διαφορά έγκειται στο ότι ενώ η κλασσική σχολή αντιμετωπίζει τις παραμέτρους σαν άγνωστες σταθερές και μεγιστοποιεί συνήθως την συνάρτηση πιθανοφάνειας για τον υπολογισμό τους, η Μπεϋσιανή σχολή τις αντιμετωπίζει σαν τυχαίες μεταβλητές. Με τον τρόπο αυτό καθορθώνει να ποσοτικοποιήσει την αμφιβολία για τις παραμέτρους

της μοντελοποίησης, εφαρμόζοντας με συνέπεια τα βασικά θεωρήματα των πιθανοτήτων. Ως εκ τούτου, η Μπεϋσιανή στατιστική παρέχει τη δυνατότητα μίας ολοκληρωμένης πιθανοτικής αντιμετώπισης του προβλήματος, με τον καθορισμό της εκ των προτέρων κατανομής των παραμέτρων (δηλαδή προτού οι παρατηρήσεις του συγκεκριμένου πειράματος γίνουν διαθέσιμες) και της εκ των υστέρων κατανομής (δηλαδή αφού οι παρατηρήσεις γίνουν διαθέσιμες). Με τον τρόπο αυτόν, η εκ των υστέρων κατανομή περιέχει το σύνολο της γνώσης μας ως προς τις παραμέτρους, ως συνδυασμός τόσο των παρατηρήσεων όσο και της εκ των προτέρων γνώσης μας για αυτές.
Στις περιπτώσεις όπου οποιαδήποτε εκ των προτέρων γνώση για τις παραμέτρους δεν είναι διαθέσιμη, η εκ των προτέρων κατανομή επιλέγεται με τρόπο ώστε να επηρεάζει όσο το δυνατόν λιγότερο την εκ των υστέρων κατανομή. Λόγω της κεφαλαιώδους σημασίας της στη Μπεϋσιανή στατιστική, η ανάλυση των εκ των προτέρων κατανομών θα παρουσιασθεί σε ειδική παράγραφο.

3.2.1 Εκτίμηση παραμέτρων

Έστω λοιπόν ένα σύνολο παρατηρήσεων $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^n$ και ένα μοντέλο M σταθερής τάξης το οποίο παραμετροποιείται από τις παραμέτρους $\theta \in \Theta$. Η εφαρμογή του κανόνα του Bayes οδηγεί στον παρακάτω τύπο

$$\pi(\theta|\mathbf{y}, M_k) = \frac{p(\mathbf{y}|\theta, M_k)\pi(\theta|M_k)}{p(\mathbf{y}|M_k)} \tag{13}$$

όπου $\pi(\theta|\mathbf{y}, M_k)$ η εκ των υστέρων συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) του θ, $\pi(\theta|M_k)$ η εκ των προτέρων σ.π.π. του θ και $p(\mathbf{y}|\theta, M_k)$ η συνάρτηση πιθανοφάνειας του θ. Η ποσότητα $p(\mathbf{y}|M)$ καλείται ολοκληρωμένη πιθανοφάνεια (integrated likelihood), υπολογίζεται ως

$$p(\mathbf{y}|M_k) = \int_{\theta\in\Theta} p(\mathbf{y}|\theta, M_k)\pi(\theta|M_k)d\theta \tag{14}$$

και ως εκ τούτου είναι ανεξάρτητη της τιμής του θ. Έτσι, σε περιπτώσεις που αρκεί η εκτίμηση-σημείου του θ (όπως η μέγιστη εκ των υστέρων πιθανότητα), ο αριθμητής της (13) επαρκεί. Όπως θα δείξουμε όμως στη συνέχεια, η ολοκληρωμένη πιθανοφάνεια του μοντέλου παίζει σημαντικότατο ρόλο στο δεύτερο επίπεδο συμπερασματολογίας, δηλαδή στην επιλογή τάξης μοντέλου.

Εκτίμηση σημείου

Όπως αναφέραμε, στόχος της Μπεϋσιανής στατιστικής είναι η εξαγωγή της εκ των υστέρων κατανομής των στατιστικών ποσοτήτων ως προς τις όποιες επιθυμούμε να συμπερασματολογήσουμε (infer). Σε πολλά προβλήματα ωστόσο αυτό δεν αρκεί και θα πρέπει να προβούμε σε μια τελική *εκτίμηση σημείου*. Ως εκτίμηση σημείου αναφερόμαστε στην πρόβλημα εξαγωγής της καταλληλότερης τιμής των ποσοτήτων αυτών με βάση την εκ των υστέρων κατανομή, σύμφωνα με ένα κριτήριο καταλληλότητας. Έτσι, η Μπεϋσιανή εκτίμηση-σημείου των παραμέτρων θ βασίζεται στην (μη κατ' ανάγκη κανονικοποιημένη) εκ των υστέρων κατανομή $\pi(\theta|\mathbf{y}, M_k) \propto p(\mathbf{y}|\theta, M_k)\pi(\theta|M_k)$ και στην επιλεγμένη συνάρτηση κόστους $C(\hat{\theta}, \theta)$. Τυπικές συναρτήσεις κόστους είναι η τετραγωνική $C(\hat{\theta}, \theta) \propto |\hat{\theta} - \theta|^2$ και η ομοιόμορφη σε μια περιοχή 2ϵ

$$C(\hat{\theta}, \theta) = \begin{cases} 1, \, |\hat{\theta} - \theta| > \epsilon \\ 0, \, \text{αλλού} \end{cases} \tag{15}$$

Η τετραγωνική συνάρτηση κόστους οδηγεί στην εκτίμηση ελάχιστου μέσου τετραγωνικού σφάλματος (MMSE), που ορίζεται ως

$$\hat{\theta} = \int_{\theta \in \Theta} \theta \, \pi(\theta|\mathbf{y}, M_k) d\theta \tag{16}$$

και όπως παρατηρούμε ισούται με την αναμενόμενη τιμή της θ, με κατανομή αναφοράς την $\pi(\theta|\mathbf{y}, M_k)$. Η εκτίμηση αυτή είναι κατάλληλη κυρίως για μονοτροπικές εκ των υστέρων κατανομές. Σε διαφορετική περίπτωση, προτιμάται η εκτίμηση μέγιστης εκ των υστέρων πιθανότητας (MAP)

$$\hat{\theta} = \underset{\theta}{\text{argmax}}\{\pi(\theta|\mathbf{y}, M_k)\} \tag{17}$$

που προκύπτει από την ομοιόφορφη συνάρτηση κόστους για $\epsilon \to 0$. Οι δύο αυτές εκτιμήσεις σημείου ταυτίζονται όταν η $\pi(\theta|\mathbf{y}, M_k)$ είναι συμμετρική και παρουσιάζει το μέγιστο στη μέση τιμή. Η επιλογή της MAP ως συνάρτησης κόστους είναι πιο συμβατή με προβλήματα που ο κίνδυνος μη επιτυχίας του στόχου είναι σημαντικότερος παράγοντας από την αναμενόμενη της διακύμανσης σε βάθος πολλών πειραμάτων.

3.2.2 Επιλογή κατάλληλης τάξης μοντέλου

Ερχόμαστε τώρα στο δεύτερο επίπεδο συμπερασματολογίας, αυτό της επιλογής της κατάλληλης τάξης του μοντέλου. Τα παρακάτω παραδείγμα είναι ενδεικτικά προβλημάτων επιλογής ταξης μοντέλου.

1. επιλογή τάξης μοντέλων παλινδρόμησης και αυτοπαλινδρόμησης,
2. επιλογή αριθμού συντελεστών σε μίγματα κανονικών κατανομών,
3. επιλογή αριθμού καταστάσεων σε υπονοούμενα Μαρκοβιανά μοντέλα,
4. επιλογή τάξης πολυβηματικής Μαρκοβιανής αλυσίδας.

Όπως ήδη αναφέραμε, στη Μπεϋσιανή σχολή επιχειρούμε να εκφράσουμε την εκ των υστέρων κατανομή των παραμέτρων που επιθυμούμε να εκτιμήσουμε, δεσμεύοντάς την ως προς τα δεδομένα και ολοκληρώνοντας ως προς τις άγνωστες μεταβλητές. Στην προκειμένη περίπτωση, επιχειρούμε να συμπερασματολογήσουμε επί της τάξης k ενός μοντέλου $\mathcal{M} = \{M_k\}_{k=1,2,\dots}, K$. Ως εκ τούτου, θα πρέπει να δεσμεύσουμε την πιθανότητα ως προς τις παρατηρήσεις $\mathbf{y}$ και να ολοκληρώσουμε ως προς τις παραμέτρους $\theta \in \Theta$. Η εκ των υστέρων συνάρτηση μάζας πιθανότητας γράφεται ως

$$\pi(M_k|\mathbf{y}, \mathcal{M}) = \frac{p(\mathbf{y}|M_k)\pi(M_k|\mathcal{M})}{\pi(\mathbf{y}|\mathcal{M})} \propto p(\mathbf{y}|M_k)\pi(M_k|\mathcal{M}) \tag{18}$$

Αντίστοιχα με το πρώτο επίπεδο συμπερασματολογίας, ο παρονομαστής

$$\pi(\mathbf{y}|\mathcal{M}) = \sum_{k=1}^{K} p(\mathbf{y}|M_k)\pi(M_k|\mathcal{M}) \tag{19}$$

είναι ανεξάρτητος της τάξης του μοντέλου. Είναι όμως η κυρίαρχη στατιστική ποσότητα στο τρίτο επίπεδο συμπερασματολογίας, αυτή της επιλογής μοντέλου.

Βλέπουμε λοιπόν ότι η επιλογή της τάξης μοντέλου κυριαρχείται από την ολοκληρωμένη πιθανοφάνεια $p(\mathbf{y}|M_k)$ που δίνεται από την (14). Η συνεισφορά της εκ των προτέρων πυκνότητα μάζας πιθανότητας $\pi(M_k|\mathcal{M})$ είναι συνήθως πολύ μικρή σε σχέση με της $p(\mathbf{y}|M_k)$, καθώς η πρώτη παραμένει σταθερή καθώς ο αριθμός των παρατηρήσεων αυξάνεται.

3.3 Οι εκ των προτέρων κατανομές

Ερχόμαστε τώρα σε ένα κομβικό ζήτημα, αυτό της επιλογής της εκ των προτέρων σ.π.π. των παραμέτρων ενός μοντέλου ορισμένης τάξης $\pi(\theta|M_k)$. Όπως δείξαμε παραπάνω, η εκ των υστέρων σ.π.π. της θ είναι το κανονικοποιημένο γινόμενο της πιθανοφάνειας και της εκ των προτέρων σ.π.π., $\pi(\theta|\mathbf{y}, M_k) \propto p(\mathbf{y}|\theta, M_k)\pi(\theta|M_k)$. Επιπλέον, η ολοκληρωμένη πιθανοφάνεια $p(\mathbf{y}|M_k)$, πάνω στην οποία βασίζεται η επιλογή τάξης μοντέλου και αυτή συνάρτηση της $\pi(\theta|M_k)$. Οι εκ των προτέρων κατανομές μπορούν να κατηγοριοποιηθούν στις ακόλουθες μη κατ΄ ανάγκη αλληλοαποκλειούμενες κατηγορίες:

1. Αντικειμενικές εκ των προτέρων κατανομές (objective priors)
2. Υποκειμενικές εκ των προτέρων κατανομές (subjective priors)
3. Ιεραρχικές εκ των προτέρων κατανομές (hierarchical priors)
4. Συζυγείς εκ των προτέρων κατανομές (conjugate priors)

3.3.1 Αντικειμενικές εκ των προτέρων κατανομές

Η χρήση των κατανομών αυτών στοχεύει στη μέγιστη δυνατή εξάλειψη του υποκειμενικού παράγοντα στη Μπεϋσιανή συμπερασματολογία. Είναι η φυσική συνέχεια της χρήσης ομοιόμορφης εκ των προτέρων πυκνότητας μάζας πιθανότητας σε όλα τα ενδεχόμενα αποτελέσματα.
Τεράστια ώθηση στην αντικειμενική Μπεϋσιανή σχολή δόθηκε από την εργασία του Jeffreys, [58]. Πιο συγκεκριμένα, ο Jeffreys πρότεινε την εξής γενική αρχή για επιλογή αντικειμενικών εκ των προτέρων κατανομών

$$\pi(\theta|M_k) \propto |\mathcal{I}(\theta)|^{1/2} \tag{20}$$

με υπονοούμενη κανονικοποιητική σταθερά $C_{\mathrm{J}} = \int_{\theta\in\Theta} |\mathcal{I}(\theta)|^{1/2} d\theta$ και

$$\mathcal{I}(\theta)_{i,j} = -\int p(\mathbf{y}|\theta, M_k) \frac{\partial^2}{\partial_i \partial_j} \log p(\mathbf{y}|\theta, M_k) d\mathbf{y} = \mathcal{E}_{p(\mathbf{y}|\theta,M_k)} \left\{ -\frac{\partial^2}{\partial_i \partial_j} \log p(\mathbf{y}|\theta, M_k) \right\} \quad (21)$$

ορίζει το (i, j) στοιχείο του πίνακα πληροφορίας του Fisher, ενώ $1 \leq i, j \leq \dim(\theta)$.

Η σημαντικότερη ιδιότητα της παραπάνω κατανομής είναι ότι παραμένει αμετάβλητη στους μετασχηματισμούς $\theta \mapsto \varphi$. Η ιδιότητα αυτή είναι θεμελιώδης, αφού μία επίπεδη (δηλαδή σταθερά) κατανομή στην παραμετροποίηση θ δεν διατηρείται σταθερά με μιά εναλλακτική παραμετροποίηση φ της ίδιας συνάρτησης κατανομής. Δεδομένου λοιπόν ότι οι θεμελιώδεις στατιστικές έννοιες είναι οι κατανομές και όχι οι εναλλακτικές παραμετροποιήσεις τους, ο στόχος θα πρέπει να είναι η ομοιομορφία ως προς την ίδια την κατανομή, ανεξαρτήτως της παραμετροποίησης που επιλέγεται ώστε να εκφρασθεί η κατανομή μαθηματικά. Έτσι, η τετραγωνική ρίζα του όγκου της καμπυλότητας στο θ αίρει τη στρεύλωση του χώρου που προκαλεί η οποιαδήποτε παραμετροποίηση της κατανομής. Για τον λόγο αυτόν, ορίζουμε το λεγόμενο στοιχείο πληροφορίας (information element) ως $d\mathcal{V} = |\mathcal{I}(\theta)|^{1/2} d\theta$. Η γραφή αυτή έχει το πλεονέκτημα ότι είναι αμετάβλητη υπό μετασχηματισμούς, αφού για μετασχηματισμό $\theta \mapsto \varphi$ θα έχουμε

$$d\mathcal{V} = |\mathcal{I}(\varphi)|^{1/2} d\varphi = \left(|\mathcal{I}(\theta)|^{1/2} \left| \frac{\partial \theta}{\partial \varphi} \right| \right) \left(d\theta \left| \frac{\partial \varphi}{\partial \theta} \right| \right) = |\mathcal{I}(\theta)|^{1/2} d\theta \quad (22)$$

Το παραπάνω στοιχείο $d\mathcal{V}$ ισούται λοιπόν με το στοιχείο όγκου του manifold των κατανομών και είναι ανεξάρτητο της παραμετροποίησης.

Το εγγενές πρόβλημα των εκ των προτέρων κατανομών του Jeffreys είναι ότι δεν οδηγούν πάντοτε σε κανονικοποιήσιμες κατανομές, καθώς το ολοκλήρωμα $C_{\mathrm{J}} = \int_{\theta\in\Theta} |\mathcal{I}(\theta)|^{1/2} d\theta$ απειρίζεται για ένα σύνολο κατανομών. Για παράδειγμα, η κατανομή του Jeffreys για την μέση τιμή κανονικής κατανομής όταν η διακύμανση είναι γνωστή είναι η ομοιόμορφη κατανομή στο σύνολο των πραγματικών αριθμών - η οποία είναι προφανώς μη κανονικοποιήσιμη (improper prior).

Πρέπει να τονίσουμε ότι η χρήση μη κανονικοποιήσιμων κατανομών δεν είναι προβληματική όταν συμπερασματολογούμε για την ίδια την παράμετρο, από τη στιγμή που η εκ των υστέρων κατανομή είναι κανονικοποιήσιμη. Αντίθετα, η χρήση μη κανονικοποιήσιμων κατανομών είναι απαγορευτική στα ανώτερα επίπεδα συμπερασματολογίας, δηλαδή στην επιλογή τάξης μοντέλου και αλλά και του ίδιου του μοντέλου. Έτσι, στις περιπτώσεις αυτές προτείνεται ο πολλαπλασιασμός της κατανομής

με μια δεύτερη κατανομή η οποία θα έχει κέντρο μία αρχική εκτίμηση $\tilde{\theta}$ και όπως θα δείξουμε στη συνέχεια, η προκύπτουσα κατανομή ταυτίζεται υπό ορισμένες συνθήκες με τη συζυγή εκ των προτέρων κατανομή.

3.3.2 Υποκειμενικές εκ των προτέρων κατανομές

Ως υποκειμενικές κατανομές ορίζονται οι κατανομές εκείνες που εμπεριέχουν τις εκ των προτέρων πεποιθήσεις μας σε σχέση με τις παραμέτρους. Οι πεποιθήσεις μας χαρακτηρίζονται τόσο από την αναμενόμενη τιμή των παραμέτρων, όσο και από την ισχύ (strength) της πεποίθησής μας για την τιμή αυτή. Επί παραδείγματι, στην εκτίμηση μίας παραμέτρου $\theta \in \Re$, μπορούμε να αναθέσουμε μία κανονική κατανομή

$$\theta \sim \mathcal{N}(m_\theta, \sigma_\theta^2) \iff \pi(\theta|M_k) = (2\pi)^{-1/2}\sigma_\theta^{-1} \exp\left(-\frac{(\theta - m_\theta)^2}{2\sigma_\theta^2}\right) \tag{23}$$

Η m_θ αντιστοιχεί στην αναμενόμενη τιμή της θ, ενώ η σ_θ^{-2} στην ισχύ της εκ των προτέρων κατανομής. Οι έννοιες αυτές αποκτούν σαφέστερη υπόσταση στην περίπτωση των συζυγών εκ των προτέρων κατανομών. Όπως θα δείξουμε στην συνέχεια, η m_θ είναι η μέση τιμή των παρατηρήσεων που στοιχειοθετούν την εκ των προτέρων γνώση μας για την θ, ενώ η σ_θ^{-2} αντιστοιχεί στο πλήθος των παρατηρήσεων που διαθέτουμε εκ των προτέρων επί τη μέση ακρίβεια ανά παρατήρηση. Οι παράμετροι $\alpha = (m_\theta, \sigma_\theta^2)$ καλούνται υπερπαράμετροι (hyperparameters), ώστε να διακρίνονται από τις παραμέτρους της συνάρτησης πιθανοφάνειας.

3.3.3 Ιεραρχικές εκ των προτέρων κατανομές

Σε πολλές μοντελοποιήσεις μιγμάτων κατανομών $\theta = \{\varphi_k\}_{k=1}^K$, γνωρίζουμε ότι οι παράμετροι που αντιστοιχούν σε κάθε συντελεστή του μίγματος προέρχονται από μία κοινή κατανομή $\pi(\theta|\alpha, M_k)$, η οποία μπορεί να είναι είτε μονοτροπική είτε να έχει έναν περιορισμένο αριθμό τρόπων, ο οποίος είναι

εκ των προτέρων γνωστός. Μια τέτοια μοντελοποίηση είναι επιθυμητή σε περιπτώσεις που θέλουμε τα $\{\varphi_k\}_{k=1}^K$ να μοιράζονται ένα σύνολο κοινών χαρακτηριστικών. Έτσι, θέτωντας μία τέτοια εκ των προτέρων κατανομή $\pi(\theta|\alpha, M_k)$, αποδίδουμε μικρή πυκνότητα πιθανότητας σε περιπτώσεις στις οποίες οι συντελεστές $\{\varphi_k\}_{k=1}^K$ απέχουν αρκετά μεταξύ τους. Σε πολλά προβληματα ωστόσο, δεν γνωρίζουμε πολλά για το κέντρο ή τα κέντρα της εκ των προτέρων κατανομής που γεννά τα $\{\varphi_k\}_{k=1}^K$. Επιπλέον, η συμπερασματολογία ενδέχεται να επιδεικνύει μεγάλη ευαισθησία ως προς την επιλογή της υπερπαραμέτρου της εκ των προτέρων κατανομής που αντιστοιχεί στα κέντρα. Μία λύση για να ελαττώσουμε την ευαισθησία είναι να μειώσουμε την αντοχή, δηλαδή να επιλέξουμε αρκετά μεγάλη τιμή στην υπερπαράμετρο σ_θ^2. Δύο είναι τα προβλήματα με αυτή τη στρατηγική. Πρώτον, η μείωση της αντοχής ισοδυναμεί με αύξηση τη εκ των προτέρων μεταβλητότητας και ώς εκ τούτου στην μη αποτύπωση της εκ των προτέρων γνώσης μας για την από κοινού γειτνύαση των $\{\varphi_k\}_{k=1}^K$. Ένα δεύτερο πρόβλημα έγκειται στο ότι η αυθαίρετη μείωση της αντοχής είναι δυνατή μόνο για ορισμένες από τις παραμέτρους. Για ένα σύνολο παραμέτρων υπάρχει ένα κάτω φράγμα στην αντοχή, το οποίο αν δεν υπερβούμε η εκ των προτέρων κατανομή που προκύπτει είναι μη κανονικοποιήσιμη.

Μία εναλλακτική λύση στο πρόβλημα είναι η προσθήκη ενός ακόμα επιπέδου στην ιεραρχία, δηλαδή να θεωρήσουμε και τις υπερπαραμέτρους ως τυχαίες μεταβλητές και να αποδώσουμε σε αυτές μια εκ των προτέρων κατανομή. Έστω, λοιπόν, $\alpha \in \mathcal{A}$ οι παράμετροι που παραμετροποιούν την εκ των προτέρων κατανομή της θ και $\beta \in \mathcal{B}$ οι παράμετροι που παραμετροποιούν την εκ των προτέρων κατανομή των υπερπαραμέτρων α. Η εκ των υστέρων κατανομή της θ θα έχει την ακόλουθη σ.π.π.

$$\pi(\theta|\beta, M_k) = \int_{\alpha\in\mathcal{A}} \pi(\theta|\mathbf{y}, \alpha)\pi(\alpha|\beta)d\alpha \tag{24}$$

Οι ιεραρχικές κατανομές δίνουν επίσης τη δυνατότητα δημιουργίας μιγμάτων εκ των προτέρων κατανομών. Ας θεωρήσουμε την ακόλουθη κατανομή

$$\pi(\alpha|\beta) = \sum_{k=1}^K w_k\delta(\alpha, q_k),\ \beta = \{w_k, q_k\}_{k=1}^K,\ \sum_{k=1}^K w_k = 1 \tag{25}$$

Η εκ των υστέρων κατανομή της θ θα λαβεί τη ακόλουθη μορφή

$$\pi(\theta|\beta, M_k) = \sum_{k=1}^K w_k\pi(\theta|\mathbf{y}, q_k) \tag{26}$$

Η χρήση τέτοιων μιγμάτων ενδείκνυται σε περιπτώσεις όπου η εκ των προτέρων γνώση μας βασίζεται σε παρατηρήσεις που προέρχονται από διαφορετικούς πληθυσμούς.
Επιπλέον, η χρήση ιεραρχικών κατανομών είναι ιδιαίτερα διαδεδομένη σε μεθόδους συμπερασματολογίας που επιτρέπουν την ανανέωση στην εκτίμηση των ίδιων των υπερπαραμέτρων α με βάση τα δεδομένα $\mathbf{y}$. Τέτοιες μέθοδοι είναι η οικογένεια δειγματοληψίας Markov Chain Monte Carlo (Metropolis-Hasting, Gibbs Sampling), καθώς και μέθοδοι προσεγγιστικής συμπερασματολογίας (Variational Bayes, Expectation-Propagation).
Αναφέρουμε, τέλος, την εκφυλισμένη (degenerated) μορφή όπου η εκ των προτέρων κατανομή των υπερπαραμέτρων προσεγγίζεται από μία συνάρτηση Dirac $\alpha \sim \delta(\alpha, \beta)$. Η τιμή της παραμέτρου α (ή ισοδύναμα της β) υπολογίζεται αναδρομικά από τις παρατηρήσεις ώστε να μεγιστοποιεί την ακόλουθη κατανομή

$$\pi(\alpha|\mathbf{y}) = \int_{\theta \in \Theta} \pi(\alpha|\theta)\pi(\theta|\mathbf{y})d\theta \tag{27}$$

Η μεθοδολογία αυτή καλείται εμπειρική μπευσιανή (Empirical Bayes) μέθοδος προσδιορισμού εκ των προτέρων κατανομών και βασίζεται αποκλειστικά στα δεδομένα $\mathbf{y}$.

3.3.4 Συζυγείς εκ των προτέρων κατανομές

Οι συζυγείς (conjugate) εκ των προτέρων κατανομές είναι οι μοναδικές κατανομές οι οποίες οδηγούν σε κλειστό τύπο υπολογισμού τόσο της εκ των υστέρων κατανομής των παραμέτρων θ, όσο και της ολοκληρωμένης πιθανοφάνειας. Επιπλέον, η εκ των υστέρων κατανομή έχει την ίδια μαθηματική έκφραση με την εκ των προτέρων κατανομή. Η συζυγία θα πρέπει να κατανοηθεί ως σχέση μεταξύ της εκ των προτέρων κατανομής και της συνάρτησης πιθανοφάνειας, ενώ οι μόνες συναρτήσεις πιθανοφάνειας που επιδέχονται συζυγή είναι αυτές που ανήκουν στην οικογένεια των εκθετικών κατανομών.

3.4 Παραδείγματα συζυγών εκ των προτέρων κατανομών

Ας δούμε λοιπόν μερικά συγκεκριμένα παραδείγματα συζυγίας, με δύο εκθετικές οικογένειες, την κανονική και την πολυωνυμική κατανομή.

3.4.1 Κανονική κατανομή με γνωστή διακύμανση

Έστω λοιπόν το παράδειγμα της εκτίμησης της μέσης τιμής μ μίας κανονικής κατανομής με βάση τις παρατηρήσεις $\mathbf{y} = \{\mathbf{y}_i\}_{i=1}^n$ καθώς και της εκ των προτέρων γνώσης μας για την μ. Θεωρούμε την περίπτωση όπου η διακύμανση σ^2 είναι γνωστή. Θεωρούμε ότι η εκ των προτέρων γνώση μας για την μ μπορεί να κωδικοποιηθεί μέσω μίας κανονικής κατανομής $\mu \sim \mathcal{N}(\mu_0, \sigma_0^2)$. Μετά από πράξεις καταλήγουμε στην ακόλουθη εκ των υστέρων σ.π.π. $\mu \sim \mathcal{N}(\tilde{\mu}, \tilde{\sigma}^2)$, όπου

$$[\tilde{\mu}, \tilde{\sigma}^2] = \left[\left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \left(\frac{\mu_0}{\sigma_0^2} + \frac{n\bar{\mathbf{y}}}{\sigma^2} \right), \left(\frac{1}{\sigma_0^2} + \frac{n}{\sigma^2} \right)^{-1} \right] \tag{28}$$

και $\bar{\mathbf{y}} = \frac{1}{n} \sum_{i=1}^n \mathbf{y}_i$ ο δειγματικός μέσος των παρατηρήσεων.
Σημειώνουμε, τέλος ότι η παραπάνω σχέση αποστά μια πιο διαισθητική μορφή, αν αντί της διακύμανσης εργαζόμαστε με την αντίστροφή της στατιστική ποσότητα, την ακρίβεια (precision). Ορίζοντας λοιπόν την ακρίβεια ως $c^2 = \sigma^{-2}$, οι παραπάνω σχέσεις έχουν ως εξής

$$[\tilde{\mu}, \tilde{c}^2] = \left[\left(c_0^2 + nc^2 \right)^{-1} \left(c_0^2 \mu_0 + nc^2 \bar{\mathbf{y}} \right), \left(c_0^2 + nc^2 \right) \right] \tag{29}$$

Από την παραπάνω σχέση προκύπτει ότι η ακρίβεια της εκ των υστέρων κατανομής είναι το άθροισμα της ακρίβειας της εκ των προτέρων κατανομής c_0^2 και αυτής της πιθανοφάνειας nc^2, ενώ οι δύο αυτές ποσότητες είναι και τα βάρη της εκ των υστέρων εκτίμησης $\tilde{\mu}$.

3.4.2 Πολυωνυμική κατανομή

Έστω τώρα το πρόβλημα της εκτίμησης των παραμέτρων $\mathbf{w} = \{w_k\}_{k=1}^K$, $\sum_{k=1}^K w_k = 1$ και $w_k \geq 0$, μίας πολυωνυμικής κατανομής, η συνάρτηση μάζας πιθανότητας (σ.μ.π.) της οποίας ορίζεται ως εξής

$$p(\mathbf{n}|\mathbf{w}, n) = \frac{n!}{\prod_{k=1}^K n_k!} \prod_{k=1}^K w_k^{n_k} \tag{30}$$

όπου $\mathbf{n} = \{n_k\}_{k=1}^K$ και $\sum_{k=1}^K n_k = n$. Για πιο συμπαγή συμβολισμό, οι περιορισμοί $\sum_{k=1}^K w_k = 1$ και $w_k \geq 0$ μπορούν να γραφούν ως $\mathbf{w} \in \Delta^{K-1}$, όπου Δ^{K-1} το $(K-1)$-σύμπλεγμα (simplex).
Η συζυγής εκ των προτέρων κατανομή των $\mathbf{w} \in \Delta^{K-1}$ είναι η Dirichlet, την οποία θα συμβολίζουμε ως $\mathbf{w} \sim \mathcal{D}ir(\boldsymbol{\alpha})$ με σ.π.π. την ακόλουθη

$$\pi(\mathbf{w}|\alpha) = \begin{cases} \frac{\Gamma(\sum_{k=1}^K \alpha_k)}{\prod_{k=1}^K \Gamma(\alpha_k)} \prod_{k=1}^K w_k^{\alpha_k - 1}, \, \mathbf{w} \in \Delta^{K-1} \\ 0, \text{ αλλού} \end{cases} \tag{31}$$

όπου

$$\Gamma(x) = \int_0^\infty t^{x-1} e^{-t} dt \tag{32}$$

η συνάρτηση Γάμμα.
Θέτοντας $\alpha_0 = \sum_{k=1}^K \alpha_k$, ο τρόπος (υπό την έννοια του mode, δηλαδή του μεγίστου) και οι δύο πρώτες κεντρικές ροπές του w_k είναι οι ακόλουθες

$$\tilde{w}_k = \frac{\alpha_k - 1}{\alpha_0 - 1} \tag{33}$$

$$\mathcal{E}\{w_k\} = \frac{\alpha_k}{\alpha_0} \tag{34}$$

$$\mathcal{E}\{w_k^2\} - (\mathcal{E}\{w_k\})^2 = \frac{\alpha_k(\alpha_0 - \alpha_k)}{\alpha_0^2(\alpha_0 + 1)} = \frac{\mathcal{E}\{w_k\}(1 - \mathcal{E}\{w_k\})}{\alpha_0 + 1} \tag{35}$$

Η εκ των υστέρων κατανομή των $\mathbf{w}$ θα είναι επίσης Dirichlet, $\mathbf{w} \sim \mathcal{D}ir(\boldsymbol{\alpha} + \mathbf{n})$. Σε αντιστοιχία με τις (33)-(35), ο τρόπος και οι πρώτες δύο κεντρικές ροπές των παραμέτρων w_k βάσει της εκ των υστέρων κατανομής θα έχουν ως εξής

$$\tilde{w}_k = \frac{\alpha_k + n_k - 1}{\alpha_0 + n - 1} \tag{36}$$

$$\mathcal{E}\{w_k\} = \frac{\alpha_k + n_k}{\alpha_0 + n} \tag{37}$$

$$\mathcal{E}\{w_k^2\} - (\mathcal{E}\{w_k\})^2 = \frac{(\alpha_k + n_k)(\alpha_0 + n - \alpha_k - n_k)}{(\alpha_0 + n)^2(\alpha_0 + n + 1)} = \frac{\mathcal{E}\{w_k\}(1 - \mathcal{E}\{w_k\})}{\alpha_0 + n + 1} \tag{38}$$

Καθώς $n \to \infty$ η αναμενόμενη τιμή των w_k βάσει της εκ των υστέρων κατανομής ισούται με την εκτίμηση μέγιστης πιθανοφάνειας (ML), η οποία ισούται με $\hat{w}_k = \frac{n_k}{n}$.

Μέσω της κατανομής Dirichlet γίνεται ακόμα πιο διαισθητική η έννοια της αντοχής (strength) της εκ των προτέρων κατανομής. Ας θεωρήσουμε ένα ζάρι K πλευρών και ότι η παράμετρος w_k αντιστοιχεί στην πιθανότητα το αποτέλεσμα μίας ρίψης να είναι η k πλευρά. Υποθέτουμε επίσης συμμετρικές υπερπαραμέτρους και έτσι τις θέτουμε ίσες μεταξύ τους, $\alpha_k - 1 = \kappa$, $k = 1, \ldots, K$. Επιλέγουμε τον συμβολισμό αυτόν, καθώς η υπερπαράμετρος $\alpha_k - 1$ αντιστοιχεί στον αριθμό των εικονικών (virtual) πειραμάτων που έχουν προηγηθεί και η έκβασή τους ήταν η k πλευρά. Παραθέτουμε μια σειρά ιδιοτήτων για τρεις διαφορετικές τιμές του κ.

Επιλογές υπερπαραμέτρων Dirichlet και οι ιδιότητές τους

- **α΄ επιλογή:** $\kappa = 0$

  Θέτοντας $\kappa = 0$, δηλαδή $\alpha_k = 1, k = 1, \ldots, K$, υποθέτουμε καμία φανταστική παρατήρηση. Η κατανομή αυτή αποδίδει ίση μάζα πιθανότητας ανά μονάδα επιφάνειας στο $(K-1)$-σύμπλεγμα. Επιπλέον, η MAP εκτίμηση (36) ισούται με την εκτίμηση ML.

- **β΄ επιλογή:** $\kappa = -1$

  Η επιλογή $\alpha_k = 0, k = 1, \ldots, K$ είναι ένα τυπικό παράδειγμα χρήσης μη-κανονικοποιήσιμης εκ των προτέρων κατανομής (καθώς $\Gamma(0) = \infty$), η οποία όμως οδηγεί σε κανονικοποιήσιμη εκ των υστέρων κατανομή. Βάσει της τελευταίας, η εκτίμηση-σημείου MMSE της w_k ταυτίζεται με την ML εκτίμηση.

- **γ΄ επιλογή:** $\kappa = -1/2$

  Η επιλογή $\alpha_k = 1/2, k = 1, \ldots, K$ είναι η κατανομή του Jeffreys για την πολυωνυμική κατανομή. Όπως παρατηρούμε, η κατανομή αυτή δεν αποδίδει ίση μάζα πιθανότητας ανά μονάδα επιφάνειας στο $(K-1)$-σύμπλεγμα. Αν θεωρήσουμε όμως τον μετασχηματισμό $w_k \mapsto$

$q_k^{1/2}, k = 1, \ldots, K$ και εφαρμόσουμε τους κανόνες αλλαγής μεταβλητών, θα παρατηρήσουμε ότι η κατανομή Jeffreys των $\mathbf{w}$ τοποθετεί ίση μάζα ανά μονάδα επιφάνειας στην $(K-1)$-σφαίρα. Στην περίπτωση αυτή λέμε ότι η κατανομή του Jeffreys στο $(K-1)$-σύμπλεγμα κληρονομείται (inherited) από την ομοιόμορφη κατανομή στην $(K-1)$-σφαίρα μέσω του μετασχηματισμού $q_k^{1/2} \mapsto w_k, k = 1, \ldots, K$.

- **δ΄ επιλογή:** $\kappa \gg 0$

  Γενικότερα, η επιλογή $\kappa \gg 0$ αυξάνει την εκ των προτέρων πυκνότητα πιθανότητας επί του Δ^{K-1} σε περιοχές όπου τα $\mathbf{w}$ είναι πιο κοντά μεταξύ τους, και λαμβάνει τη μέγιστη τιμή για $w_k = K^{-1}$, $k = 1, \ldots, K$. Όσο αυξάνουμε το κ, τόσο ενισχύουμε στη συμπερασματολογία την εκ των προτέρων πίστη μας ότι το ζάρι είναι μη πολωμένο.

- **ε΄ επιλογή:** $\kappa \ll 0$

  Τέλος, τιμές $\kappa \ll 0$ εκφράζουν την εκ των προτέρων πίστη μας ότι το ζάρι είναι πολωμένο, καθώς τοποθετούμε στις περιοχές κοντά στις K γωνίες του Δ^{K-1} μεγάλη πυκνότητα πιθανότητας. Έτσι, είναι κατάλληλη όταν γνωρίζουμε ότι τα δεδομένα θα κυριαρχούνται από μία από τις K κατηγορίες, χωρίς ωστόσο να γνωρίζουμε εκ των προτέρων ποιά είναι αυτή.

3.5 Εκ των προτέρων κατανομές και γεωμετρία της πληροφορίας

Ας εξετάσσουμε τώρα τις εκ των προτέρων κατανομές της Μπεϋσιανής στατιστικής με βάση τη γεωμετρία πληροφορίας που σχετίζεται με τις εκθετικές κατανομές. Θα ξεκινήσουμε την ανάλυση με μια σύντομη αναφορά στην οικογένεια των δ-αποκλίσεων. Ο παρακάτω συμβολισμός είναι γενικός καθότι αναφέρεται στο σύνολο των πεπερασμένων θετικών μέτρων, δηλαδή $p, q \in \tilde{\mathcal{P}}$ υποσύνολο του οποίου αποτελούν τα πιθανοτικά μέτρα, $\mathcal{P} \subset \tilde{\mathcal{P}}$. Ορίζουμε τη δ-απόκλιση ως εξής

$$D_\delta(p||q) = \begin{cases} \frac{1}{\delta(1-\delta)} \left(1 - \int p^\delta q^{1-\delta} d\mathbf{y}\right), \text{ εάν } \delta \in (0,1) \\ \int \log(\frac{p}{q}) p d\mathbf{y}, \text{ εάν } \delta = 1 \\ \int \log(\frac{q}{p}) q d\mathbf{y}, \text{ εάν } \delta = 0 \end{cases} \tag{39}$$

Σημειώνεται ότι $\frac{p^\delta}{\delta} \to \log p$, $\delta \to 0$. Για $\delta = 1$ έχουμε την Kullback-Leibler απόκλιση, ενώ για $\delta = 0$ έχουμε και πάλι την Kullback-Leibler απόκλιση, αλλά με αντίστροφα ορίσματα. Επιπλέον, η μόνη συμμετρική απόκλιση είναι η περίπτωση $\delta = 1/2$, όπου

$$D_{1/2}(p||q) = 2\int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mathbf{y} \tag{40}$$

δηλαδή το διπλάσιο της τετραγωνικής απόστασης Hellinger.

Με βάση τις παραπάνω αποκλίσεις, λοιπόν, μπορούμε να ορίσουμε μία ευρεία οικογένεια εκ των προτέρων κατανομών. Παραθέτουμε την μορφή της σ.π.π., η οποία έχει ως ακολούθως

$$\Pi^{\alpha}_{\delta,\nu}(p_\theta; t_0) \propto \begin{cases} |\mathcal{I}(\boldsymbol{\theta})|^{1/2}[1 + \alpha\nu D_\delta(p_\theta||t_0)]^{-\frac{1}{\nu}}, \text{ εάν } \nu \in (0,1] \\ |\mathcal{I}(\boldsymbol{\theta})|^{1/2} e^{-\alpha D_\delta(p_\theta||t_0)}, \text{ εάν } \nu = 0 \end{cases} \tag{41}$$

Η t_0, αν και μπορεί να εκφράζει μια οσοδήποτε σύνθετη στατιστική οντότητα, στην απλούστερη περίπτωση συμβολίζει την εκ των προτέρων μέση τιμή των παραμέτρων.

Οι παραπάνω εκφράσεις έχουν μια βαθιά γεωμετρική ερμηνεία, καθώς προκύπτει από την ελαχιστοποίηση της ακόλουθης συνάρτησης κόστους

$$\mathcal{J}^{\alpha}_{\delta,\nu}(\Pi) = \gamma_e \int \Pi(\theta; t_0) D_\delta(p_\theta||t_0) d\theta + \gamma_u D_{1-\nu}(\Pi(\theta; t_0)|| \, |\mathcal{I}(\theta)|^{1/2}) \tag{42}$$

όπου

$$\alpha \leftarrow \frac{\gamma_e}{\gamma_u} \tag{43}$$

με χρήση λογισμού των μεταβολών (calculus of variations).

Η $\mathcal{J}^{\alpha}_{\delta,\nu}(\Pi)$ αποτελείται, λοιπόν, από δύο όρους. Ο πρώτος όρος αντιπροσωπεύει την αναμενόμενη δ-απόκλιση της p_θ από την κατανομή t_0. Αντίστροφα, ο δεύτερος όρος είναι η $(1-\nu)$-απόκλιση της $\Pi(\theta; t_0)$ από την ομοιόμορφη κατανομή - δηλαδή της κατανομής του Jeffreys $\propto |\mathcal{I}(\theta)|^{1/2}$. Για $\nu = 0$, ο δεύτερος όρος ταυτίζεται με τη (αρνητική) διαφορική εντροπία. Η σταθερά γ_e εκφράζει το βαθμό εμπιστοσύνης μας στην κατανομή t_0, η οποία είναι ανάλογη με τον αριθμό των παρατηρήσεων του τμήματος ομιλίας - υποθέτοντάς τες ανεξάρτητες και ομοιόμορφα κατανεμημένες. Αντίθετα, η γ_u εκφράζει το βάθμο στον οποίον επιθυμούμε η $\Pi(\theta; t_0)$ να μη φέρει πληροφορία. Εύκολα προκύπτει ότι οι δύο αυτές σταθερές μπορούν να εκφρασθούν με μία μόνο μεταβλητή $\alpha \leftarrow \frac{\gamma_e}{\gamma_u}$,

δηλαδή τον λόγο τους. Μία βασική ελεύθερη παράμετρος της έκφρασης, πέραν των (δ, ν) είναι λοιπόν η α, που ισοδυναμεί με τον αριθμό των εικονικών παρατηρήσεων.

Τέλος, η περίπτωση $(\delta, \nu) = (0, 0)$ είναι η συζυγής κατανομή των εκθετικών οικογενειών.

Οι εκθετικές οικογένειες έχουν ως σ.π.π. την παρακάτω μορφή

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) \tag{44}$$

όπου

$$\psi(\boldsymbol{\theta}) = \log \int_{\mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y})) h(\mathbf{y}) d\mathbf{y} \tag{45}$$

είναι η λογαριθμική κανονικοποιητική σταθερά (γνωστή και ως συνάρτηση *log-partition*) και $h(\mathbf{y})d\mathbf{y}$, $h : \mathcal{Y} \mapsto R^+$ το μέτρο αναφοράς. Επιπλέον, συμβολίζουμε με $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^P$ το P-διάστατο διάνυσμα των *φυσικών παραμέτρων* (*natural parameters*), και με $\mathbf{t}(\mathbf{y})$ το επίσης P-διάστατο διάνυσμα των *επαρκών στατιστικών* (*sufficient statistics*) των $\mathbf{y}$, δηλαδή μία απεικόνιση (mapping) $\mathcal{Y} \mapsto \Re^P$.

Με βαση λοιπόν την απόκλιση Kullback-Leibler, η σ.π.π. για ένα δείγμα $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^n$ γράφεται και ως

$$p(\{\mathbf{y}^{(i)}\}_{i=1}^n|\boldsymbol{\theta}) = h(\mathbf{y}) \exp(-n(D_0(\boldsymbol{\theta}||\mathbf{t}(\mathbf{y})) - \varphi(\mathbf{t}(\mathbf{y})))) \tag{46}$$

όπου $\varphi(\mathbf{t}(\mathbf{y}))$ η αρνητική εντροπία της κατανομής που ορίζουν οι επαρκείς στατιστικές $\mathbf{t}(\mathbf{y})$. Πολλαπλασιάζοντας λοιπόν την πιθανοφάνεια με την $(\delta, \nu) = (0, 0)$ εκ των προτέρων κατανομή θα λάβουμε

$$\frac{p(\{\mathbf{y}^{(i)}\}_{i=1}^n|\boldsymbol{\theta})\pi(\boldsymbol{\theta};\boldsymbol{\eta},\alpha)}{\int_\Theta p(\{\mathbf{y}^{(i)}\}_{i=1}^n|\boldsymbol{\theta})\pi(\boldsymbol{\theta};\boldsymbol{\eta},\alpha)d\boldsymbol{\theta}} \propto |\mathcal{I}(\boldsymbol{\theta})|^{1/2} \exp\left(-(n+\alpha)(D_0(\boldsymbol{\theta}||\tilde{\boldsymbol{\eta}}))\right) \tag{47}$$

όπου

$$\tilde{\boldsymbol{\eta}} = \frac{a\mathbf{t}(\mathbf{y}) + \alpha\boldsymbol{\eta}_0}{n+\alpha} \tag{48}$$

ο εκ των υστέρων μέσος (posterior mean) των παραμέτρων. Παρατηρούμε λοιπόν τη συζυγία μεταξύ πιθανοφάνειας και εκ των προτέρων κατανομών. Τα νέα δείγματα $\{\mathbf{y}^{(i)}\}_{i=1}^n$ προστίθενται στις εικονικές α το πλήθος παρατηρήσεις και δημιουργούν τις υπερπαραμέτρους της εκ των υστέρων πλέον κατανομής, $(\tilde{\boldsymbol{\eta}}, n+\alpha)$.

3.5.1 Ανακεφαλαίωση

Στο κεφάλαιο αυτό αναφερθήκαμε σε μερικές από τις θεμελιώδεις έννοιες και μεδόδους της Μπεϋσιανής στατιστικής. Δείξαμε την γενική μεθοδολογία με την οποία αντιμετωπίζεται η εκτίμηση παραμέτρων και η επιλογή τάξης ενός μοντέλου. Κατόπιν, εξετάσαμε τις εκ των προτέρων κατανομές και παρουσιάσαμε παραδείγματα συζυγών κατανομών. Τέλος, εστιάσαμε σε μία γεωμετρική ερμηνεία μιας οικογένειας κατανομών και δείξαμε με εναλλακτικό τρόπο του πώς επιτυγχάνεται η συζυγία με τη συνάρτηση πιθανοφάνειας, στην περίπτωση που η τελευταία αποτελεί σ.π.π. εκθετικής οικογένειας κατανομών. Μία αναλυτικότερη περιγραφή των παραπάνω, καθώς και των θεμελιωδών εννοιών της Γεωμετρίας της Πληροφορίας μπορεί να βρεθεί στο Παράρτημα.

4 ΣΥΜΜΙΞΗ ΔΥΑΔΙΚΩΝ ΧΑΡΑΚΤΗΡΙΣΤΙΚΩΝ ΜΕΣΩ ΕΚΘΕΤΙΚΩΝ ΜΟΝΤΕΛΩΝ ΚΑΙ ΧΡΗΣΗΣ ΤΗΣ ΑΡΧΗΣ ΤΗΣ ΜΕΓΙΣΤΗΣ ΕΝΤΡΟΠΙΑΣ ΜΕ ΣΤΟΧΟ ΤΗΝ ΑΥΤΟΜΑΤΗ ΟΜΑΔΟΠΟΙΗΣΗ ΟΜΙΛΗΤΩΝ

4.1 Εισαγωγή

Στο κεφάλαιο αυτό παρουσιάζουμε μία προσέγγιση για το πρόβλημα της ομαδοποίησης, όπως στο [1]. Η κεντρική ιδέα της προτεινόμενης μεθόδου είναι ο συνδυασμός εναλλακτικών παραμετρικών χώρων, μοντέλων, μετρικών αλλά και κατωφλιώσεων σε ένα εκθετικό μοντέλο, η εκπαίδευση του οποίου βασίζεται στην αρχή της μέγιστης εντροπίας (ΜΕ) του Jaynes, [59]. Για την εκπαίδευση του μοντέλου απαιτείται ένα σύνολο εκμάθησης, κάθε δείγμα του οποίου αποτελείται από δύο τμήματα ομιλίας, τη μακροσκοπική κλάση στην οποία τα τμήματα αυτά ανήκουν καθώς και την ένδειξη αν αυτά τα τμήματα προέρχονται ή όχι από τον ίδιο ομιλητή. Η απόκριση του μοντέλου είναι ένα βαθμωτό μέγεθος στο διάστημα $[0, 1]$ και προσεγγίζει την εκ των υστέρων πιθανότητα τα τμήματα να προέρχονται από *διαφορετικό* ομιλητή.

4.1.1 Η προτεινόμενη μεθόδος και τα πλεονεκτήματά της

Στην εισαγωγή της παρούσας διατριβής παρουσιάσθηκε ένα σύνολο από μετρικές στατιστικής απόκλισης μεταξύ δύο τμημάτων ομιλίας. Επιπλέον, έγινε αναφορά στους εναλλακτικούς παραμετρικούς χώρους, με έμφαση στα MFCC και LSF καθώς και στις δυνατές μοντελοποίησεις της πυκνότητας πιθανότητας. Ο οποιοσδήποτε συνδυασμός παραμετρικού χώρου, μοντέλου, μετρικής απόκλισης και κατωφλίου συνιστά και από έναν ταξινομητή, αφού λαμβάνει ως είσοδο ένα σύνολο δεδομένων - τα δύο συνολα παρατηρήσεων που αντιστοιχούν σε κάθε τμήμα - και το ταξινομεί είτε στην κλάση *ίδιοι* ομιλητές είτε στην κλάση *διαφορετικοί* ομιλητές. Υιοθετώντας αυτή την έννοια της

κλάσης, ένα φυσικό ερώτημα που τίθεται είναι αν θα μπορούσαμε να συνδυάσουμε αποτελεσματικά αυτούς τους ταξινομητές, με τρόπο ώστε να οδηγηθούμε, α) σε μεγαλύτερη ακρίβεια στην ομαδοποίηση συγκριτικά με τη χρήση ενός και μόνο ταξινομητή και β) σε μία πιθανοτική ταξινόμηση, υπό την έννοια της απόδοσης σε κάθε ζεύγος της πιθανότητας να ανήκει σε μία από τις δύο κλάσεις. Ένα δεύτερο ζήτημα σχετίζεται με την κατάτμηση του χώρου εισόδου με βάση α) τη μακροσκοπική κλάση στην οποία ανήκουν τα δύο τμήματα ομιλίας, την οποία υποθέτουμε ίδια για κάθε ομιλητή ξεχωριστά, διαφορετικά θα θεωρούμε ότι τα τμήματα ανήκουν σε διαφορετικούς ομιλητές και β) το πλήθος των παρατηρήσεων που διαθέτουμε για το κάθε τμήμα ομιλίας. Μέσω της κατάτμησης του χώρου εισόδου θα μπορέσουμε να εκπαιδεύσουμε ένα μοντέλο ανά κατηγορία, το οποίο θα έχει εκπαιδευτεί με το αντίστοιχο σύνολο εκμάθησης, αναδεικνύοντας τους ταξινομητές εκείνους που λειτουργούν καλύτερα για κάθε κατηγορία. Μάλιστα, λόγω της πιθανοτικής μοντελοποίησης, η απόκριση του μοντέλου για νέα ζεύγη θα είναι με τη σειρά της ένας συνδυασμός της απόκρισης κάθε μοντέλου κατηγορίας, με βάρος ανάλογο της εκ των υστέρων πιθανότητας του ζεύγους να ανήκει στην κατηγορία αυτή (model averaging).

4.1.2 Οι εφαρμογές του μοντέλου σε πεδία αναγνώρισης προτύπων

Το μοντέλο που παρουσιάζουμε έχει βρει εφαρμογή σε τουλάχιστον δύο πεδία της αναγνώρισης προτύπων, όπως αυτό της Επεξεργασίας Φυσικής Γλώσσας (NLP, [22], [60]), της ανίχνευσης αλλαγής θεματολογίας σε εκπομπές ειδησεογραφικού χαρακτήρα (News Story Segmentation, [21], [20]) καθώς και σε επεξεργασία εικόνας ([61]).

Σε εφαρμογές επεξεργασίας φωνής, ενδιαφέρον παρουσιάζει η χρήση του μοντέλου για αναγνώριση φωνημάτων που προτείνεται στο [62]. Έχοντας στη διάθεσή μας πολλαπλούς παραμετρικούς χώρους (MFCC, PLP, RASTA-PLP), επιχειρείται να βρεθεί το κατάλληλο διάνυσμα χαρακτηριστικών το οποίο οδηγεί σε βέλτιστο διαχωρισμό μεταξύ των φωνημάτων. Το τελικό μοντέλο επιτυγχάνει μείωση της διάστασης των χαρακτηριστικών στο 33% ενώ η απόκρισή του μπορεί να ενσωματωθεί σε συστήματα αναγνώρισης φωνής με Υπονοούμενα Μαρκοβιανά Μοντέλα.

4.1.3 Συμβολισμοί μεταβλητών

Στο σημείο αυτό εισάγουμε τους απαραίτητους συμβολισμούς. Θεωρούμε αρχικά την περίπτωση ενός μοντέλου, οι παράμετροι του οποίου θα πρέπει να εκτιμηθούν από ένα ενιαίο σύνολο εκμάθησης $z_j = (x_j, b_j), j = 1, \ldots, n$. Έστω x ένα ζεύγος τμημάτων ομιλίας και $b \in \{0, 1\}$ η δυαδική μεταβλητή που ορίζει εάν τα τμήματα ανήκουν στουν ίδιο ομιλητή ($b = 0$) ή σε διαφορετικό ($b = 1$). Επιπλέον, έστω N_f δυαδικοί ταξινομητές $f_i(x, b) \in \{0, 1\}$, $f_i(x, b) = 1_{\{g_i(x)=b\}}$, όπου $1_{\{\cdot\}}$ η συνάρτηση ένδειξης (indicator function) και $g_i(x)$ η i-οστή συνάρτηση πρόβλεψης. Η συνάρτηση $g_i(x)$ ισούται με μονάδα αν ο i-οστός ταξινομητής αποφαίνεται ότι το ζεύγος x ανήκει στην κλάση *διαφορετικοί* ομιλητές και διαφορετικά μηδέν.
Το εκθετικό μοντέλο έχει την παρακάτω μορφή

$$q_\lambda(b|x) = \frac{1}{Z_\lambda(x)} \exp\left(\sum_{i=1}^{N_f} \lambda_i f_i(x, b)\right) \tag{49}$$

όπου $Z_\lambda(x)$ η σταθερά κανονικοποίησης

$$Z_\lambda(x) = \sum_b \exp\left(\sum_{i=1}^{N_f} \lambda_i f_i(x, b)\right). \tag{50}$$

Όπως θα δείξουμε στη συνέχεια, οι παράμετροι του μοντέλου $\lambda = [\lambda_1, \lambda_2, \ldots, \lambda_{N_f}]$ αντιστοιχούν στους πολλαπλασιαστές Lagrange ενός προβλήματος βελτιστοποίησης.
Ορίζουμε επίσης την εμπειρική κατανομή των δεδομένων $\tilde{p}(x, b)$, η οποία ορίζεται ως $\tilde{p}(\zeta) = \frac{1}{n}\sum_{j=1}^{n} \delta(\zeta, z_j)$ όπου $z_j = (x_j, b_j), j = 1, \ldots, n$ το σύνολο εκμάθησης. Η εμπειρική κατανομή αποτελείται από n το πλήθος συναρτήσεις Dirac στα σημεία του χώρου (x, b) όπου έχουμε δεδομένα εκμάθησης, με κατάλληλη κανονικοποίηση (δηλαδή n^{-1}) ώστε το ολοκλήρωμά της να ισούται με τη μονάδα.

4.2 Το μοντέλο ως λύση μέγιστης εντροπίας

Με βάση τους παραπάνω ορισμούς, η λογαριθμική πιθανοφάνεια του μοντέλου στο σύνολο εκμάθησης εκφράζεται ως

$$\mathcal{L}(\tilde{p}|q_\lambda) = \sum_{x,b} \tilde{p}(x,b) \sum_i \lambda_i f_i(x,b) - \sum_x \tilde{p}(x) \log \sum_b \exp \sum_i \lambda_i f_i(x,b) \tag{51}$$

Διαφορίζοντας την παραπάνω σχέση ως προς λ_i λαμβάνουμε

$$\begin{aligned} \frac{\partial \mathcal{L}(\tilde{p}|q_\lambda)}{\partial \lambda_i} &= \sum_{x,b} \tilde{p}(x,b) f_i(x,b) - \sum_{x,b} \tilde{p}(x) q_\lambda(b|x) f_i(x,b) & (52) \\ &= < \tilde{f}_i > - < f_i > . & (53) \end{aligned}$$

Στην παραπάνω σχέση συμβολίζουμε με $< \tilde{f}_i >$ την αναμενόμενη τιμή του $f_i(x,b)$ ως προς την εμπειρική κατανομή $\tilde{p}(x,b)$, ενώ με $< f_i >$ την αντίστοιχη τιμή ως προς την κατανομή $\tilde{p}(x)q_\lambda(b|x)$. Θέτοντας την παράγωγο ίση με μηδέν προκύπτει ότι οι δύο αυτές ποσότητες της σχέσης (52) θα πρέπει να είναι ίσες.

Η παραπάνω ιδιότητα κάθε άλλο παρά τυχαία είναι. Τουναντίον, το εκθετικό μοντέλο προκύπτει μέσω της μεγιστοποίησης της εντροπίας με τους παραπάνω N_f το πλήθος περιορισμούς.

Η δεσμευμένη ως προς x (conditional) εντροπία του μοντέλου στα δεδομένα εκμάθησης έχει την ακόλουθη μορφή

$$M(q_\lambda) = -\sum_{x,b} \tilde{p}(x) q_\lambda(b|x) \log q_\lambda(b|x) \tag{54}$$

Το εκθετικό μοντέλο προκύπτει από τη μεγιστοποίηση της $M(q_\lambda)$, με τους N_f περιορισμούς. Έχουμε, λοιπόν, το πρόβλημα

$$q_\lambda = \underset{q \in \mathcal{C}}{\operatorname{argmax}}\, M(q_\lambda) \tag{55}$$

όπου $\mathcal{C}$ το υποσύνολο του χώρου των συναρτήσεων κατανομών πιθανότητας $\mathcal{Q}$, που αποτελείται από τις κατανομές που υπακούουν στους N_f περιορισμούς. Στη θεωρία βελτιστοποίησης, το πρόβλημα αυτό καλείται πρωταρχικό (primal) και η επίλυσή του βασίζεται στο θεώρημα Kuhn-Tucker.

Με χρήση των πολλαπλασιαστών Lagrange το πρόβλημα βελτιστοποίησης γράφεται ως

$$q_\lambda = \underset{q \in \mathcal{Q}}{\operatorname{argmax}}\, \Lambda(q, \lambda) \tag{56}$$

όπου

$$\Lambda(q,\lambda) = M(q_\lambda) + \sum_i \lambda_i \left(< \tilde{f}_i > - < f_i > \right) \tag{57}$$

η Λαγκρανσιανή. Το παραπάνω πρόβλημα έχει ως λύση το εκθετικό μοντέλο (49), το οποίο και συμβολίζουμε με q_λ.

Έστω τώρα $\Psi(\lambda) = \Lambda(q_\lambda, \lambda)$ το μέγιστο της Λαγκρανσιανής. Η εκτίμηση των παραμέτρων λ^* προκύπτει ως λύση του δυικού (dual) προβλήματος

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, \Psi(\lambda) \tag{58}$$

Τοποθετώντας το μοντέλο q_λ στην $\Psi(\lambda)$, έχουμε

$$\Psi(\lambda) = -\sum_x \tilde{p}(x) \log Z_\lambda(x) + \sum_i \lambda_i < \tilde{f}_i > \tag{59}$$

Η παραπάνω συνάρτηση είναι η λογαριθμική πιθανοφάνεια του μοντέλου στα δεδομένα εκμάθησης, που δίνεται στη σχέση (51). Έτσι, η επίλυση του δυικού προβλήματος ισοδυναμεί με τη μεγιστοποίηση ως προς λ της λογαριθμικής δεσμευμένης πιθανοφάνειας

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, \Lambda(\tilde{p}|q_\lambda), \tag{60}$$

4.3 Εκπαίδευση του εκθετικού μοντέλου

4.3.1 Εισαγωγή

Παρά την εκθετική του μορφή, πρέπει να τονισθεί ότι η συνάρτηση πιθανοφάνειας των ολοκληρωμένων δεδομένων $z = (x, b)$ του μοντέλου *δεν* ανήκει στην εκθετική οικογένεια. Η σ.μ.π. γράφεται ως

$$q_\lambda(b|x) = \exp\left(\sum_{i=1}^{N_f} \lambda_i f_i(x,b) - \log Z_\lambda(x) \right) \tag{61}$$

όπου $Z_\lambda(x)$ η σταθερά κανονικοποίησης

$$Z_\lambda(x) = \sum_b \exp\left(\sum_{i=1}^{N_f} \lambda_i f_i(x,b) \right). \tag{62}$$

Για να ανήκει στην εκθετική οικογένεια, θα πρέπει η συνάρτηση $\log Z_\lambda(x)$ να μπορούσε να εκφρασθεί σαν άθροισμα $\log Z_\lambda(x) = \log u(\lambda) + \log v(x, b)$, το οποίο δεν ισχύει. Το γεγονός αυτό έχει ως αποτέλεσμα την μη ύπαρξη κλειστού τύπου υπολογισμού των παραμέτρων λ. Επομένως, θα πρέπει να καταφύγουμε σε αναδρομικούς τύπους υπολογισμού, ώστε να μεγιστοποιήσουμε την πιθανοφάνεια.

Ταυτόχρονα, όμως, η πιθανοφάνεια έχει ένα και μοναδικό μέγιστο, αφού ο πίνακας με στοιχείο (i, j)

$$\frac{\partial^2}{\partial\lambda_i \partial\lambda_j} q_\lambda(b|x) \tag{63}$$

είναι αρνητικά ορισμένος. Γνωρίζουμε λοιπόν ότι υπάρχει μία και μόνη λύση στο προβλήμα της εκτίμησης μέγιστης πιθανοφάνειας $\hat{\lambda}$.

4.3.2 Γενική μορφή του αναδρομικού αλγορίθμου

Μελετούμε τώρα μεθόδους εκπαίδευσης του παραπάνω μοντέλου. Όπως αναφέραμε, έχουμε στη διάθεσή μας ένα σύνολο εκπαίδευσης $z_j = (x_j, b_j), j = 1, \ldots, n$. Καθένας από τους ταξινομητές $f_i(x, b)$ λαμβάνει ως είσοδο το ζεύγος x και επιτελεί τον μετασχηματισμό $x \mapsto \{0, 1\}$ σύμφωνα με τον χώρο χαρακτηριστικών, μοντέλο, μετρική και κατώφλι που του αντιστοιχούν. Οι συνδυασμοί όλων των πιθανών συνδυασμών είναι ένας πολύ μεγάλος αριθμός και ως εκ τούτου χρειαζόμαστε ένα κατάλληλα επιλεγμένο υποσύνολό τους. Έτσι, πέραν της εκτίμησης του βέλτιστου λ (φάση εκτίμησης) θα χρειασθούμε και έναν αλγόριθμο επιλογής των πιο κατάλληλων ταξινομητών (φάση επιλογής).

Η διαδικασία εκμάθησης του αλγορίθμου είναι αναδρομική. Κατά τη φάση εκτίμησης, ο αλγόριθμος έχει στη διάθεσή του έναν αριθμό ταξινομητών και με βάση το σύνολο εκμάθησης βελτιστοποιεί το παραμετρικό διάνυσμα λ, έτσι ώστε

$$\lambda^* = \underset{\lambda}{\operatorname{argmax}}\, \mathcal{L}(\lambda|\tilde{p}) \tag{64}$$

Κατά τη φάση επιλογής, ο αλγόριθμος επιλέγει τον ταξινομητή εκείνον ο οποίος προστιθέμενος στο ήδη υπάρχον μοντέλο επιφέρει το μεγαλύτερο κέρδος. Το κέρδος αυτό ορίζεται ως εξής,

$$G_{a,g} = D(\tilde{p}||q_\lambda) - D(\tilde{p}||q_{a.g}) \tag{65}$$

όπου

$$q_{a.g}(b|x) = \frac{\exp(ag(x,b))q_\lambda(b|x)}{Z_a(x)} \tag{66}$$

και $D(\tilde{p}||q_{a.g})$ η Kullback-Leibler απόκλιση μεταξύ της εμπειρικής κατανομής $\tilde{p}$ και του μοντέλου $q_{\cdot}$. Αναπτύσσοντας την $D(\tilde{p}||q_{\cdot})$ ως

$$D(\tilde{p}||q_{\cdot}) = \mathbb{E}_{\tilde{p}}\{\log \tilde{p}\} - \mathbb{E}_{\tilde{p}}\{\log q_{\cdot}\} \tag{67}$$

και παρατηρώντας ότι ο πρώτος όρος είναι ανεξάρτητος του μοντέλου, το κέρδος $G_{a,g}$ δεν είναι άλλο από την αύξηση που προκαλείται στη λογαριθμική πιθανοφάνεια του μοντέλου με την συμπερίληψη του ταξινομητή g, δηλαδή

$$G_{a,g} = \mathcal{L}(\lambda, a, g|\tilde{p}) - \mathcal{L}(\lambda|\tilde{p}). \tag{68}$$

4.3.3 Οι φάσεις εκτίμησης και επιλογής ταξινομητών αναλυτικότερα

Για τη φάση εκτίμησης του λ^* τρεις είναι οι πιο δημοφιλείς αλγόριθμοι. Ο πρώτος είναι ο αλγόριθμος γενικευμένης αναδρομικής κλιμάκωσης (Generalized Iterative Scaling, GIS) ο δεύτερος είναι αλγόριθμος βελτιωμένης αναδρομικής κλιμάκωσης (Improved Iterative Scaling, IIS) και ο τρίτος είναι ο αλγόριθμος περιορισμένης μνήμης μεταβλητής μετρικής (Limited-Memory Variable Metric, απαντάται στη βιβλιογραφία ως L-BFGS). Στη γενική τους μορφή, όλοι οι αλγόριθμοι έχουν ως ακολούθως.

Γενική μορφή αλγορίθμων αναδρομικής κλιμάκωσης

1. Έστω $\lambda^0 = [\lambda_1^0, \ldots, \lambda_k^0]$ η αρχική τιμή του k-διάστατου διανύσματος συντελεστών του μοντέλου.

2. Για $i = \{1, 2, \ldots, k\}$:

   (α′) Έστω $\Delta\lambda_i$ η λύση μίας εξίσωσης που διαφέρει ανάλογα με τον αλγόριθμο.

   (β′) Ενημέρωση: $\lambda_i \leftarrow \lambda_i + \Delta\lambda_i$

3. Αν $i = k$ τερματισμός αλγορίθμου.

Παρουσιάζουμε τώρα τους τρόπους υπολογισμού του $\Delta\lambda_i$ που προτείνει ο κάθε αλγόριθμος. Ορίζουμε αρχικά την ποσότητα $f^{\#}(x, b) \equiv \sum_i f_i(x, b)$ η οποία είναι το άθροισμα των $f_i(x, b)$ για κάθε ένα από τα δείγματα του συνόλου εκμάθησης. Ο αλγόριθμος γενικευμένης αναδρομικής κλιμάκωσης (GIS) υπολογίζει το βήμα ως εξής

$$\Delta\lambda_i = \frac{1}{C} \log \left(\frac{\sum_{x,b} \tilde{p}(x, b) f_i(x, b)}{\sum_{x,b} \tilde{p}(x) q_\lambda(b|x) f_i(x, b)} \right) \tag{69}$$

Στην περίπτωση που ισχύει ότι $f^{\#}(x, b) = c$, δηλαδή το άθροισμα των $f_i(x, b)$ είναι ίδιο σε κάθε δείγμα, θέτουμε $C = c$, διαφορετικά θέτουμε $C = \max\{f^{\#}(x, b)\}$. Στην τελευταία περίπτωση, ο αλγόριθμος παρουσιάζει αργή σύγκλιση στο λ^*.

Ο αλγόριθμος βελτιωμένης αναδρομικής κλιμάκωσης (IIS) επιλέγει το $\Delta\lambda_i$ επιλύοντας την εξίσωση

$$\sum_{x,b} \tilde{p}(x) q_\lambda(b|x) f_i(x, b) \exp\left(\Delta\lambda_i f^{\#}(x, b)\right) = \sum_{x,b} \tilde{p}(x, b) f_i(x, b) \tag{70}$$

όπου $f^{\#}(x, b) \equiv \sum_i f_i(x, b)$. Στην περίπτωση που το $f^{\#}(x, b)$ είναι σταθερό, η λύση της (70) είναι αναλυτική, διαφορετικά επιστρατεύεται η μέθοδος του Newton. Γενικά, ο (IIS) προσφέρει ταχύτερη σύγκλιση σε σχέση με τον (GIS).

Αναφέρουμε τέλος και τον αλγόριθμο περιορισμένης μνήμης μεταβλητής μετρικής (L-BFGS). Ο αλγόριθμος αυτός χρησιμοποιεί την quasy-Newton μέθοδο για την εκτίμηση των $\Delta\lambda_i$. Η μέθοδος του Newton ανήκει στις μεθόδους βελτιστοποίησης δευτέρου βαθμού, αφού κάνει χρήση του πίνακα Hessian $H(\lambda)$ της λογαριθμικής πιθανοφάνειας

$$\Delta\lambda_i = H^{-1}(\lambda) G(\lambda) \tag{71}$$

όπου $G(\lambda) = \frac{\partial \mathcal{L}(\lambda)}{\partial \lambda}$ η παράγωγος της λογαριθμικής πιθανοφάνειας ως προς το διάνυσμα λ. Δεδομένου ότι ο υπολογισμός και η αντιστροφή του πίνακα Hessian είναι ιδιαίτερα χρονοβόρες διαδικασίες, ο αλγόριθμος L-BFGS προβαίνει σε προσέγγιση του Hessian σε μία περιοχή γύρω από τιμές $\Delta\lambda_i$ των αμέσως προηγούμενων αναδρομών. Έτσι, αποφεύγει την αποθήκευση στη μνήμη παραμέτρων που απαιτούν άλλες προσεγγιστικές μέθοδοι, ενώ προσφέρει ταχεία σύγκλιση.
Για μία αναλυτική παρουσίαση των αλγορίθμων και σύγκριση μεταξύ τους παραπέμπουμε στο [63].

Φάση επιλογής ταξινομητή Ερχόμαστε τώρα στη φάση επιλογής του νέου ταξινομητή. Οι εξισώσεις (65) και (66) δείχνουν ότι η επιλογή βασίζεται στο ποιός από τους εναπομείναντες ταξινομητές θα επιφέρει το μεγαλύτερο κέρδος αν προστεθεί στο ήδη υπάρχον μοντέλο. Για να εκτιμηθεί το κέρδος συμπερίληψης τού κάθε ταξινομητή, το συνολικό επαυξημένο μοντέλο $\lambda^+ = [\lambda, a]$ θα πρέπει να επανεκτιμηθεί εκ νέου, διαδικασία η οποία θα πρέπει να επαναλαμβάνεται τόσες φορές όσοι και οι εναπομείναντες ταξινομητές. Μια τέτοια διαδικασία είναι ιδιαίτερα χρονοβόρα και ως εκ τούτου υιοθετούμε μία ταχύτερη αλλά υποβέλτιστη στρατηγική. Έτσι, αντί να εκτιμούμε το επαυξημένο μοντέλο $\lambda^+ = [\lambda, a]$, εκτιμούμε μόνο τον συντελεστή που αντιστοιχεί στον εκάστοτε υποψήφιο ταξινομητή, δηλαδή το a, διατηρώντας το λ σταθερό. Όταν η φάση επιλογής περατωθεί, η φάση εκτίμησης είναι πλέον υπεύθυνη για την εκτίμηση του βέλτιστου επαυξημένου πλέον μοντέλου, με αρχική τιμή την $\lambda^+ = [\lambda, a]$.

4.4 Μπεϋσιανές και ημι-μπεϋσιανές μέθοδοι

4.4.1 Εισαγωγή

Έχει ενδιαφέρον να εστιάσουμε αρχικά στη μεθοδολογία της μέγιστης εντροπίας. Είδαμε παραπάνω ότι το ίδιο το εκθετικό μοντέλο προκύπτει ως λύση μίας συνάρτησης βελτιστοποίησης. Συγκεκριμένα, μεγιστοποιήσαμε την εντροπία της κατανομής, με περιορισμούς που σχετίζονται με τις αναμενόμενες τιμές των χαρακτηριστικών. Δεδομένου ότι ο δειγματικός χώρος της συνάρτη-

σης είναι ο διακριτός χώρος $b \in \{0, 1\}$, η παραπάνω διαδικασία ισοδυναμεί με την ελαχιστοποίηση της απόκλισης KL μεταξύ της q και της ομοιόμορφης κατανομής $\bar{\pi}(b) = 1/|b|$, όπου $|b| = 2$. Η απόκλιση αναλύεται ως

$$D(q||\bar{\pi}) = \sum_b q(b) \log q(b) - \sum_b q(b) \log \bar{\pi} \tag{72}$$

και παρατηρούμε ότι ο δεύτερος όρος είναι ανεξάρτητος της q. Έτσι, η λύση μέγιστης εντροπίας ταυτίζεται με τη λύση ελάχιστης απόκλισης KL από την ομοιόμορφη κατανομή. Σημειώνουμε την ομοιότητα της στρατηγικής αυτής με το πρόβλημα της επιλογής εκ των προτέρων κατανομών στη Μπεϋσιανή στατιστική, η οποία θα παρουσιασθεί στο κεφάλαιο όπου παρουσιάζουμε τον αλγόριθμο μετατόπισης τους μέσου.

Η μεγιστοποίηση της εντροπίας λοιπόν δρα ως ένας ρυθμιστής (regularizer) στον χώρο των κατανομών, μέσω του οποίου προκρίνονται λύσεις οι οποίες δεν υπερταιριάζουν (overfitting) στα δεδομένα εκμάθησης. Τυπικά φαινόμενα υπερταιριάσματος είναι οι πολύ μεγάλες τιμές συντελεστών που αντιστοιχούν σε κάποια χαρακτηριστικά, λόγω του μη επαρκούς ή και θορυβώδους δείγματος εκπαίδευσης. Το πρόβλημα είναι ότι η συγκεκριμένη στρατηγική σπάνια είναι επαρκής ώστε το φαινόμενο του υπερταιριάσματος τελικά να αποφευχθεί. Στο [64], οι συγγραφείς έδειξαν ότι η χρήση ημι-μπεϋσιανών τεχνικών ρύθμισης (κανονική εκ των προτέρων κατανομή μηδενικής μέσης τιμής στα λ_i και εκτίμηση MAP) βελτιώνει σημαντικά την αξιοπιστία γλωσσικών μοντέλων. Στο κεφάλαιο αυτό λοιπόν θα εστιάσουμε σε μεθόδους ρύθμισης με τις οποίες θα μπορέσουμε να εκτιμήσουμε πιο στιβαρά τις παραμέτρους λ_i και θα εξάγουμε την νέα μορφή που θα έχει η συνάρτηση βελτιστοποίησης ώστε να ενσωματώνει την εκ των προτέρων κατανομή.

4.4.2 Ημι-Μπεϋσιανή μοντελοποίηση και εκτίμηση μέγιστης εκ των υστέρων πιθανότητας

Το πρόβλημα της εκτίμησης των συντελεστών λ_i είναι ένα τυπικό παράδειγμα εκτίμησης σημείου, αφού θεωρούμε ότι η συνολική μάζα πιθανότητας των λ_i είναι συγκεντρωμένη στην τιμή MAP, δηλαδή $\pi(\lambda|\mathbf{z}) = \delta(\lambda, \hat{\lambda})$. Η εκτίμηση αυτή δεν αποτελεί ωστόσο τον τελικό στόχο του προ-

βλήματος, ο οποίος είναι η πρόβλεψη του b^{n+1} δεδομένου του x^{n+1} και του συνόλου εκμάθησης $\mathbf{z}$. Δεδομένου λοιπόν ότι η Μπεϋσιανή συμπερασματολογία προβαίνει σε εκτιμήσεις σημείου (αν αυτό απαιτείται) μόνο κατά το τελικό στάδιο (εν προκειμένω, το στάδιο της πρόβλεψης), η συγκεκριμένη στρατηγική μπορεί να χαρακτηρισθεί μόνο ως ημι-Μπεϋσιανή.

Όπως θα δούμε στη συνέχεια, πολλές τεχνικές ρύθμισης που χρησιμοποιούνται στην κλασσική (συχνοτική) στατιστική, είναι ισοδύναμες με την εκτίμηση MAP και την επιβολή μίας συγκεκριμένης εκ των προτέρων κατανομής. Τέτοιοι ρυθμιστές μπορούν να στοχεύουν είτε στην αποφυγή μεγάλων τιμών στις $\hat{\lambda}$, είτε στην εξαγωγή αραιών (sparse) λύσεων, όπου προκρίνονται λύσεις με λίγα μόνο από τα διαθέσιμα χαρακτηριστικά. Στην πρώτη περίπτωση, συνήθης πρακτική είναι η προσθήκη όρου ρύθμισης τετραγωνικής νόρμας L_2, ενώ στη δεύτερη η προσθήκη νόρμας ρυθμιστή L_1, [65]. Χρήσιμο είναι επίσης να τονίσουμε τους δύο ρυθμιστές για γραμμικά μοντέλα παλινδρόμηση, την παλινδρόμηση Ringe και LASSO (Least Absolute Shrinkage and Selection Operator) για τις αντίστοιχες νόρμες, οι οποίες παρουσιάζουν τις παραπάνω ιδιότητες, [66].

Στην περίπτωσή μας, ωστόσο, οι συντελεστές λαμβάνουν μόνο θετικές τιμές. Έτσι, η εκ των προτέρων κατανομή θα πρέπει να πρέπει να αποδιδει μηδενική μάζα πιθανότητας στην περιοχή όπου ένας και πλέον συντελεστές έχουν αρνητική τιμή. Έτσι, επιλέγεται η χρήση της εκθετικής κατανομής

$$\pi(\lambda_i) = \alpha_i \exp(-\alpha_i \lambda_i) \tag{73}$$

έναντι της διπλής εκθετικής (ή αλλιώς Λαπλασιανής) κατανομής. Επιπλέον, θέτουμε μία κοινή τιμή στις υπερπαραμέτρους, δηλαδή $\alpha_i = \alpha, i = 1, 2, \ldots, k$.

Ο αλγόριθμος γενικευμένης αναδρομικής κλιμάκωσης θα έχει τώρα τον ακόλουθο τύπο ανανέωσης (update)

$$\lambda_i \leftarrow \max[0, \lambda_i + \frac{1}{f^{\#}} \log \frac{< \tilde{f}_i > -\alpha}{< f_i >}] \tag{74}$$

έναντι της εκτίμησης μέγιστης πιθανοφάνειας

$$\lambda_i \leftarrow \lambda_i + \frac{1}{f^{\#}} \log \frac{< \tilde{f}_i >}{< f_i >} \tag{75}$$

Η παραπάνω αναδρομή προκύπτει άμεσα από την μεγιστοποίηση μας η εκ των υστέρων κατανομή των λ_i, $\pi(\lambda|\mathbf{z})$, ως προς λ_i. Μετά από πράξεις καταλήγουμε στις συνθήκες

$$< \tilde{f}_i >=< f_i > -\alpha,\ i = 1, 2, \ldots, k \tag{76}$$

Η τεχνική αυτή είναι γνωστή ως discounting, και όπως είναι εμφανές ελαττώνει την αξιοπιστία των χαρακτηριστικών, μειώνοντας τις ορθές προβλέψεις στο σύνολο εκμάθησης κατά α σε σχέση με την εκτίμηση μέγιστης πιθανοφάνειας. Επιπλέον, όσο ο συντελεστής α αυξάνεται, τόσο περισσότερα χαρακτηριστικά αποκτούν μηδενική τιμή, οδηγώντας έτσι σε πιο αραιές εκτιμήσεις.

4.5 Η κατάτμηση του χώρου εισόδου και η μέθοδος κατωφλίωσης των αποστάσεων

4.5.1 Τα οφέλη της κατάτμησης του χώρου εισόδου

Ερχόμαστε τώρα να δούμε ένα από τα πλεονεκτήματα της προτεινόμενης μεθόδου, αυτό της κατάτμησης του χώρου εισόδου σε προεπιλεγμένες περιοχές και του πιθανοτικού συνδυασμού των αποκρίσεων των εξειδικευμένων ανά περιοχή μοντέλων σε ένα ενιαίο υπερμοντέλο. Το πρόβλημα που αντιμετωπίζουμε είναι αυτό της συμπερασματολογίας (inference) για το αν δύο τμήματα ομιλίας προέρχονται από τον ίδιο ($b = 0$) ή διαφορετικό ($b = 1$) ομιλητή. Αναφέρουμε την έννοια συμπερασματολογία, δεδομένου ότι μας ενδιαφέρει η μοντελοποίηση της εκ των υστέρων πιθανότητας του b και όχι απλώς η απόδοση του ζεύγους x σε μια από τις δύο κλάσεις. Δείξαμε ότι το εκθετικό μοντέλο είναι ένα εργαλείο επίλυσης ενός τέτοιου προβλήματος συμπερασματολογίας, καθώς έχει την ικανότητα να συνδυάζει πολλαπλούς ταξινομητές - οι οποίοι μπορεί να προέρχονται και από διαφορετικές ροές πληροφορίας (πολυτροπική σύμμιξη) - καθώς και τη δυνατότητα ενσωμάτωσης εκ των προτέρων πληροφορίας όσον αφορά στα βάρη των ταξινομητών.

Η κατάτμηση του χώρου εισόδου είναι μια στρατηγική *διαίρει και βασίλευε* η οποία ενδείκνυται για περιπτώσεις όπου απαιτείται εξειδικευμένη επιλογή χαρακτηριστικών και εκτίμηση παραμέτρων για κάθε περιοχή. Στο πρόβλημα που αντιμετωπίζουμε, τέτοιες περιοχές είναι η φύση του διαύλου

(εύρος ζώνης), οι συνθήκες ηχογράφησης, η ύπαρξη ή μη παρεμβολών όπως μουσικής ή ηχητικών εφέ, κ.ο.κ. Ένα επιπλέον χαρακτηριστικό κατάτμησης σχετίζεται με τις διάρκειες των δύο τμημάτων ομιλίας. Τμήματα φωνής μικρής διάρκειας (< 5 δευτερολέπτων) μοντελοποιούνται πιο στιβαρά με χρήση μίας κανονικής κατανομής, σε αντίθεση με τμήματα διάρκειας > 30 δευτερολέπτων, για τα οποία μπορούμε να χρησιμοποιήσουμε πιο σύνθετα μοντέλα, όπως μίγματα κανονικών κατανομών. Επιπλέον, η βέλτιστη κατωφλίωση των μετρικών απόκλισης διαφέρει ανάλογα με τις διάρκειες των τμημάτων ομιλίας. Εξαιρώντας το Μπεϋσιανό Κριτήριο Πληροφορίας (BIC), οι υπόλοιπες μετρικές απόκλισης (Kullback-Leibler, Arithmetic-Harmonic Sphericity, κ.ά.) δεν λαμβάνουν υπόψη τους τις διάρκειες των τμημάτων ομιλίας και ως εκ τούτου τη μεταβλητότητα στην εκτίμηση των παραμέτρων των κατανομών. Έτσι, το βέλτιστο κατώφλι είναι μεγαλύτερο για μικρά τμήματα ομιλίας συγκριτικά με μεγαλύτερα τμήματα. Μία λύση για αυτό το πρόβλημα είναι η συμπερίληψη ταξινομητών των οποίων η μέθοδος κατωφλίωσης προκύπτει πιο φυσικά, ως απόρροια της ίδιας της μοντελοποίησης (π.χ. το BIC στις διάφορες μορφές του). Μία άλλη λύση είναι η κατάτμηση του χώρου εισόδου σε περιοχές με βάση τις διάρκειες των τμημάτων. Στην τελευταία περίπτωση, η εκτίμηση του βέλτιστου κατωφλίου μπορεί να γίνει για κάθε μία από αυτές ξεχωριστά, σύμφωνα με ένα εξειδικευμένο, ανά περιοχή εισόδου, σύνολο εκμάθησης. Μία ανάλογη στρατηγική ακολουθείται και στο [67], με μετρική απόκλισης τον γενικευμένο λόγο πιθανοφάνειας.

4.5.2 Το υπερμοντέλο ως πιθανοτικός συνδυασμός των μοντέλων περιοχών εισόδου

Υποθέτουμε αρχικά ότι έχουμε κατατμήσει τον χώρο εισόδου σε M το πλήθος περιοχές και έστω $q_{\lambda^m}(b|x), m = 1, \ldots, M$ τα αντίστοιχα μοντέλα. Το υπερμοντέλο ορίζεται ως γραμμικός συνδυασμός των επιμέρους μοντέλων, ως εξής

$$q_\Lambda(b|x) = \sum_{m=1}^{M} w_m(x) q_{\lambda^m}(b|x), \tag{77}$$

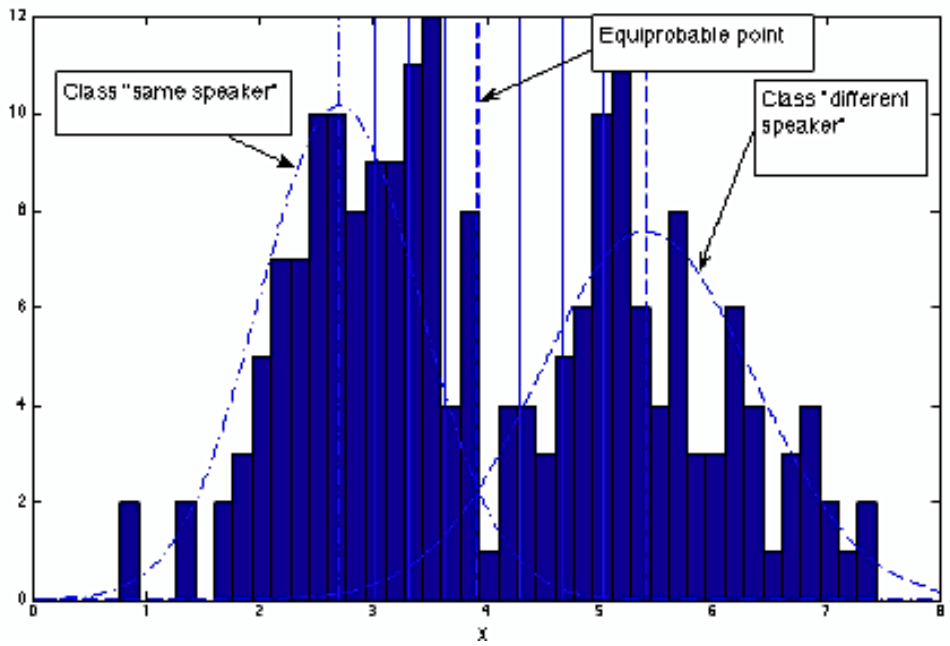
όπου

$$w_m(x) = p(m|x, \Theta), \sum_{m=1}^{M} w_m(x) = 1. \tag{78}$$

Παρατηρούμε ότι τα βάρη $w_m(x)$ ορίζονται ως οι εκ των υστέρων πιθανότητες κατάταξης του διανύσματος εισόδου x σε κάθε μία από τις M περιοχές. Λόγω της γραμμικότητας του $q_\Lambda(b|x)$ ως προς τα q_{λ^m}, με βάρη που αθροίζουν στη μονάδα, ισχύει ότι $\sum_b q_\Lambda(b|x) = 1$. Επομένως, το υπερμοντέλο συνεχίζει να εκφράζει την εκ των υστέρων πιθανότητα κατάταξης του ζεύγους x σε μία εκ των κλάσεων $b = \{0, 1\}$. Επιπλέον, υπερβαίνει την αποκλειστική κατάταξη του x σε μία μόνο ομάδα, αλλά παρέχει τη δυνατότητα πιθανοτικής κατάταξης του διανύσματος εισόδου x στις περιοχές εισόδου και συναπόφασης των υπομοντέλων (model averaging).

4.5.3 Η μέθοδος κατωφλίωσης των αποστάσεων

Στην παράγραφο αυτή περιγράφουμε τη μέθοδο πολλαπλής κατωφλίωσης των αποστάσεων. Μέσω των κατωφλιώσεων εξάγουμε N_{th} δυαδικούς ταξινομητές για κάθε απόσταση, όπου N_{th} ο αριθμός των κατωφλίων, τον οποίο εμπειρικά θέσαμε $N_{th} = 9$. Έστω, λοιπόν, n ζεύγη τμημάτων ομιλίας $z_j = (x_j, b_j), j = 1, \ldots, n$ από την κατηγορία εισόδου m. Χρησιμοποιούμε τη μεταβλητή x_j για να συμβολίσουμε το j-οστό ζεύγος και $b_j = \{0, 1\}$ για να συμβολίσουμε την κλάση *ίδιος* ή *διαφορετικός* ομιλητής. Για κάθε μία κλάση υπολογίζουμε τη μέση τιμή και τη μεταβλητότητα (μ_b^m, v_b^m) και υποθέτουμε ότι ακολουθεί μία κανονική κατανομή. Κατόπιν, υπολογίζουμε το ισοπίθανο σημείο και το ορίζουμε ως το κεντρικό κατώφλι. Τα υπόλοιπα κατώφλια υπολογίζονται έτσι ώστε το μικρότερο και το μεγαλύτερο να αντιστοιχούν στις μέσες τιμές μ_0^m και μ_1^m αντίστοιχα. Η διαδικασία που περιγράψαμε απεικονίζεται στο Σχήμα 5. Οι μπλε μπάρες δείχνουν το ιστόγραμμα των αποστάσεων, οι κάθετες ευθείες τα σημεία κατωφλίωσης ενώ με διακεκομμένες γραμμές σκιαγραφείται η κανονική κατανομή με την οποία προσεγγίζουμε το ιστόγραμμα κάθε κλάσης.

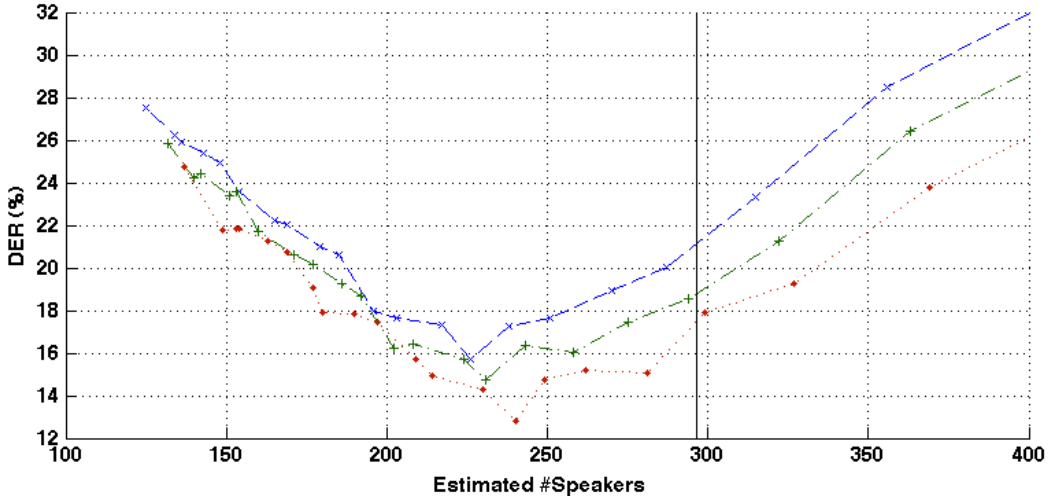


Σχῆμα 5: *Παράδειγμα πολλαπλών κατωφλιώσεων της KL2 απόκλισης. Κάθε ένα κατώφλι ορίζει και από έναν ταξινομητή.*

4.6 Πειραματικά αποτελέσματα

Για να εκπαιδεύσουμε το προτεινόμενο μοντέλο, χρησιμοποιήθηκε το σύνολο εκμάθησης της βάσης ESTER. Το σύνολο αυτό αποτελείται από 14 εκπομπές λόγου από το Γαλλικό ραδιόφωνο. Από το σύνολο αυτό εξάγαμε ζεύγη τμημάτων ομιλίας, τα οποία κατηγοριοποιήθηκαν βάσει του φύλου, των συνθηκών ηχογράφησης (στούντιο και εξωτερική ηχογράφηση) καθώς και των διαρκειών των δύο τμημάτων.

Από το Σχ. 6 παρατηρούμε ότι από μόνη της η κατηγοριοποίηση της εισόδου επιφέρει ένα κέρδος της τάξης του 1%. Η διαφορά μεταξύ μπλέ και πράσινης καμπύλης έγκειται στο ότι στην τελευταία η επιλογή της παραμέτρου λ πραγματοποιείται ανα κατηγορία ηχογράφησης. Επιπλέον, βλέπουμε ότι η αύξηση των χαρακτηριστικών σε 5 ανά κατηγορία βελτιώνει περαιτέρω τα αποτε-



Σχῆμα 6: *Εκτιμώμενος αριθμός ομιλητών και DER (%) για τις 14 εκπομπές της βάσης ESTER (σύνολο ανάπτυξης). Μπλε καμπύλες με 'x': Τοπικό-BIC, Πράσινες καμπύλες με σταυρούς: ένα χαρακτηριστικό ανά κατηγορία, Κόκκινες καμπύλες με τελείες: πέντε χαρακτηριστικά ανά κατηγορία.*

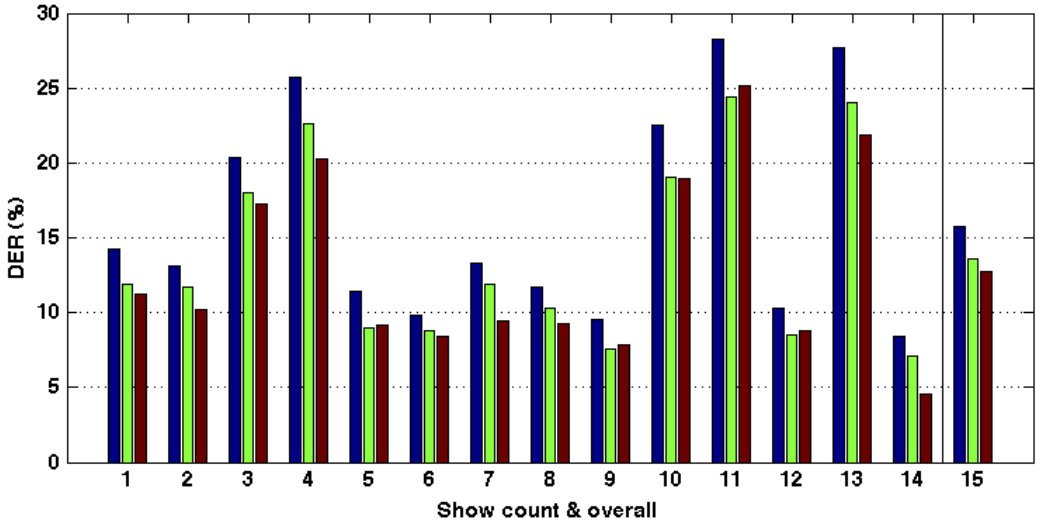
λέσματα.

Τα σφάλματα ομαδοποίησης ανά αρχείο παρουσιάζονται στο Σχ. 7 (σύνολο εκμάθησης) και Σχ. 8 (σύνολο αξιολόγησης).

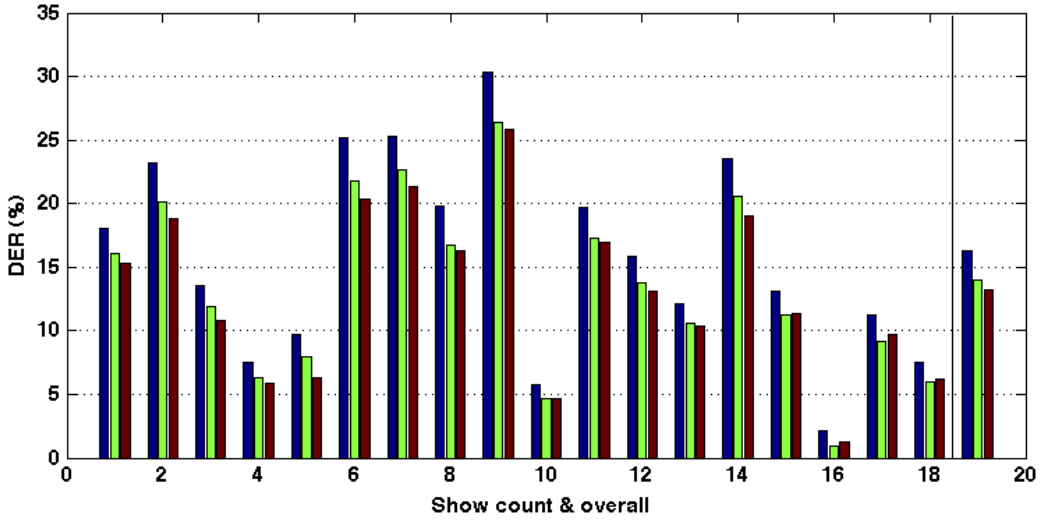
4.7 Επίλογος κεφαλαίου

4.7.1 Ανακεφαλαίωση

Στο κεφάλαιο αυτό προτείναμε μια μέθοδο συνδυασμού ενός συνόλου δυαδικών χαρακτηριστικών, με στόχο την προσέγγιση της εκ των υστέρων πιθανοφάνειας $p(b|x)$. Δείξαμε ότι το εκθετικό μοντέλο είναι ικανό να συνδυάσει επιτυχώς τέτοια χαρακτηριστικά, τα οποία μπορούν να προκύψουν από ένα σύνολο παραμετρικών χώρων (MFCC, LSF, κ,ά), μοντελοποιήσεων, στατιστικών αποστάσεωνκαι κατωφλιώσεων. Για να επιτύχουμε μια τέτοια μοντελοποίηση ορίσαμε τις αφηρημένες κλάσεις *ίδιος ομιλητής* και *διαφορετικοί ομιλητές* ($b = 0$ και $b = 1$ αντίστοιχα), και ως εκ τούτου



Σχῆμα 7: *Βέλτιστη επίδοση σε DER (%) για τις 14 εκπομπές της βάσης ESTER (σύνολο ανάπτυξης). Αριστερές, μπλε μπάρες: Τοπικό-BIC, Μέσες-πράσινες μπάρες: ένα χαρακτηριστικό ανά κατηγορία, Δεξιές μπάρες με κόκκινο χρώμα: πέντε χαρακτηριστικά ανά κατηγορία.*



Σχῆμα 8: *DER (%) για τις 18 εκπομπές της βάσης ESTER (σύνολο αξιολόγησης). Αριστερές, μπλε μπάρες: Τοπικό-BIC, Μέσες-πράσινες μπάρες: ένα χαρακτηριστικό ανά κατηγορία, Δεξιές μπάρες με κόκκινο χρώμα: πέντε χαρακτηριστικά ανά κατηγορία.*

μετατρέψαμε την ομαδοποίηση (clustering) ομιλητών σε πρόβλημα ταξινόμησης (classification). Έτσι, με βάση ένα επαρκές και αντιπροσωπευτικό προταξινομημένο δείγμα εκμάθησης (labeled training set) και ένα εκθετικό μοντέλο προσεγγίζουμε τη σ.μ.π. $q_\lambda(b|x)$.

Στα χαρακτηριστικά αυτής της προσέγγισης είναι η πιθανοτική μοντελοποίηση της $b|x$. Σε αντίθεση με τις μεθόδους οι οποίες χρησιμοποιούν μία και μόνη απόσταση και ένα προεπιλεγμένο κατώφλι, η προτεινόμενη μέθοδος είναι πιθανοτική, και ως εκ τούτου το κατώφλι απόφασης μπορεί να τεθεί σύμφωνα με τη θεωρία απόφασης. Έτσι, μπορούμε να το θέσουμε ίσο 0.5 εφόσον τα δεδομένα είναι κατάλληλα σταθμισμένα. Εναλλακτικά, μπορούμε να επιλέξουμε διαφορετική κατωφλίωση, σε περιπτώσεις που η συνάρτηση κόστους για τους δύο τύπους λαθών είναι μη-συμμετρική.

Επιπλέον, η πιθανοτική μοντελοποίηση επιτρέπει τη δημιουργία μιας ιεραρχικής δομής, ώστε το ανώτερο στρώμα να επιτελεί κατάτμηση του χώρου εισόδου $\mathcal{X}$ σε προκαθορισμένες κατηγορίες. Για κάθε κατηγορία εισόδου ένα διαφορετικό μοντέλο εκπαιδεύεται, το οποίο θα περιέχει τα χαρακτηριστικά εκείνα που είναι καταλληλότερα για κάθε κατηγορία. Για το νέο δείγμα x^{n+1}, η τελική κατανομή του $b^{n+1}|x^{n+1}$ προκύπτει ως μέσος όρος των κατανομών των μοντέλων κάθε κατηγορίας, με βάρη τις εκ των υστέρων κατανομές του x^{n+1} για κάθε κατηγορία.

Τέλος, δείξαμε ότι Μπεϋσιανές τεχνικές εκτίμησης σημείου των παραμέτρων μπορούν να εφαρμοσθούν. Συγκεκριμένα, εξετάσαμε την εκτίμηση MAP θέτοντας μια εκ των προτέρων κατανομή στις παραμέτρους του μοντέλου. Με τη μεγιστοποίηση της εκ των υστέρων πιθανότητας των παραμέτρων έναντι της εκτίμησης μέγιστης πιθανοφάνειας αποφεύγουμε το υπερταίριασμα των παραμέτρων στο συγκεκριμένο σύνολο εκμάθησης και επιτυγχάνουμε έτσι μικρότερο σφάλμα γενίκευσης σε νέα δεδομένα.

4.7.2 Προτεινόμενες επεκτάσεις

Η προτεινόμενη μέθοδος μπορεί να επεκταθεί ή και να μεταβληθεί με διάφορους τρόπους. Σε σχέση με τα χαρακτηριστικά που προτείναμε, χρήσιμο θα ήταν να συμπεριληφθούν και συνδυαστικά χαρακτηριστικά, τα οποία προκύπτουν από δυαδικές πράξεις μεταξύ των πρωτογενών χαρακτηριστικών.

Με τον τρόπο αυτών θα μπορούσαμε να ελαττώσουμε τον αριθμό των χαρακτηριστικών που ο αλγόριθμος τελικά επιλέγει, καθώς μοντελοποιείται και η συνδιακύμανση των πρωτογενών χαρακτηριστικών. Επιπλέον, πέραν των χαρακτηριστικών που βασίζονται στο ηχητικό φάσμα, το μοντέλο επιδέχεται και χαρακτηριστικά τα οποία προκύπτουν από άλλα ηχητικά χαρακτηριστικά (προσωδία, ύπαρξη σιωπών, κ.ά.) καθώς και από άλλες συνιστώσες, αν αυτές είναι διαθέσιμες (π.χ. εικονική συνιστώσα).

Τέλος, θα πρέπει να εξετασθούν μέθοδοι οι οποίες είναι αμιγώς Μπεϋσιανές. Οι μέθοδοι αυτές αποφεύγουν την εκτίμηση σημείου των παραμέτρων λ και επιχειρούν ολοκληρωμένη συμπεσματολογία ως προς αυτές. Στη συμπεσματολογία μάλιστα μπορεί να ενταχθεί και η ίδια η διάσταση του λ, ο αριθμός δηλαδή των χαρακτηριστικών τα οποία τελικά επιλέγονται. Η συμπερασματολογία επιτυγχάνεται είτε με τη χρήση του αλγορίθμου Reversible-jump Markov Chain Monte-Carlo είτε με τη χρήση μοντέλων άπειρης χωρητικότητας. Με τον τρόπο αυτόν, η εκτίμηση για το νέο δείγμα εξάγεται με ολοκλήρωση ως προς τις παραμέτρους αυτές βάσει της εκ των υστέρων κατανομής των λ.

Επιπλέον, η Μπεϋσιανή μοντελοποίηση είναι κατάλληλη για τη μοντελοποίηση της ιεραρχικής δομής. Πιο συγκεκριμένα, μας παρέχει τη δυνατότητα να προκρίνουμε επιλογές χαρακτηριστικών τα οποία να είναι αραιά (sparse), γεγονός που μειώνει τις υπολογιστικές απαιτήσεις και κυρίως βοηθάει στο να αποφύγουμε το υπερταίριασμα. Επιπλέον, μέσω της ιεραρχικής δομής μπορούμε να επιβάλουμε το μοίρασμα χαρακτηριστικών μεταξύ διαφορετικών μοντέλων. Έτσι, στην περίπτωση χρήσης της κατάτμησης του χώρου εισόδου, μπορούμε να επιβάλουμε στα επιμέρους μοντέλα τα χαρακτηριστικά τους να είναι από κοινού αραιά, να μοιράζονται δηλαδή αρκετά από τα χαρακτηριστικά τους, έτσι ώστε το σύνολο των επιλεγόμενων χαρακτηριστικών να είναι $o(M)$, να αυξάνεται δηλαδή υπογραμμικά με τον αριθμό των επιμέρους μοντέλων.

5 ΕΠΑΝΑΠΡΟΣΕΓΓΙΖΟΝΤΑΣ ΤΟ ΜΠΕΫΣΙΑΝΟ ΚΡΙΤΗΡΙΟ ΠΛΗΡΟΦΟΡΙΑΣ ΓΙΑ ΤΟ ΠΡΟΒΛΗΜΑ ΤΗΣ ΟΜΑΔΟΠΟΙΗΣΗΣ ΟΜΙΛΗΤΩΝ

Περίληψη

Στο κεφάλαιο αυτό αναφερόμαστε στη χρήση του Μπεϋσιανού Κριτηρίου Πληροφορίας (BIC) στο πρόβλημα της ομαδοποίησης ομιλητών. Οι βασικοί μας στόχοι είναι η εξέταση των διαφορών των μεταξύ των δύο βασικών εναλλακτικών διατυπώσεών του (ολικού και τοπικού) και η παρουσίαση μιας εναλλακτικής του μορφής, του τμηματικού. Στη συνέχεια εξετάζουμε τις ασυμπτωτικές ιδιότητες του BIC στην περίπτωση όπου τα στατιστικά μοντέλα που χρησιμοποιούμε είναι ελλειπώς ορισμένα (misspecified). Το βασικό μας αποτέλεσμα είναι μια έκδοση του BIC, το τμηματικό-BIC-τετραγωνικής ρίζας, με το οποίο επιτυγχάνουμε ακρίβεια ομαδοποίησης πολύ κοντά σε αυτήν που επιτυγχάνουν συστήματα ιδιαίτερα απαιτητικά σε χρόνο επεξεργασίας.

5.1 Εισαγωγή

Μία από τις θεμελιώδεις ιδιότητες της Μπεϋσιανής στατιστικής είναι η συστηματικότητα με την οποία μπορούν να προσεγγισθούν προβλήματα δευτέρου επιπέδου συμπερασματολογίας. Σε αυτό εντάσσονται η επιλογή της κατάλληλης τάξης μοντέλου (model selection), καθώς και του συνδυασμού μοντέλων διαφορετικών τάξεων (model averaging) με στόχο μια πιο ολοκληρωμένη προσέγγιση της κατανομής πρόβλεψης για νέα δεδομένα. Όπως αναφέρουμε στο Κεφ. 2, η ποσότητα-κλειδί της Μπεϋσιανής στατιστικής κατά το δεύτερο επίπεδο συμπερασματολογίας είναι η ολοκληρωμένη (ή οριακή) πιθανοφάνεια. Η ποσότητα αυτή προκύπτει από την ολοκλήρωση του γινομένου της συνάρτησης πιθανοφάνειας επί την εκ των προτέρων κατανομή των παραμέτρων του μοντέλου. Το BIC είναι ένα κριτήριο επιλογής μοντέλου που χρησιμοποιείται κατά κόρον σε περιπτώσεις

όπου είτε η ολοκληρωμένη πιθανοφάνεια δεν έχει κλειστό τύπο υπολογισμού, είτε επιθυμούμε να αποφύγουμε τον ορισμό μίας εκ των προτέρων κατανομής. Εισήχθη από τον G. Schwarz στο [68], όπου απέδειξε ότι για την εκθετική οικογένεια κατανομών, το κριτήριο προσεγγίζει την λογαριθμική ολοκληρωμένη πιθανοφάνεια με σφάλμα $\mathcal{O}(1)$.

Εντός της κοινότητας που ασχολείται με την ομαδοποίηση ομιλητών, το BIC χρησιμοποιείται ευρύτατα τόσο για την κατάτμηση όσο και για την ομαδοποίηση των τμημάτων ομιλίας, [69], [25]. Παρά την ευρύτατη χρήση του, όμως, αρκετά σχετικά θέματα δεν έχουν εξετασθεί επαρκώς. Ποια η βασική διαφορά μεταξύ ολικού και τοπικού BIC; Υφίσταται Μπεϋσιανή αιτιολόγηση πίσω από τη χρήση της παραμέτρου ρύθμισης; Μπορεί να προκύψει κάποιο ώφελος από τη χρήση εκ των προτέρων κατανομών στον χώρο των ομαδοποιήσεων; Και το σημαντικότερο, μπορούμε να εξετάσουμε νέους όρους ποινής από αυτούς που προτείνει το BIC, οι οποίοι να είναι πιο συμβατοί με τη φύση των δεδομένων μας;

Η συνέχεια του κεφαλαίου έχει ως ακολούθως. Στην παράγραφο 5.2, η χρήση του BIC στο πρόβλημα της κατάτμησης ομιλητών εξετάζεται, μαζί με τις εκφράσεις των διαφόρων εναλλακτικών προσεγγίσεών του. Στην παράγραφο 5.4, η μαθηματική εξαγωγή του κριτηρίου παρατίθεται, ενώ αρκετά ακόμα θεωρητικά ζητήματα εξετάζονται. Στις παραγραφους 5.5 και 5.6, διατυπώνουμε το αναθεωρημένο κριτήριο, και χρησιμοποιούμε παραδείγματα για να το θεμελιώσουμε και πειραματικά. Τέλος, πειραματικά αποτελέσματα παρουσιάζονται στην παράγραφο 5.7, βασισμένα σε επίσημες βάσεις κατάτμησης και ομαδοποίησης ομιλητών.

5.2 Η χρήση του BIC στο πρόβλημα της ομαδοποίησης ομιλητών

5.2.1 Συμβολισμοί μεταβλητών

Αρχικά παραθέτουμε τους συμβολισμούς των μεταβλητών που θα χρησιμοποιήσουμε. Ορίζουμε τα διανύσματα παρατήρησης ως $\mathbf{y} = [\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(n)}] \in \mathcal{Y}^n \subseteq \Re^{d \times n}$ όπου $i = 1, \ldots, n$ αντιστοιχεί στον χρόνο ενώ d είναι η διάσταση του χώρου χαρακτηριστικών. Το αντίστοιχο σύνολο των κρυφών με-

ταβλητών (ή ομαδοποιήσεων, ή ακολουθιών καταστάσεων στη Μαρκοβιανή ορολογία) τον ορίζουμε ως $\mathbf{x} = [x^{(1)}, \ldots, x^{(n)}] \in \mathcal{X}^n$, όπου $x^{(i)} \in [1, \ldots, K]$ ο αύξων αριθμός της ομάδας (ή κατάστασης στην Μαρκοβιανή ορολογία) που ανήκει το διάνυσμα $\mathbf{y}^{(i)}$ και $\max(\mathbf{x}) = K$, υποθέτοντας μία ομαδοποίηση του $\mathbf{y}$ σε K ομάδες. Χρησιμοποιούμε τον δείκτη $\mathbf{y}_k$ για να αναφερθούμε στα διανύσματα παρατήρησης τα οποία ομαδοποιούνται στην k-ομάδα, δηλαδή $\mathbf{y}_k = \{\mathbf{y}^{(i)} : x^{(i)} = k\}$. Επιπλέον, $n_k = \sum_{i=1}^{n} \delta(x^{(i)}, k)$ όπου $\delta(\cdot, \cdot)$ είναι η συνάρτηση δέλτα του Kronecker και $\sum_{k=1}^{K} n_k = n$. Η συνάρτηση πυκνότητας πιθανότητας (σ.π.π.) που αντιστοιχεί στην k-ομάδα ορίζεται ως $f(\cdot|\varphi_k)$ και παραμετροποιείται από τις $\varphi_k \in \Phi \subseteq \Re^P$. Δεδομένου ότι θα ασχοληθούμε μόνο με κανονικές κατανομές (και όχι με μίγματα), ισχύει $\varphi_k = (\mu_k, \Sigma_k)$ και επομένως $P_k = P = d + d(d+1)/2$. Θα αναφερόμαστε στις παραμέτρους αυτές σαν τις *εσωτερικές* παραμέτρους του μοντέλου, σε αντίθεση με τις *εξωτερικές* παραμέτρους οι οποίες αντιστοιχούν στον πίνακα μεταβάσεων μεταξύ καταστάσεων, σε ορολογία Μαρκοβιανών μοντέλων. Το παραμετρικό διάνυσμα ορίζεται ως εξής $\varphi = [\varphi_1, \ldots, \varphi_K]$, $\varphi \in \Phi^K$.

Στο κεφάλαιο αυτό θα προτείνουμε τις δικές μας τροποποιήσεις του BIC. Κεντρικό ρόλο στις προτάσεις μας έχει η ανάλυση των Kass-Wasserman [70] που αφορά στις εκ των προτέρων κατανομές που υπονοούνται από το BIC. Η ανάλυση αυτή αποτελεί ένα χρήσιμο εργαλείο το οποίο δεν έχει εξετασθεί με επάρκεια στη βιβλιογραφία της ομαδοποίησης ομιλητών. Βάσει αυτής, θα δείξουμε ότι η διαφορά μεταξύ ολικού και τοπικού BIC μπορεί να γεφυρωθεί, μέσω μιας αναθεώρησης των εκ των προτέρων κατανομών, η οποία θα έχει επίπτωση στον όρο ποινής. Το νέο κριτήριο θα το ονομάσουμε Τμηματικό-BIC, (Segmental-BIC) και θα προέλθει από την απαίτησή μας να δρα τοπικά ως αυτόνομη μετρική απόστασης (όπως το τοπικό) ενώ ταυτόχρονα να προκύπτει ως προσέγγιση της ολοκληρωμένης πιθανοφάνειας (όπως το ολικό). Επιπλέον, θα επιχειρήσουμε να ενσωματώσουμε την παράμετρο ρύθμισης του BIC στη Μπεϋσιανή μοντελοποίηση, υιοθετώντας εκ των προτέρων κατανομές που εξαρτώνται από το πλήθος των παρατηρήσεων (sample-size dependent priors). Με βάση αυτήν την παραδοχή, θα εξετάσουμε εκ νέου τον όρο ποινής και το τροποποιημένο πλέον κριτήριο θα το ονομάσουμε τμηματικό-BIC ρίζας. Τέλος, θα μελετήσουμε τρόπους ενσωμάτωσης των Μαρκοβιανών δυναμικών στο κριτήριο όπως αυτές ορίζονται από τις εξισώσεις συστήματος των HMMs, έτσι ώστε να αποδίδουμε μία μάζα πιθανότητας $\pi(\mathbf{x})$ σε κάθε

ομαδοποίηση $\mathbf{x} \in \mathcal{X}^n$. Η εξαγωγή της πιθανότητας αυτής θα προκύπτει μέσω Dirichlet εκ των προτέρων κατανομών σε κάθε γραμμή του πίνακα μετάβασης μεταξύ καταστάσεων. Η τελευταία ανάλυση είναι μια συστηματική απόπειρα απόδοσης μικρής πιθανότητας σε ομαδοποιήσεις $\mathbf{x} \in \mathcal{X}^n$ που περιέχουν ταχείες μεταβάσεις μεταξύ των καταστάσεων και οι οποίες είναι μη ρεαλιστικές για το πρόβλημα ομαδοποίησης ομιλητών.

5.3 Η εξαγωγή του BIC ως προσέγγιση της ολοκληρωμένης πιθανοφάνειας ενός μοντέλου

5.3.1 Τα πεδία εφαρμογής του BIC

Για να προχωρήσουμε στην ανάλυση των προτάσεών μας, είναι απαραίτητη η εξέταση του BIC ως κριτηρίου εκτίμησης της φειδωλής τάξης μιας μοντελοποίησης, με βάση τα δεδομένα $\mathbf{y} \in \mathcal{Y}^n$, $\mathcal{Y} \subset \mathbb{R}^d$. Θα πρέπει να τονίσουμε ότι το BIC δεν προτάθηκε αρχικά ως κριτήριο εκτίμησης τάξης για προβλήματα ομαδοποίησης. Τα προβλήματα τα οποία επιχείρησε να αντιμετωπίσει ήταν η εκτίμηση της φειδωλής τάξης πολυωνυμικής παλινδρόμησης (multinomial regression), πολυβηματικών μαρκοβιανών αλυσίδων (multi-step Markov Chain) και άλλων παρεμφερών προβλημάτων. Η ασθενής συνέπεια (weak consistency) του BIC στην εκτίμηση της πραγματικής τάξης K έχει αποδειχθεί για μια σειρά τέτοιων προβλημάτων, όχι όμως και για προβλήματα ομαδοποίησης, όπως τα Μίγματα Κατανομών ή τα HMMs. Αυτό σημαίνει ότι ακόμα και για $n \to \infty$, το BIC δεν εκτιμάει το πραγματικό πλήθος των ομάδων με πιθανότητα 1.

5.3.2 Μπεϋσιανή σύγκριση μοντέλων και οριακή πιθανοφάνεια

Εξετάζουμε τώρα την εξαγωγή του BIC ως προσέγγιση της οριακής πιθανοφάνειας ενός μοντέλου. Έστω $\mathcal{M}$ ο χώρος των μοντέλων που αποτελείται από τα μοντέλα $M_j \in \mathcal{M}, j = 1, \ldots, |\mathcal{M}|$, με παραμέτρους $\theta \in \Theta \subseteq \mathbb{R}^{k_j}$, όπου k_j αντιστοιχεί στον αριθμό των ελεύθερων παραμέτρων του μοντέλου M_j. Στη Μπεϋσιανή συμπερασματολογία, η *εκ των υστέρων* πιθανότητα $p(M_j|\mathbf{y})$ ενός μοντέλου M_j περικλείει όλη τη γνώση που αποκτούμε από τον συνδυασμό της εκ των προτέρων γνώσης και αυτής που προκύπτει από τις παρατηρήσεις $\mathbf{y}$. Η εκ των υστέρων πιθανότητα του μοντέλου M_j είναι ανάλογη της από κοινού πιθανότητας (joint probability)

$$p(M_j|\mathbf{y}) \propto p(M_j)p(\mathbf{y}|M_j) \tag{79}$$

Υποθέτοντας ομοιόμορφη εκ των προτέρων κατανομή πιθανότητας στο σύνολο των εναλλακτικών υποθέσεων, το πρόβλημα καταλήγει στην εύρεση της υπόθεσης με τη μέγιστη οριακή πιθανοφάνεια (marginal likelihood ή evidence) $p(\mathbf{y}|M_j) = \int_\Theta p(\mathbf{y}|\theta, M_j)\pi_j(\theta)d\theta$. Εκτός από ορισμένες περιπτώσεις, το παραπάνω ολοκλήρωμα δεν έχει αναλυτική φόρμα υπολογισμού και έτσι είναι αναγκαία η χρήση προσεγγίσεων. Σε αντίθεση με την Variational Bayes εκμάθηση ή τις τεχνικές Monte-Carlo, το ολοκλήρωμα προσεγγίζεται με τη μέθοδο Laplace.

5.3.3 Η μέθοδος Laplace

Έστω ότι επιθυμούμε να προσεγγίσουμε το ολοκλήρωμα ενός μη-κανονικοποιημένου πιθανοτικού μέτρου $Z_P = \int \bar{P}(x)dx$, η οποία μεγιστοποιείται στο σημείο x_0. Αναπτύσσοντας τον λογάριθμο της $\bar{P}(x)$ γύρω από το μέγιστο έχουμε

$$\log \bar{P}(x) \approx \log \bar{P}(x_0) - \frac{c}{2}(x - x_0)^2 + \ldots \tag{80}$$

όπου

$$c = -\frac{\partial^2}{\partial x^2} \log \bar{P}(x) \mid_{x=x_0} \tag{81}$$

Με τον τρόπο αυτόν, προσεγγίζουμε την $\bar{P}(x)$ μέσω μίας κανονικής κατανομής,

$$\bar{Q}(x) = \bar{P}(x_0) \exp\left(-\frac{c}{2}(x - x_0)^2\right) \tag{82}$$

της οποία η σταθερά κανονικοποίησης είναι γνωστή,

$$Z_Q = \bar{P}(x_0)\sqrt{\frac{2\pi}{c}} \tag{83}$$

Η επέκταση της μεθόδου Laplace για d-διάστατες μεταβλητές $y \in \mathbb{R}^d$ είναι ευθεία. Ορίζουμε τον πίνακα των παραγώγων β′ τάξης της $-\log \bar{P}(x)$ (Hessian) στο μέγιστο $\mathbf{y}_0$ ως H, όπου

$$H_{ij} = -\frac{\partial^2}{\partial y_i \partial y_j} \log \bar{P}(\mathbf{y}) \mid_{\mathbf{y}=\mathbf{y}_0} \tag{84}$$

και η σταθερά κανονικοποίησης προκύπτει ως

$$Z_P \approx Z_Q = \bar{P}(\mathbf{y}_0)\sqrt{\frac{(2\pi)^d}{|H|}} \tag{85}$$

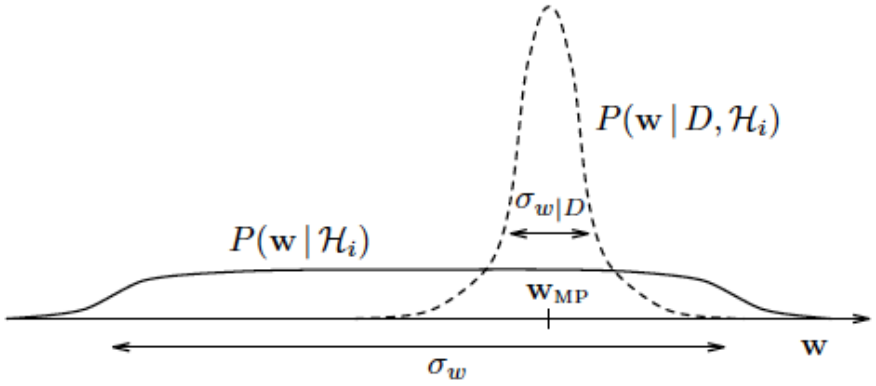
5.3.4 Οριακή πιθανοφάνεια και ο κανόνας του Occam

Πριν προχωρήσουμε στην εξαγωγή του κριτηρίου του Schwarz, είναι χρήσιμο να εξετάσουμε τον τρόπο με τον οποίο η οριακή πιθανοφάνεια εκφράζει την επιθυμητή διελκυστίνδα μεταξύ της ικανότητας ενός μοντέλου να περιγράφει τα δεδομένα (μέσω της πιθανοφάνειας) και ταυτόχρονα να χρησιμοποιεί τους ελάχιστους δυνατούς πόρους για την επίτευξη του σκοπού αυτού (ποινή στην πολυπλοκότητα της μοντελοποίησης). Η παραπάνω διελκυστίνδα καλείται και αρχή (ή ξυράφι) του Occam που προτείνει ότι μεταξύ των εξηγήσεων οι οποίες εκφράζουν με (σχεδόν) ισοδύναμη ακρίβεια ένα σύνολο δεδομένων, θα πρέπει να επιλέξουμε την απλούστερη εξ αυτών.
Η εφαρμογή της μεθόδου Laplace οδηγεί στην ακόλουθη προσέγγιση

$$P(D|M_j) \approx P(D|\hat{\mathbf{w}}, M_j) \times P(\hat{\mathbf{w}}|M_j)\sigma_{\mathbf{w}|D} \tag{86}$$

όπου συμβολίζουμε τα μονοδίαστατα δεδομένα ως D και τις παραμέτρους υπό την υπόθεση M_j ως $\mathbf{w}$. Η ποσότητα $P(D|\hat{\mathbf{w}}, M_j)$ εκφράζει την ικανότητα του μοντέλου να περιγράψει τα δεδομένα

D και ταυτίζεται ασυμπτωτικά με τη μέγιστη πιθανοφάνεια του $\mathbf{w}$. Για να δείξουμε πώς ο όρος $P(\hat{\mathbf{w}}|M_j)\sigma_{\mathbf{w}|D}$ ποινικοποιεί την πολυπλοκότητα της μοντελοποίησης που εκφράζει η M_j, θα θεωρήσουμε ότι η εκ των προτέρων κατανομή των παραμέτρων είναι ομοιόμορφη για μια περιοχή $\sigma_{\mathbf{w}}$. Όσο πιο πολύπλοκη είναι η μοντελοποίηση, τόσο πιο ευρεία θα είναι και αυτή η περιοχή. Επομένως, η $P(\hat{\mathbf{w}}|M_j) = \frac{1}{\sigma_{\mathbf{w}}}$ θα μειώνεται, όσο η πολυπλοκότητα της M_j αυξάνεται. Ο παράγοντας του Occam εκφράζεται σχηματικά ως $\frac{\sigma_{\mathbf{w}|D}}{\sigma_{\mathbf{w}}} < 1$ και παρουσιάζεται στο Σχήμα 9.



Σχῆμα 9: *Αναπαράσταση του κανόνα του Occam*

.

Για πολυδιάστατες παραμέτρους, η *εκ των υστέρων αμφιβολία* (posterior uncertainty) που εκφράζεται σχηματικά ως $\sigma_{\mathbf{w}|D}$ είναι πλέον ο Hessian H. Ο H καλείται πίνακας παρατηρούμενης πληροφορίας περιέχεται στα δεδομένα για τις παραμέτρους και εκφράζει την τοπική κυρτότητα (curvature) της εκ των υστέρων κατανομής. Είναι χρήσιμο να παρατηρήσουμε ότι για μη-πολωμένη εκτίμηση, το στοιχείο του $H^{-1}_{i,i}$ είναι το κάτω φράγμα της μεταβλητότητας της εκτίμησης της i-οστής παραμέτρου, το γνωστό Cramer-Rao φράγμα. Με βάση το φράγμα αυτό και τη θεωρία πληροφορίας του Shannon, ο Rissenan (′78) πρότεινε την αρχή του Ελάχιστου Μήκους Περιγραφής (Minimum Description Length, [71]) και κατέληξε σε ένα κριτήριο πολυπλοκότητας σχεδόν ταυτόσημο με το Μπεϋσιανό Κριτήριο Πληροφορίας. Η δική μας ανάλυση πάντως θα βασισθεί αποκλειστικά στην Μπεϋσιανή εκδοχή του.

5.3.5 Οι μαθηματικές εκφράσεις των κριτηρίων

Αρχικά, παρουσιάζουμε τις εκφράσεις των διαφορετικών κριτηρίων. Έστω ένα αρχείο από n διανύσματα παρατήρησης $\mathbf{y}$ και μία ομαδοποιήσή του $\mathbf{x}$ σε K ομάδες. Η έκφραση του ολικού-BIC έχει ως εξής

$$BIC^G = l(\hat{\varphi}; \mathbf{x}|\mathbf{y}) - \lambda K \frac{P}{2} \log n \tag{87}$$

ενώ η αντίστοιχη του τμηματικού ως

$$BIC^S = l(\hat{\varphi}; \mathbf{x}|\mathbf{y}) - \lambda \frac{P}{2} \sum_{k=1}^{K} \log n_k \tag{88}$$

Η συνάρτηση $l(\hat{\varphi}; \mathbf{x}|\mathbf{y})$ συμβολίζει τη λογαριθμική πιθανοφάνεια (classification log-likelihood, [9]) των φ δεδομένων των $\mathbf{x}$, υπολογισμένη στην εκτίμηση μέγιστης πιθανοφάνειας (ML) $\hat{\varphi}$ για κάθε ομάδα, δηλαδή

$$l(\hat{\varphi}; \mathbf{x}|\mathbf{y}) = \sum_{i=1}^{n} \log f(\mathbf{y}^{(i)}|\hat{\varphi}_{x^{(i)}}) \tag{89}$$

Σημειώνουμε επίσης ότι η ποσότητα $\hat{\varphi}$ θα πρέπει να ερμηνευθεί ως K διαφορετικές εκτιμήσεις $\{\hat{\varphi}_k\}_{k=1}^K$ (δηλαδή μία για κάθε ομάδα), δεδομένων των $\mathbf{y}$ και μιάς εκτίμησης των $\mathbf{x}$. Αυτό ισχύει γιατί δεν εφαρμόζουμε κάποιον EM αλγόριθμο ώστε να θεωρείται το $\hat{\varphi}$ ως (έστω τοπικό) μέγιστο δεδομένου του K, αλλά απλώς υπολογίζουμε τη δεσμευμένη ως προς $\mathbf{x}$ πιθανοφάνεια των παραμέτρων στα δεδομένα, όπου το $\mathbf{x}$ μπορεί να λάβει οποιαδήποτε τιμή $\mathbf{x} \in \mathcal{X}^n$.

Σαν απόρροια της παρατήρησης αυτής, και σε συνδυασμό με άλλα χαρακτηριστικά του προβλήματος, προτείνουμε μία εναλλακτική στρατηγική και περιλαμβάνουμε στους όρους ποινής μόνο το ποσοστό των δειγμάτων που χρησιμοποιούνται για την εκτίμηση κάθε παραμέτρου. Αυτό αποτυπώνεται στη σχέση (88). Παρατηρείστε ότι στις παραμέτρους φ_k αντιστοιχούμε όρο ποινής ίσο με $P \log n_k$, ενώ το ολικό κριτήριο αντιστοιχεί $P \log n$ για κάθε ομάδα.

Και τα δύο κριτήρια προσεγγίζουν την ολοκληρωμένη πιθανοφάνεια κατάταξης, δηλαδή

$$p(\mathbf{y}|\mathbf{x}) = \int_{\Phi^K} e^{l(\varphi; \mathbf{x}|\mathbf{y})} \pi(\varphi|\mathbf{x}) d\varphi \tag{90}$$

Υποθέτουμε τώρα ότι για να εκτιμήσουμε το $\mathbf{x}$ επιλέγουμε να χρησιμοποιήσουμε τα παραπάνω κριτήρια με έναν αλγόριθμο που βασίζεται σε αποστάσεις μεταξύ ζευγών. Επιλογές είναι η ιεραρχική

ομαδοποίηση, η φασματική ομαδοποίηση, [72] και άλλες. Σύμφωνα με τα παραπάνω, οι αποστάσεις αυτές (ΔBIC) θα έχουν ως εξής

$$\Delta BIC^G = \log GLR(\mathbf{y}_k, \mathbf{y}_l) - \lambda \frac{P}{2} \log n \tag{91}$$

και

$$\Delta BIC^S = \log GLR(\mathbf{y}_k, \mathbf{y}_l) - \lambda \frac{P}{2} \log \frac{n_k \times n_l}{n_k + n_l} \tag{92}$$

Η συντομογραφία GLR συμβολίζει τον γενικευμένο λόγο πιθανοφάνειας (Generalized Likelihood Ratio), που ορίζεται ως ακολούθως

$$GLR(\mathbf{y}_k, \mathbf{y}_l) = \frac{\prod_{\mathbf{y}^{(i)} \in \mathbf{y}_k} f(\mathbf{y}^{(i)}|\hat{\varphi}_k) \prod_{\mathbf{y}^{(i)} \in \mathbf{y}_l} f(\mathbf{y}^{(i)}|\hat{\varphi}_l)}{\prod_{\mathbf{y}^{(i)} \in \mathbf{y}_{k \cup l}} f(\mathbf{y}^{(i)}|\hat{\varphi}_{k \cup l})} \tag{93}$$

όπου $\mathbf{y}_{k \cup l}$ σημαίνει $\mathbf{y}_k \cup \mathbf{y}_l$ και $\hat{\varphi}_{k \cup l}$ τις εκτιμήσεις ML δεδομένων των $\mathbf{y}_{k \cup l}$.
Οι εκφράσεις των ΔBIC είναι ο λογαριθμικός λόγος των ολοκληρωμένων πιθανοφανειών των δύο μοντέλων.
Σημειώνουμε ότι η ερμηνεία που δίνει η σχολή των κλασσικών (συχνοτικών) στατιστικών είναι ο έλεγχος υποθέσεων

- $\mathcal{H}_0$: Τα δύο εναλλακτικά μοντέλα έχουν ασυμπτωτικά την ίδια ικανότητα περιγραφής των δεδομένων.
- $\mathcal{H}_1$: Το πιο σύνθετο μοντέλο έχει ασυμπτωτικά ανώτερη ικανότητα περιγραφής των δεδομένων.

Πάνω στον λόγο των πιθανοφανειών (και όχι των ολοκληρωμένων πιθανοφανειών) η σχολή αυτή επιχειρεί να εξετάσει την ασυμπτωτική συμπεριφορά του. Στη Μπεϋσιανή σχολή αντίθετα, ο λόγος που χρησιμοποιείται για τον έλεγχο υποθέσεων είναι ο λόγος των ολοκληρωμένων πιθανοφανειών (γνωστός και ως Παράγοντας Bayes) και ως εκ τούτου μπορούμε να θεωρήσουμε το μηδέν σαν κατώφλι. Έτσι, αν ΔBIC< 0, απορρίπτουμε την εναλλακτική υπόθεση $\mathcal{H}_1$ και θεωρούμε πιο πιθανή

την $\mathcal{H}_0$, όπου με ΔBIC εννοούμε ΔBIC ($\mathcal{H}_1/\mathcal{H}_0$).
Το τοπικό -ΔBIC ορίζεται ως

$$\Delta BIC^L = \log GLR(\mathbf{y}_k, \mathbf{y}_l) - \lambda \frac{P}{2} \log(n_k + n_l) \tag{94}$$

δηλαδή περιλαμβάνει μόνο το άθροισμα του πλήθους των δειγμάτων των δύο ομάδων αντί του n.

5.3.6 Το τμηματικό-BIC σαν απόπειρα συνδυασμού των ιδιοτήτων ολικού και τοπικού

Από τις παραπάνω εκφράσεις, μπορούμε να επισημάνουμε δύο βασικές διαφορές μεταξύ του ολικού και τοπικού.

- Το ολικό BIC ορίζει ένα μέγεθος για το $\mathbf{x}$, που προσεγγίζει την λογαριθμική πιθανοφάνεια κατάταξης του μοντέλου, ενώ το τοπικό ορίζει μόνο μία ΔBIC έκφραση.
- Το τοπικό ΔBIC είναι μία αυτόνομη μετρική απόστασης, δηλαδή ορίζεται πλήρως από τις επαρκείς στατιστικές των παραμέτρων και του πλήθους δειγμάτων των $\mathbf{y}_k$ και $\mathbf{y}_l$, ενώ το ολικό-ΔBIC απαιτεί το n για να ορισθεί.

Πολλαπλά πειράματα σε επίσημες βάσεις αξιολόγησης αποδεικνύουν ότι μία αυτόνομη μετρική απόκλισης έχει καλύτερα αποτελέσματα, [44], [51]. Ωστόσο, το τοπικό ΔBIC δείχνει να είναι αρκετά περιοριστικό, αφού ορίζει μόνο μια ΔBIC έκφραση. Ως εκ τούτου, μπορεί να χρησιμοποιηθεί μόνο με αλγόριθμους που βασίζονται σε ανά δύο αποστάσεις. Δύο ομαδοποιήσεις $\mathbf{x}^a, \mathbf{x}^b$ του $\mathbf{y}$ δεν μπορούν να συγκριθούν, παρά μόνο αν διαφέρουν σε μία και μόνο μία ένωση ενός ζεύγους. Το μειονέκτημα αυτό είναι αποτέλεσμα του γεγονότος ότι το τοπικό-ΔBIC δεν εξάγεται από μια έκφραση του BIC και έτσι δεν αντιστοιχεί σε αύξηση της ολοκληρωμένης πιθανοφάνειας κατάταξης. Εν μέρει, λοιπόν, η πρότασή μας (τμηματικό BIC) είναι μια απόπειρα εκπλήρωσης και

των δύο ιδιοτήτων, χωρίς τους περιορισμούς του τοπικό ΔBIC. Από τις σχέσεις (88) και (92), επαληθεύεται άμεσα ο ισχυρισμός μας.

5.4 Το BIC ως προσέγγιση της λογαριθμικής ολοκληρωμένης πιθανοφάνειας κατάταξης

Η παράγραφος αυτή καλύπτει ορισμένα από τα θεωρητικά ζητήματα της χρήσης των παραπάνω κριτηρίων, συμπεριλαμβανομένων των υπονοούμενων από αυτά των εκ των προτέρων κατανομών (implied priors), την πιθανοφάνεια κατάταξης και τον ρόλο της παραμέτρου ρύθμισης λ στη μοντελοποίηση.

5.4.1 Η εκ των υστέρων πιθανότητα ομαδοποίησης και το BIC

Συμβατότητα μεταξύ διαδικασίας και συνάρτησης κόστους

Ένα κεφαλαιώδες ζήτημα σε σχέση με τη χρήση του BIC στην ομαδοποίηση ομιλητών έγκειται στην στατιστική ποσότητα την οποία επιχειρούμε να εκτιμήσουμε. Με τη χρήση της πιθανοφάνειας κατάταξης $p(\mathbf{y}|\mathbf{x})$ αντί της ολοκληρωμένης πιθανοφάνειας του μοντέλου στα δεδομένα $p(\mathbf{y}|K)$, η ποσότητα αυτή γίνεται η ομαδοποίηση $\mathbf{x}$ έναντι της τάξης του μοντέλου K. Η επιλογή αυτή είναι σε πλήρη ευθυγράμμιση με τη συνάρτηση κόστους (loss function) που υπονοείται από το σφάλμα με το οποίο αξιολογούμε το σύστημα, δηλαδή το σφάλμα ομαδοποίησης, (Diarization Error Rate, DER). Το σφάλμα αυτό ορίζεται ως η απόσταση Hamming μεταξύ της εκτίμησης $\hat{\mathbf{x}}$ και της πραγματικής $\mathbf{x}^*$. Ως εκ τούτου, για να είμαστε συμβατοί με το σφάλμα ομαδοποίησης η συνάρτηση κόστους θα πρέπει και αυτή να ορίζεται στο χώρο των εναλλακτικών ομαδοποιήσεων $\mathcal{X}$ και όχι στον χώρο της τάξης του μοντέλου.

Ομαδοποιήσεις και τάξη μοντέλου

Ωστόσο, η εκτίμηση του K μέσω της ομαδοποίησης μέγιστης εκ των υστέρων πιθανότητας συνιστά μια έμμεση διαδικασία συμπερασματολογίας. Η ορθή αντιμετώπιση ενός τέτοιου προβλήματος απαιτεί τη μεγιστοποίηση της εκ των υστέρων πιθανότητας του K. Για να επιτευχθεί κάτι τέτοιο απαιτείται η ολοκλήρωση ως προς όλες τις ομαδοποιήσεις που είναι συμβατές με τη συγκεκριμένη τάξη του μοντέλου $\{\mathbf{x} \in \mathcal{X}^n | \max(\mathbf{x}) = K\}$ (ή ακόμα καλύτερα $\{\mathbf{x} \in \mathcal{X}^n | \max(\mathbf{x}) \leq K\}$, αφού η τυχαία ακολουθία $\mathbf{x} \in \mathcal{X}^n$ μπορεί κάλιστα να μην επισκεφθεί κάθε κατάσταση του μοντέλου, [73]), ορίζοντας μια εκ των προτέρων κατανομή ομαδοποίησης $\mathbf{x}$ δεδομένου του K και μία εκ των προτέρων κατανομής του K, $\pi(\mathbf{x}|K)$ και $\pi(K)$, αντίστοιχα. Η τελευταία αυτή προσέγγιση είναι συμβατή με μια συνάρτηση κόστους με βάση το K και τον πραγματικό αριθμό των ομιλητών K^*, είναι λοιπόν συμβατή με τη κλασσική χρήση του BIC ως μέσου εκτίμησης της τάξης του μοντέλου. Υπό αυτή την έννοια, όλες οι εκδοχές του BIC που εξετάζουμε δεν είναι αυστηρά συμβατές με την αρχική διατύπωση του BIC (βλ. επίσης [74]).

Εκ των προτέρων κατανομή στο χώρο των ακολουθιών καταστάσεων

Για να εξάγουμε τις εκφράσεις των κριτηρίων (87) και (88), ξεκινούμε ορίζοντας την εκ των υστέρων πιθανότητα της ομαδοποίησης $\mathbf{x}$ ως $\pi(\mathbf{x}|\mathbf{y}) \propto p(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})$ και υποθέτουμε ότι η $\pi(\mathbf{x})$ είναι ομοιόμορφη επί του υποσυνόλου $\mathcal{X}_c^n \subset \mathcal{X}^n$ το συνόλου των ομαδοποιήσεων $\mathcal{X}^n$. Στην ομαδοποίηση ομιλητών, το υποσύνολο αυτό αποτελούν όλες εκείνες οι ομαδοποιήσεις $\mathbf{x} \in \mathcal{X}^n$ οι οποίες πληρούν μία σειρά από χρονικούς περιορισμούς. Τυπικά, επιβάλλουμε ως ελάχιστη διάρκεια παραμονής σε κάθε κατάσταση τα 1 με 2 δευτερόλεπτα, και έτσι αποδίδουμε μηδενική μάζα πιθανότητας στις ομαδοποιήσεις $\mathbf{x} \in \mathcal{X}^n$ που περιέχουν έστω και μία ταχεία μετάβαση από ομάδα σε ομάδα. Κατόπιν, ολοκληρώνουμε ως προς τις εξωτερικές παραμέτρους του μοντέλου, υποθέτοντας μια εκ των προτέρων σ.π.π. $\pi(\varphi|\mathbf{x})$ στο Φ^K

$$p(\mathbf{y}|\mathbf{x}) = \int_{\Phi^K} f(\mathbf{y}|\varphi, \mathbf{x})\pi(\varphi|\mathbf{x})d\varphi \tag{95}$$

Πρωτού αναπτύξουμε το πώς το BIC προσεγγίζει την (95), πρέπει να τονίσουμε ότι η διαδικασία αυτή (μεγιστοποίηση της δεσμευμένης ως προς $\mathbf{x}$ ολοκληρωμένης πιθανοφάνειας και χρήση ομοι-

όμορφης κατανομής ως προς $\mathbf{x} \in \mathcal{X}_c^n$) δεν είναι απολύτως Μπεϋσιανή. Δεδομένης μιάς αρχικής κατάτμησης του $\mathbf{y}$, η διαδικασία ισοδυναμεί με μία προσέγγιση ΄τσάντας από τμήματα΄ (bag-of-segments), αφού δεν λαμβάνει υπ΄όψιν τη σχετική θέση των τμημάτων μεταξύ τους. Έτσι, μία προσέγγιση πιο συμβατή με τη Μπεϋσιανή σχολή που οδηγεί στον εμπλουτισμό της συμπερασματολογίας του $\mathbf{x}$ με τη χρονική πληροφορία θα είναι η ολοκλήρωση της πιθανοφάνειας και ως προς τις εξωτερικές παραμέτρους του μοντέλου (δηλαδή ως προς τον πίνακα μεταβάσεων) με χρήση της εκ των προτέρων κατανομής Dirichlet σε κάθε γραμμή του πίνακα μετάβασεων, [73]. Η ιδέα αυτή αναπτύσσεται στο Παράρτημα 1.

5.4.2 Μαθηματική εξαγωγή των εκφράσεων των κριτηρίων

Σαν κριτήριο, το BIC δεν είναι τίποτε άλλο παρά η προσέγγιση Laplace ολοκληρωμάτων και η απόρριψη κάθε όρου δεν αλλάζει τάξη μεγέθους (δηλαδή παραμένει $\mathcal{O}(1)$) καθώς το πλήθος των δειγμάτων αυξάνεται. Για να προσεγγίσουμε λοιπόν το ολοκλήρωμα της (95) χωρίς να ορίσουμε κάποια εκ των προτέρων πιθανότητα επί των φ, χρησιμοποιούμε την προσέγγιση Laplace, [75], [76]. Η ιδέα είναι να προσεγγίσουμε την από κοινού πιθανότητα $f(\mathbf{y}_k|\varphi)\pi(\varphi_k)$ με μία κανονική κατανομή γύρω από την εκτίμηση MAP $\tilde{\varphi}$. Σημειώνουμε ότι ασυμπτωτικά, η κανονικότητα της κατανομής επαληθεύεται από το θεώρημα κεντρικού ορίου. Για επαρκώς μεγάλο n έχουμε ότι $\tilde{\varphi} \to \hat{\varphi}$ και έτσι καταλήγουμε στην παρακάτω προσέγγιση

$$p(\mathbf{y}|\mathbf{x}) \approx (2\pi)^{KP/2} |\mathcal{I}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})|^{-1/2} \exp\left(l(\hat{\varphi}; \mathbf{x}|\mathbf{y})\right) \pi(\hat{\varphi}) \tag{96}$$

όπου με $\mathcal{I}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})$ ορίζουμε την παρατηρούμενη πληροφορία για την φ που φέρουν τα $\mathbf{y}$, υπολογισμένη στην τιμή $\varphi = \hat{\varphi}$, και δεσμευμένη ως προς την ομαδοποίηση $\mathbf{x}$.
Δεσμεύοντας την πιθανοφάνεια ως προς $\mathbf{x}$, είναι εμφανές ότι η παρατηρούμενη πληροφορία $\mathcal{I}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})$ αποκτά δομή μπλοκ-διαγωνίου πίνακα, δηλαδή

$$\mathcal{I}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\varphi) = \mathcal{I}_{\varphi_1}^{\mathbf{y}_1}(\varphi) \oplus \mathcal{I}_{\varphi_2}^{\mathbf{y}_2}(\varphi) \oplus \ldots \oplus \mathcal{I}_{\varphi_K}^{\mathbf{y}_K}(\varphi) \tag{97}$$

αφού μόνο οι n_k παρατηρήσεις $\mathbf{y}_k = \{\mathbf{y}^{(i)} : x^{(i)} = k\}$ φέρουν πληροφορία για την φ_k. Το k μπλοκ ορίζεται ως ακολούθως

$$\mathcal{I}_{\varphi_k}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi}) = -\sum_{i=1}^{n} \delta(x^{(i)}, k) \nabla_{\varphi_k}^2 \log f(\mathbf{y}^{(i)}|\varphi_k)\Big|_{\varphi_k=\hat{\varphi}_k} \tag{98}$$

όπου $(\nabla_\theta^2)_{i.j} = \frac{\partial^2}{\partial\theta_i \partial\theta_j}$. Η προσέγγιση του ολικού-BIC είναι η παρακάτω

$$|\mathcal{I}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})|^{-1/2} = n^{-KP/2} |\mathcal{J}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})|^{-1/2} \tag{99}$$

όπου

$$\mathcal{J}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\varphi) = \mathcal{J}_{\varphi_1}^{\mathbf{y}|\mathbf{x}}(\varphi) \oplus \mathcal{J}_{\varphi_2}^{\mathbf{y}|\mathbf{x}}(\varphi) \oplus \ldots \oplus \mathcal{J}_{\varphi_K}^{\mathbf{y}|\mathbf{x}}(\varphi) \tag{100}$$

και το k μπλοκ έχει ως εξής

$$\mathcal{J}_{\varphi_k}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi}) = -\frac{1}{n}\sum_{i=1}^{n} \delta(x^{(i)}, k) \nabla_{\varphi_k}^2 \log f(\mathbf{y}^{(i)}|\varphi_k)\Big|_{\varphi_k=\hat{\varphi}_k} \tag{101}$$

Αντίθετα, η προτεινόμενη μέθοδος είναι η παραγοντοποίηση της παρατηρούμενης πληροφορίας ως

$$|\mathcal{I}_{\varphi}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})|^{-1/2} = \prod_{k=1}^{K} n_k^{-P/2} |\mathcal{J}_{\varphi_k}^{\mathbf{y}_k}(\hat{\varphi}_k)|^{-1/2} \tag{102}$$

όπου

$$\mathcal{J}_{\varphi_k}^{\mathbf{y}_k}(\hat{\varphi}_k) = -\frac{1}{n_k}\sum_{i=1}^{n} \delta(x^{(i)}, k) \nabla_{\varphi_k}^2 \log f(\mathbf{y}^{(i)}|\varphi_k)\Big|_{\varphi_k=\hat{\varphi}_k} \tag{103}$$

δηλαδή $\mathcal{J}_{\varphi_k}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi}) = \frac{n_k}{n}\mathcal{J}_{\varphi_k}^{\mathbf{y}_k}(\hat{\varphi}_k)$. Οι εκφράσεις των ανωτέρω κριτηρίων μπορούν πλέον να εξαχθούν με αντικατάσταση στην (96) των εξισώσεων (99) και (102) αντίστοιχα και χρησιμοποιώντας την (97). Καταλήγουμε στις προσεγγίσεις $\log p(\mathbf{y}|\mathbf{x}) \approx BIC^G + T^G$ και $\log p(\mathbf{y}|\mathbf{x}) \approx BIC^S + T^S$ όπου T^G και T^S οι σταθεροί όροι που προκύπτουν από τον μη-ορισμό εκ των προτέρων κατανομών.

5.4.3 Οι υπονοούμενες εκ των προτέρων κατανομές

Εξετάζοντας τους όρους αυτούς, οι υπονοούμενες εκ των προτέρων κατανομές εξάγονται εύκολα. Αρχικά, θεωρούμε $\lambda = 1$ και καταλήγουμε ότι

$$T^G = \frac{KP}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K} \log |\mathcal{J}_{\varphi_k}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})| + \log \pi(\hat{\varphi}|\mathbf{x}) \tag{104}$$

και

$$T^S = \frac{KP}{2}\log 2\pi - \frac{1}{2}\sum_{k=1}^{K}\log|\mathcal{J}_{\varphi_k}^{\mathbf{y}_k}(\hat{\varphi}_k)| + \log\pi(\hat{\varphi}|\mathbf{x}) \tag{105}$$

Έτσι, οι υπονοούμενες εκ των προτέρων κατανομές του ολικού-BIC έχουν ως εξής

$$\varphi_k \sim \mathcal{N}(\hat{\varphi}_k, \mathcal{J}_{\varphi_k}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})^{-1}) \tag{106}$$

όπου $\mathcal{N}(m, C)$ η κανονική κατανομή με μέση τιμή m και πίνακα συμμεταβλητότητας C. Οι εκ των προτέρων κατανομές (106) είναι η κατανομές μοναδιαίας πληροφορίας (*unit-information priors*) που προτάθηκαν από τους Kass & Wasserman στο [70] και μπορούν να ορισθούν ως οι κατανομές που προκύπτουν κάνοντας χρήση ενός αντιπροσωπευτικού διανύσματος παρατήρησης. Για να γίνει η ιδιότητα αυτή κατανοητή, σημειώνουμε ότι η $\mathcal{J}_{\varphi_k}^{\mathbf{y}|\mathbf{x}}(\hat{\varphi})$ είναι η αναμενόμενη πληροφορία (δηλαδή ανά παρατήρηση) που φέρει το $\mathbf{y} \in \mathcal{Y}^n$ για την φ_k. Τοποθετώντας την (106) στη (104) ο όρος T^G εξαλείφεται.

Θα πρέπει ωστόσο να τονισθεί ότι οι συγκεκριμένες κατανομές προέρχονται από μοντέλα παλινδρόμησης, όπου η συζυγία μεταξύ πιθανοφάνειας και εκ των προτέρων κατανομής ισχύει. Η χρήση τους ως εκ των προτέρων κατανομών για τις παραμέτρους $\varphi_k = (\mu_k, \Sigma_k)$ είναι προβληματική, καθώς α) παραβιάζουν την ελάχιστη απαιτούμενη ανοχή (χρειάζονται τουλάχιστον $> d-1$ εικονικά δείγματα για την κατανομή του Σ_k) και β) αποδίδουν θετική εκ των προτέρων πυκνότητα σε μη θετικά ορισμένους Σ_k. Μια τυπική Μπεϋσιανή μοντελοποίηση χρησιμοποιεί αντίστροφες Wishart κατανομές Σ_k, με τουλάχιστον d βαθμούς ελευθερίας, [77], [78]. Παρ΄ όλα αυτά, θα συνεχίσουμε την ανάλυσή μας με τις εν λόγω εκ των προτέρων κατανομές, καθώς επιθυμούμε α) να είμαστε συμβατοί με την ανάλυση του BIC και β) να εστιάσουμε στην ανοχή των κατανομών προκειμένου να αναδείξουμε τις διαφορές μεταξύ των κριτηρίων, και όχι στο σχήμα τους.

Με τον ίδιο τρόπο, μπορούμε να εξάγουμε και τις υπονοούμενες κατανομές του τμηματικού-BIC ως εξής,

$$\varphi_k \sim \mathcal{N}(\hat{\varphi}_k, \mathcal{J}_{\varphi_k}^{\mathbf{y}_k}(\hat{\varphi}_k)^{-1}) \tag{107}$$

Η κατανομή (107) δείχνει ότι η βασική διαφοροποίηση έγκειται στο ότι το τμηματικό-BIC αντιστοιχεί στις παραμέτρους $\varphi_k = (\mu_k, \Sigma_k)$ πληροφορία που φέρεται σε μία αντιπροσωπευτική παρατήρηση ανά ομάδα (δηλαδή ανά $\mathbf{y}_k$) σε αντίθεση με το ολικό, το οποίο αντιστοιχεί στο φ μία παρατήρηση

ανά αρχείο (δηλαδή ανά $\mathbf{y}$).

Τέλος, το τοπικό-BIC, δηλαδή ο λογαριθμικός παράγοντας Bayes $\mathrm{BF}_{1,0}$ μεταξύ της εναλλακτικής $\mathcal{H}_1$ (διαφορετικοί ομιλητές) και της μηδενικής υπόθεσης $\mathcal{H}_0$ (ίδιος ομιλητής), χρησιμοποιεί μία παρατήρηση ανά ζεύγος τμημάτων $\mathbf{y}_{k\cup l}$.

Όπως παρατηρούμε από την παραπάνω ανάλυση, το προτεινόμενο κριτήριο παραβιάζει τη γενική στρατηγική της Μπεϋσιανής σύγκρισης μοντέλων, που θέλει την ανοχή μεταξύ των μοντέλων σταθερή, [77]. Θα δείξουμε ότι η παραβίαση αυτής της αρχής μπορεί να οδηγήσει σε κριτήρια που είναι πιο συμβατά με το πρόβλημά μας.

5.4.4 Η παράμετρος ρύθμισης λ ως μέρος της μοντελοποίησης

Όπως έχουμε αναφέρει, ο λόγος που τοποθετούμε την παράμετρο λ είναι η ελλειπώς ορισμένη συνάρτηση εκπομπής (κανονική κατανομή) με την οποία μοντελοποιούμε τις παρατηρήσεις κάθε ομιλητή. Τα πειραματικά αποδεικνύουν ότι θέτοντας $\lambda = 1$ ελάχιστα τμήματα ομιλίας θα συνενωθούν ώστε να δημιουργήσουν ομάδες, [79].

Ένα ερώτημα λοιπόν που τίθεται είναι το κατά πόσον είναι εφικτή η ενσωμάτωση της παραμέτρου λ στη μοντελοποίηση. Αν θεωρήσουμε $\lambda > 1$, μία προφανής και ευρετική μέθοδος είναι η θεώρηση ενός συνόλου παραμέτρων $\bar{\varphi}$ συνολικού αριθμού $|\bar{\varphi}| = (\lambda - 1)KP$ το οποίο δεν έχει καμία συνεισφορά στους λόγους πιθανοφάνειας.

Μία εναλλακτική στρατηγική προκύπτει εύκολα αν θεωρήσουμε εκ των προτέρων κατανομές εξαρτώμενες από το πλήθος των παρατηρήσεων (sample size dependent priors). Η ιδιότητα αυτή, αν και δεν είναι συνήθης στη Μπεϋσιανη σχολή, μας δίνει τη δυνατότητα να εκφράσουμε μαθηματικά τη σ.π.π των κατανομών αυτών. Θεωρώντας $\lambda > 1$ μπορούμε να εκφράσουμε την κατανομή των παραμέτρων της k ομάδας ως

$$\varphi_k \sim \mathcal{N}(\hat{\varphi}_k, n_k^{\lambda-1}\mathcal{J}_{\varphi_k}^{\mathbf{y}_k}(\hat{\varphi}_k)^{-1}) \tag{108}$$

Η εξίσωση (108) δείχνει ότι η λ μπορεί να θεωρηθεί ως υπερπαράμετρος μεταβολής της ανοχής (strength) της εκ των προτέρων κατανομής των φ_k σε σχέση με το πλήθος των παρατηρήσεων n_k.

Πιο συγκεκριμένα, η κατανομή γίνεται όλο και πιο επίπεδη στην παραμετροποίηση $\varphi_k = (\mu_k, \Sigma_k)$ όσο το n_k αυξάνεται.

Αυτό σημαίνει ότι το κεντράρισμα (centering) της εκ των προτέρων κατανομής στην τιμή $\varphi_k = \hat{\varphi}_k$ είναι επίσης προβληματική και ενδεχομένως οδηγεί σε υπερταίριασμα (overfitting) του μοντέλου στα δεδομένα, δηλαδή δημιουργία πολλών μικρών σε αριθμό δειγμάτων ομάδων. Παρατηρούμε όμως το εξής. Το σφάλμα που προκαλεί το κεντράρισμα στην ML εκτίμηση αντί σε μιά άλλη θέση φ_k^0 (η οποία δεν θα εξαρτάται από το συγκεκριμένο δείγμα $\mathbf{y}_k$) θα εξαλείφεται ταχύτατα, λόγω της εξάρτησης της ανοχής της από το n_k. Θέτοντας $\lambda > 1$, η κατανομή της φ_k γίνεται πιο επίπεδη όσο το n_k αυξάνεται. Έτσι, το σφάλμα που εισάγεται λόγω του κεντραρίσματος της εκ των προτέρων κατανομής στην $\hat{\varphi}_k$ (έναντι της όποιας φ_k^0) στην MAP εκτίμηση $\tilde{\varphi}_k$ εξαλείφεται με ρυθμό $r(n_k, \lambda) \propto n_k^{-\lambda}$. Έτσι, μπορούμε να το αγνοήσουμε ακόμα και για μετρίου πλήθους ομάδες και απλώς να θέσουμε $\tilde{\varphi}_k = \hat{\varphi}_k$. Ασυμπτωτικά, ο ρυθμός με τον οποίο η κατανομή γίνεται πιο επίπεδη είναι αυτός που κυριαρχεί - σε συνδυασμό με την πιθανοφάνεια. Στην παράγραφο 5.5.1, η συζήτηση συνεχίζεται με την εξέταση της φύσης των δεδομένων όταν επιδιώκουμε την ομαδοποίηση βάσει των ομιλητών και όχι των φωνημάτων.

5.5 Ομαδοποιήσεις με βάση τους ομιλητές και BIC

Στην παράγραφο αυτή, εξετάζουμε μερικές από τις ιδιότητες του προβλήματος της ομαδοποιήσης με βάση τους ομιλητές σε σχέση με τη χρήση του BIC. Η συζήτηση θα συνεχιστεί και στην παράγραφο 5.6.3, όπου θα εξετάσσουμε την επίπτωση στην ασυμπτωτική συμπεριφορά της στατιστικής logGLR σε πειραματικά δεδομένα.

5.5.1 Ενδεχόμενο υπερταίριασμα των παραμέτρων στα δεδομένα

Ένα ερώτημα που πρέπει να απαντηθεί είναι το κατά πόσον το προτεινόμενο κριτήριο οδηγεί σε υπερταίριασμα των παραμέτρων στα δεδομένα, λόγω του εναλλακτικού τρόπου που υπολογίζει την

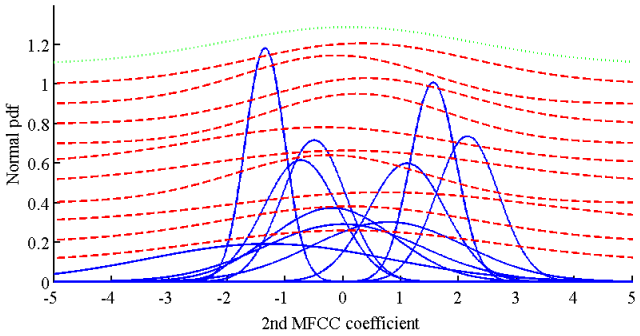
πολυπλοκότητα του μοντέλου. Όπως εξηγήθηκε στην παράγραφο 5.4, η προσέγγισή μας χρησιμοποιεί κατανομές διαφορετικής ανοχής (strength) για κάθε διαφορετική ομαδοποίηση $\mathbf{x} \in \mathcal{X}$ και αντίθετα εστιάζει στο να διατηρεί την αντοχή σταθερή για κάθε ομάδα ξεχωριστά. Από τη στιγμή που οι κατανομές του BIC είναι από τη φύση τους εξαρτώμενες από τα δεδομένα, έχει ενδιαφέρον να δούμε την πληροφορία που φέρουν σε σχέση με τα δεδομένα αυτά.

Για να επιδείξουμε ότι το προτεινόμενο κριτήριο είναι σε ευθυγράμμιση με το πρόβλημά μας θα πρέπει να επισημάνουμε ότι η κατάτμηση των δεδομένων (τυπικά MFCC) με βάση τους ομιλητές δεν αποτελεί τη φυσική ομαδοποίηση των $\mathbf{y} \in \mathcal{Y}^n$. Αντιθέτως, μπορεί να επιτευχθεί μόνο αν θεωρήσουμε κάποιον μηχανισμό που θα επιβάλει την μακρά παραμονή στην παρούσα κατάσταση (bias towards self-transition, [80]), ή - στην περίπτωση των βήμα-προς-βήμα αλγορίθμων - επιβάλοντας περιορισμούς ελάχιστης παραμονής στην παρούσα κατάστατη, διαδικασία που λαμβάνει χώρα κατά την κατάτμημη του αρχείου σε τμήματα. Και στις δύο περιπτώσεις, το αποτέλεσμα είναι οι κανονικές κατανομές να καλύπτουν ευρύτατα τμήματα στον παραμετρικό χώρο, σε σχέση με την αντίστοιχη κάλυψη για το ίδιο $\mathbf{y} \in \mathcal{Y}^n$ όταν οι περιορισμοί αυτοί δεν τίθενται - όταν δηλαδή εφαρμόζουμε π.χ. ιεραρχική ομαδοποίηση απευθείας στις παρατηρήσεις και όχι σε τμήματα ομιλίας. Αυτό με τη σειρά του σημαίνει ότι η πληροφορία η οποία περικλείεται στην εκ των προτέρων κατανομή των φ_k είναι ιδιαίτερα ασαφής (vague), συγκρινόμενη με την αντίστοιχη (φυσική) ομαδοποίηση με βάση τα φωνήματα. Η σοβαρή αυτή διαφοροποίηση μεταξύ των δύο ομαδοποιήσεων σκιαγραφείται στο Σχ. 10.

5.5.2 Γιατί επιθυμούμε να διατηρήσουμε αναλλοίωτες τις ανά δύο αποστάσεις

Όπως αναφέραμε παραπάνω, η πρότασή μας (τμηματικό BIC) προήλθε από την ιδέα να ορίσουμε ένα κριτήριο ικανό να αποδίδει μία τιμή σε συνολικές ομαδοποιήσεις (όπως η ολική του εκδοχή), ενώ ταυτόχρονα, στην ΔBIC μορφή του, να λειτουργεί ως μία αυτόνομη άπόσταση΄ (σαν την τοπική εκδοχή του).

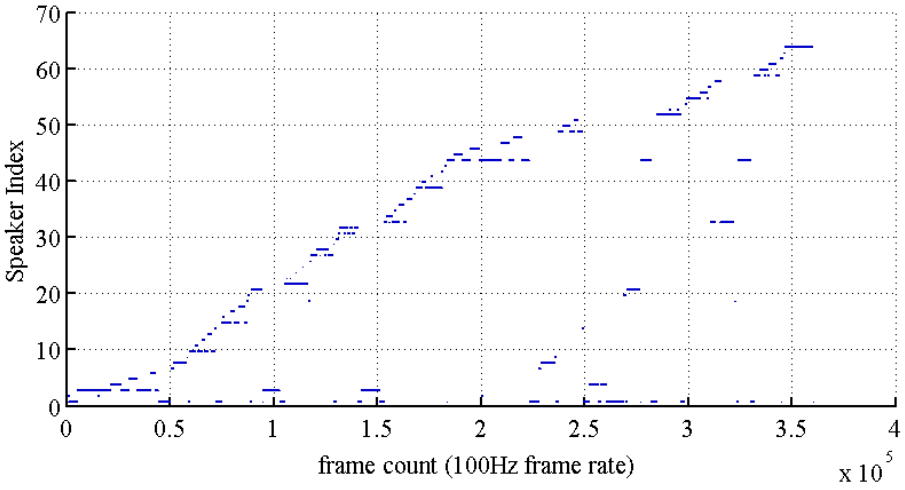
Γιατί λοιπόν η τελευταία ιδιότητα είναι επιθυμητή; Ένα διαισθητικό παράδειγμα είναι το ακόλουθο.



Σχῆμα 10: *Οι διαφορές μεταξύ ομαδοποίησης βάσει ομιλητών έναντι της φυσικής ομαδοποίησης. Οι καμπύλες αντιστοιχούν στις περιθώριες κατανομές των δεύτερων συντελεστών MFCC. Έχουμε 10 γυναικίες φωνές σε συνθήκες στούντιο. Μπλέ - κανονικές καμπύλες: φυσική ομαδοποίηση (μέσω αλγορίθμου ΕΜ), Κόκκινες καμπύλες με παύλες: Ομαδοποίηση με βάση τους ομιλητές, Πράσινες καμπύλες με τελείες (πάνω μέρος): Προσέγγιση της σ.π.π. του συνολικού αρχείου με χρήση μίας μονο κανονικής κατανομής. Για λόγους οπτικοποίησης, κάθε καμπύλη ομιλητή κεντράρεται σε διαφορετική y-συντεταγμένη.*

Υποθέστε ότι επιθυμούμε να ομαδοποιήσουμε τμήματα από μια εκπομπή ειδησεογραφικού χαρακτήρα διάρκειας 60 λεπτών, και έστω S_k και S_l δύο διακριτοί ομιλητές. Έστω επιπλέον ότι οι δύο ομιλητές, μιλούν στην ίδια ιστορία (π.χ. ρεπορτάζ) του δελτίου και επομένως οι ακουστικές συνθήκες είναι παρόμοιες. Ας επιχειρήσουμε δύο πειράματα, ένα χρησιμοποιώντας τα τμήματα ομιλίας του συνολικού δελτίου και το δεύτερο χρησιμοποιώντας τα τμήματα ομιλίας της συγκεκριμένης ιστορίας. Είναι φανερό ότι το ολικό $\Delta\text{BIC}^G > 0$ μεταξύ των δύο ομάδων που αντιστοιχούν στους δύο ομιλητές θα είναι διαφορετικό. Καθώς $n \to \infty$, οποιοδήποτε ομιλητές εμφανίζονται μόνο μια φορά θα ενωθούν, καθώς ο όρος ποινής απειρίζεται. Αυτό έρχεται σε αντίφαση με τη φύση των ομάδων σε εκπομπές ειδησεογραφικού χαρακτήρα. Σε αντίθεση με τα φωνήματα, οι περισσότεροι ομιλητές παρουσιάσουν ελάχιστη διάχυση στον χρόνο του δελτίου, καθώς εμφανίζονται σε μία μόνο ιστορία. Για να επιτευχθεί αυτή η αυτόνομη απόσταση, η προσέγγιση του τοπικού-BIC είναι να εστιάζουμε μόνο στο ζεύγος που συγκρίνουμε, αγνοώντας το υπόλοιπο τμήμα του δελτίου. Αντίθετα, η πρότασή μας είναι ο επαναπροσδιορισμός των εκ των προτέρων κατανομών ώστε να οδηγούν σε μία αυτόνομη ΔBIC απόσταση.

Για να ανακεφαλαιώσουμε, η βασική αρχή του ολικού-BIC είναι η θεώρηση εκ των προτέρων κατανομών σταθερής ανοχής (δηλαδή συνολικών εικονικών παρατηρήσεων) ανά αρχείο ίδιου μήκους (ή για αρχείο γενικότερα αν $\lambda = 1$), ενώ η βασική αρχή του τμηματικού-BIC είναι η θεώρηση εκ των προτέρων κατανομών σταθερής ανοχής ανά τμήμα ίδιου μήκους (ή για τμήμα γενικότερα αν $\lambda = 1$). Το ολικό-BIC, υιοθετώντας την παραπάνω στρατηγική τείνει να υποβαθμίζει τις διαφορές μεταξύ διακριτών ομιλητών σταθερής διάρκειας ομιλίας καθώς η διάρκεια του συνολικού αρχείου αυξάνεται, καθώς προτείνει ποινή ίση με $\lambda P \log n$ για κάθε ομιλητή. Δείχνει λοιπόν πιο συμβατό με προβλήματα όπου ο αριθμός των ομάδων K αυξάνεται πολύ αργά σε σχέση με τον αριθμό των παρατηρήσεων n. Τέτοια παραδείγματα είναι διάλογοι με μικρό αριθμό ομιλητών ή το πρόβλημα της ομαδοποίησης με βάση τα φωνήματα. Αντίθετα, το προτεινόμενο δείχνει ικανό να αντιμετωπίσει και διαδικασίες με ρυθμούς δημιουργίας νέων ομάδων από $K = \mathcal{O}(n)$ έως $K = o(n)$. Ένα τυπικό παράδειγμα ειδησεογραφικού δελτίου παρουσιάζεται στο Σχ. 11. Παρατηρήστε τον σχεδόν γραμμικό ρυθμό δημιουργίας νέων ομάδων και το βαθμό της χρονικής τους εστίασης.



Σχῆμα 11: *Πραγματική ομαδοποίηση ειδησεογραφικού δελτίου από το Γαλλικό Ραδιόφωνο (βάση ESTER). Ο ρυθμός δημιουργίας νέων ομάδων μπορεί να χαρακτηρισθεί ως* $K = \mathcal{O}(n)$.

5.6 Το Τμηματικό BIC τετραγωνικής ρίζας

5.6.1 Φωλιασμένα μοντέλα, ελλειπώς ορισμένα μοντέλα και συνέπεια

Στην παράγραφο αυτή, αναθεωρούμε το παραπάνω κριτήριο. Όπως θα δείξουμε στη συνέχεια, η νέα έκφραση του κριτηρίου

$$BIC_{SR}^{S} = l(\hat{\varphi}; \mathbf{x}|\mathbf{y}) - \lambda \frac{P}{2} \sum_{k=1}^{K} \sqrt{n_k} \log n_k \tag{109}$$

θα αυξήσει κατακόρυφα την ακρίβεια της ομαδοποίησης. Η ερμηνεία αυτού του πιο αυστηρού όρου ποινής βασίζεται σε ένα αποτέλεσμα που προέρχεται από οικονομετρικά μοντέλα, όταν τα συγκρινόμενα μοντέλα δεν είναι φωλιασμένα (nested). Σε τέτοιες περιπτώσεις, όπως αποδείχθηκε στο [81], η συνέπεια του BIC στην επιλογή του κατάλληλου μοντέλου δεν ισχύει, καθώς υπό την μηδενική υπόθεση (δηλαδή όταν δύο μοντέλα επιδεικνύουν ασυμπτωτικά ισάξια ικανότητα περιγραφής των δεδομένων) ένας επιπλέον περιορισμός θα πρέπει να τεθεί για τον όρο ποινής.

Στην περίπτωσή μας, το μείζον πρόβλημα είναι τα ελλειπώς ορισμένα μοντέλα με τα οποία επιχειρούμε να περιγράψουμε την κατανομή των παρατηρήσεων ανά ομιλητή. Ας επαναλάβουμε τις υποθέσεις μας. Η υπόθεση ότι τα δείγματα κάθε ομιλητή είναι ανεξάρτητα και όμοια κατανεμημένα (i.i.d.) είναι απλοϊκή καθώς αγνοούμε τις μαρκοβιανές ιδιότητες των σημάτων φωνής. Επιπλέον, η μονοτροπικότητα της κατανομής των δειγμάτων είναι ξεκάθαρα λαθεμένη υπόθεση. Η κατανομή είναι ξεκάθαρα πολυτροπική, και επομένως ένα μίγμα κανονικών κατανομών (ή μη) είναι πιο ορθή μοντελοποίηση, ή ακόμα καλύτερα ένα μίγμα κατανομών διαδικασίας Dirichlet, αφού η τάξη του μοντέλου είναι και αυτή μια μεταβλητή, [80]. Τέλος, τα τμήματα ομιλίας του ίδιου ομιλητή παρουσιάζουν μεγάλη μεταβλητότητα σε σχέση με την προσωδία, τον ρυθμό παραγωγής των φωνημάτων και άλλων χαρακτηριστικών. Η βασικότερη συνέπεια της χρήσης ενός τόσο απλοϊκού μοντέλου είναι η αργή σύγκλιση του $\hat{\varphi}_k$ στην $\hat{\varphi}_k^*$, αφού η ανθρώπινη ομιλία δεν μπορεί να παραμετροποιηθεί με ένα τέτοιο μοντέλο $\hat{\varphi}_k^*$.

Μη φωλιασμένα μοντέλα και συνέπεια στην επιλογή μοντέλου Το βασικό αποτέλεσμα της ανάλυσης των Sin & White στο [81] έχει ως εξής. Έστω δύο υποψήφια μοντέλα και έστω ότι ισχύει η εναλλακτική υπόθεση $\mathcal{H}_1$. Για ασθενή συνέπεια στην επιλογή του πιο σύνθετου μοντέλου, απαιτείται $c_{\tilde{n}} = o(\tilde{n})$, δηλαδή ο όρος ποινής $c_{\tilde{n}}$ να αυξάνεται με ρυθμό μικρότερο από γραμμικό. Στην περίπτωσή μας, η ακολουθία $c_{\tilde{n}}$ αντιστοιχεί στον όρο ποινής του ΔBIC και ο δείκτης $\tilde{n}$ στο άθροισμα των δειγμάτων, δηλαδή $\tilde{n} = n_k + n_l$. Το αποτέλεσμα αυτό ισχύει ανεξάρτητα από το αν τα μοντέλα είναι φωλιασμένα ή μη, και βασίζεται στο ότι αν η $\mathcal{H}_1$ ισχύει, η διαφορά μεταξύ των λογαριθμικών πιθανοφανειών αυξάνεται γραμμικά με το $\tilde{n}$. Έστω τώρα ότι ισχύει η $\mathcal{H}_0$. Αν τα μοντέλα είναι φωλιασμένα, το κριτήριο είναι ασθενώς συνεπές αν $P(c_{\tilde{n}} \to \infty) = 1$, δηλαδή προκρίνει το απλούστερο μοντέλο με πιθανότητα που τείνει στη μονάδα. Έτσι για φωλιασμένα μοντέλα, οι συνθήκες συνέπειας είναι $c_{\tilde{n}} \to \infty$ και $c_{\tilde{n}} = o(\tilde{n})$. Αντιθέτως, αν τα μοντέλα δεν είναι φωλιασμένα, για ασθενή συνέπεια απαιτείται επιπλέον $P(\tilde{n}^{-1/2} c_{\tilde{n}} \to \infty) = 1$, δηλαδή ο όρος ποινής θα πρέπει να αυξάνεται πιο γρήγορα από $\sqrt{\tilde{n}}$.

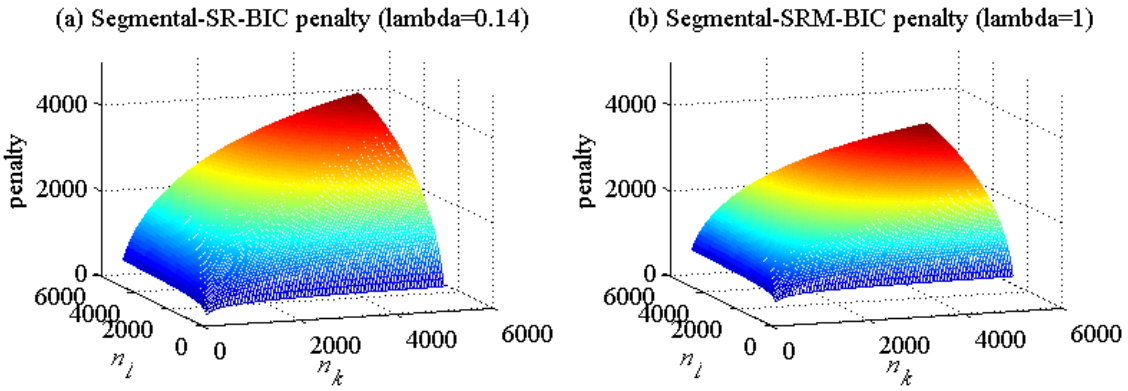
5.6.2 Το τμηματικό BIC τετραγωνικής ρίζας και οι εκ των προτέρων κατανομές του

Ας επιχειρήσουμε να εφαρμόσουμε αυτούς τους περιορισμούς στο κριτήριό μας. Για να διατηρήσουμε την κεντρική ιδέα, δηλαδή ένα κριτήριο το οποίο θα αποδίδει ένα σκορ για ολοκληρωμένες ομαδοποιήσεις των $\mathbf{y}$ ενώ ταυτόχρονα θα οδηγεί σε ένα αυτόνομο ΔBIC, λειτουργούμε ως εξής. Πολλαπλασιάζουμε τους K επιμέρους όρους ποινής με το αντίστοιχο ${n_k}^{1/2}$ και καταλήγουμε στο (109), το οποίο και αποκαλούμε τμηματικό BIC τετραγωνικής ρίζας. Ο όρος ποινής του $\Delta\text{BIC}_{SR}^{S}$ σκιαγραφείται στο Σχ. 12, σαν συνάρτηση των n_k και n_l.

Οι εκ των προτέρων κατανομές του κριτηρίου έχουν ως εξής

$$\varphi_k \sim \mathcal{N}\left(\hat{\varphi}_k, n_k^{\lambda\sqrt{n_k}-1} \mathcal{J}_{\varphi_k}^{-1}(\hat{\varphi}_k)\right) \tag{110}$$

και σε όρους κατανομών μοναδιαίας πληροφορίας, η κατανομή φ_k χρησιμοποιεί μόλις το $n_k^{1-\lambda\sqrt{n_k}}$ μίας παρατήρησης. Αφού η μεταβλητότητα της κατανομής αυξάνεται τόσο γρήγορα με το n_k, το



Σχῆμα 12: *Προτεινόμενοι όροι ποινής για το* ΔBIC, *ως συνάρτηση του αριθμού παρατηρήσεων* $50 \leq n_k, n_l \leq 6000$. (α) τμηματικό BIC τ. ρίζας ($\lambda = 0.14$), (β) τμηματικό BIC τ. ρίζας, μέσων τιμών ($\lambda = 1$).

κεντράρισμα της κατανομής στην τιμή $\hat{\varphi}_k$ ή σε κάποια προκαθοριμένη τιμή φ_k^0 καθίσταται επουσιώδες. Αυτό που κυριαρχεί είναι ο ρυθμός με τον οποίον η μεταβλητότητά του αυξάνεται (δείτε παράγραφο 5.4.4).

Παρατηρήστε ότι οι παραπάνω περιορισμοί εχουν τεθεί μόνο στο ΔBIC, δηλαδή όταν εκτιμούμε αν δύο ομάδες προέρχονται από τον ίδιο ή διαφορετικούς ομιλητές. Για μοντέλα με περισσότερους ομιλητές, οι περιορισμοί ισχύουν μόνο για την αναμενόμενη τιμή των όρων ποινής . Επιτρέπουμε να συμβεί κάτι τέτοιο αφού όπως αναλύσαμε στην παράγραφο 5.4.1, τα κριτήρια αποδίδουν σκορ σε ομαδοποιήσεις $\mathbf{x} \in \mathcal{X}_c^n$ οι οποίες υπονοούν τάξη μοντέλου και όχι απευθείας στην τάξη του μοντέλου.

Τέλος, παρατηρήσθε ότι για να είμαστε συμβατοί με την ανάλυση της παραγράφου 5.4.3 και την ερμηνεία που απόδίδουμε τόσο για το λ όσο και για την τ. ρίζα (μέσω των εκ των προτέρων κατανομών) θα πρέπει να εφαρμόσουμε τις παραπάνω ιδέες μόνο στις μέσες τιμές και όχι στους πίνακες συμμεταβλητότητας. Διαφορετικά, η εκ των προτέρων κατανομή της παραμέτρου Σ_k θα καθίστατο μη-ολοκληρώσιμη καθώς το n_k αυξάνεται. Το κριτήριο αυτό έχει ως εξής

$$BIC_{SRM}^S = l(\hat{\varphi}; \mathbf{x}|\mathbf{y}) - \sum_{k=1}^{K} \frac{\lambda P_\mu}{2} \sqrt{n_k} \log n_k + \frac{P_\Sigma}{2} \log n_k \tag{111}$$

όπου $P_\mu = d$ και $P_\Sigma = d(d+1)/2$. Η έκδοση αυτή έχει ένα επιπλεόν πλεονέκτημα σε σχέση με την (109). Η βέλτιστη επίδοση στο πρόβλημά μας επιτυγχάνεται (κατόπιν πειραμάτων) για $\lambda = 1$.[1]

5.6.3 Η επίπτωση της χρήσης ελλειπώς ορισμένων μοντέλων

Η έως τώρα ανάλυση βασίστηκε σε μια πιο θεωρητική ανάλυση των διαφορών μεταξύ του ολικού και του τμηματικού κριτηρίου. Στην παράγραφο αυτήν παρουσιάζουμε τις διαφορές του σε σχέση με το πρόβλημα της χρήσης ελλειπώς ορισμένων μοντέλων, χρησιμοποιώντας τόσο πραγματικά δείγματα φωνής, όσο και συνθετικά. Μια βασική ερώτηση η οποία χρήζει απάντησης είναι αν η εξάρτηση του όρου ποινής από τα $\{n_k\}_{k=1}^K$ είναι απλώς ένα παράγωγο της απόπειρας να παντρέψουμε ορισμένες από τις ιδιότητες του ολικού και του τμηματικού κριτηρίου. Θα πρέπει να παρατηρήσουμε ότι για σταθερά K και n, ο όρος ποινής μεγιστοποιείται όταν $n_k = n/K, k = 1, \ldots, K$. Αυτό αποδεικνύεται εύκολα, λόγω του ότι η συνάρτηση $f(x) = \sqrt{x}\log x$ είναι κοίλη. Θα πρέπει να δείξουμε λοιπόν γιατί η ομοιόμορφη ομαδοποίηση $n_k = n/K, k = 1, \ldots, K$ πρέπει να έχει τη μέγιστη ποινή·
Για να αιτιολογήσουμε την προσέγγιση μας και επιπλέον να είμαστε σε θέση να τη συγκρίνουμε με το τοπικό κριτήριο, εστιάζουμε στο ΔBIC και εξετάζουμε τη συμπεριφορά της στατιστικής logGLR. Χρησιμοποιούμε λοιπόν δύο ηχογραφήσεις από το Ελληνικό κοινοβούλιο, και συγκεκριμένα ομιλίες του νυν και του πρώην πρωθυπουργού. Για να παρατηρήσουμε την στατιστική logGLR όταν η $\mathcal{H}_1$ ισχύει, τμήματα ομιλίας εξάγονται τυχαία από τα αρχεία ήχου. Η συμπεριφορά της στατιστικής μας σκιαγραφείται στο Σχ. 13(a), συναρτήσει του αθροιστικού αριθμού των δειγμάτων και των αναλογιών. Ο x-άξονας ορίζει τον αθροιστικό αριθμό των δειγμάτων $\tilde{n} = n_k + n_l$, ενώ το χρώμα και το στύλ γραμμής την αναλογία $w_k = \frac{n_k}{\tilde{n}}, w_l = 1 - w_k$, όπου χωρίς βλάβη της γενικότητας θεωρούμε $w_k \leq 0.5$. Η ίδια διαδικασία επαναλαμβάνεται όταν η $\mathcal{H}_0$ ισχύει. Στην περίπτωση αυτή, χρησιμοποιούμε μόνο τη μία ηχογράφηση με προσοχή ώστε τα δύο τμήματα ομιλίας που

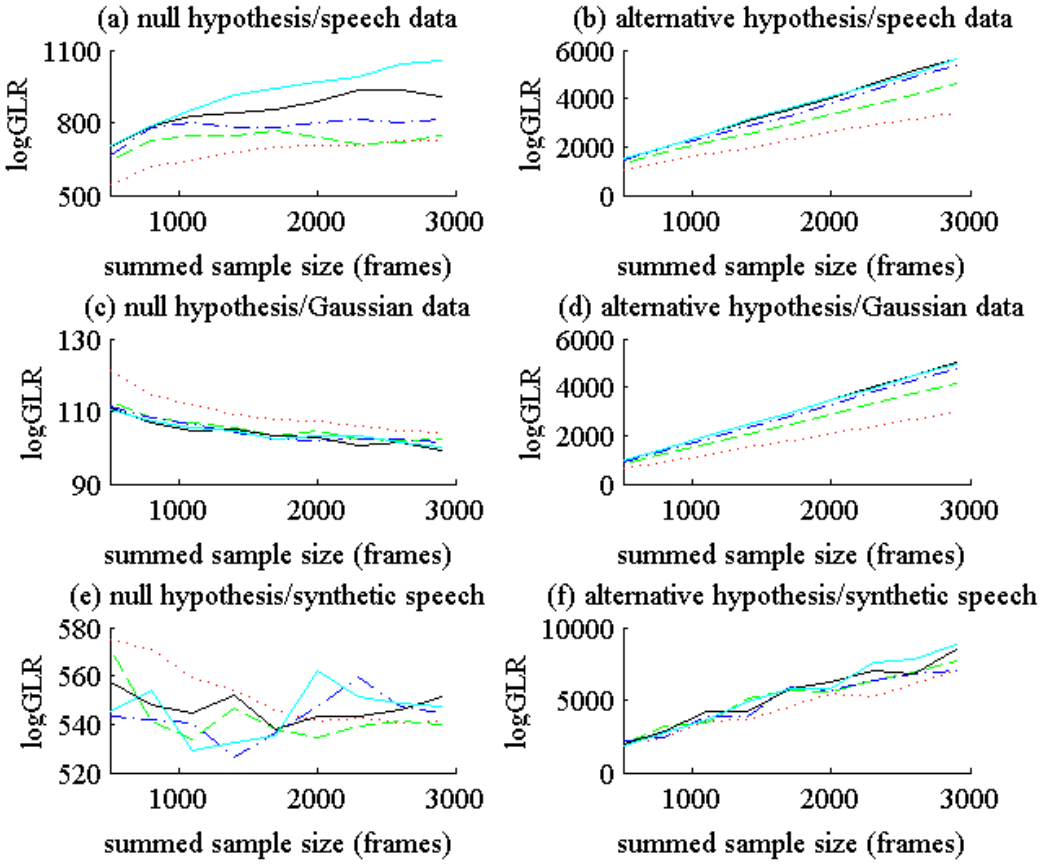
[1] Αν επιθυμούμε την περίληψη και διαφορικών όρων MFCC πέραν των στατικών, οι μέσες τιμές τους δεν πρέπει να θεωρηθούν ελεύθερες παράμετροι, δηλαδή. P_μ αντιστοιχοί μόνο στο πλήθος των στατικών όρων. Αντιθέτως, οι μέσες τιμές θα πρέπει να τίθενται ίσες με τις αναμενώμενες τιμές τους (δηλαδή μηδέν) και οι πίνακες συμμεταβλητότητας να υπολογίζονται με τη φόρμουλα γνωστής μέσης τιμής, (δηλαδή σ_{n-1} έναντι της σ_n)

εξετάζονται κάθε φορά να μην παρουσιάζουν επικάλυψη μεταξύ τους (Σχ. 13b). Για την οπτικοποίηση του τρόπου με τον οποίον η παραδοχή της κανονικότητας επηρεάζει τη στατιστική μας, το ίδιο πείραμα επαναλαμβάνεται με συνθετικά δεδομένα. Δημιουργούμε δύο πίνακες παρατήρησεις μηδενικής μέσης τιμής και μοναδιαίου πίνακα συμμεταβλητότητας και ο κάθε πίνακας μετασχηματίζεται γραμμικά (affine transform) οι δύο πρώτες ροπές να ταυτισθούν με τις αντίστοιχες των πραγματικών ηχογραφήσεων. Η συμπεριφορά της στατιστικής μας για την περίπτωση των Γκαουσιανών δεδομένων σκιαγραφείται στο Σχ. 13(c-d). Τέλος, συμπεριλαμβάνουμε και ένα πείραμα με συνθετική φωνή, η οποία δημιουργήθηκε χρησιμοποιώντας συνθέτη συγκολλητικής αρχιτεκτονικής (concatenative synthesis). Σημειώνουμε ότι ο συνθέτης έχει ρυθμισθεί με τρόπο ώστε το αποτέλεσμα να είναι αρκετά επίπεδο σε σχέση με την προσωδία και τον ρυθμό εκπομπής φωνημάτων. Οι αντίστοιχες καμπύλες παρουσιάζονται στο Σχ. 13 (e-f).

Το Σχ. 13 σκιαγραφεί τη δραστική επίπτωση της χρήσης ελλειπώς ορισμένων μοντέλων στη στατιστική μας όταν ισχύει η $\mathcal{H}_0$. Από τις καμπύλες προκύπτει ότι η στατιστική μας δεν μπορεί να χαρακτηρισθεί ως $\mathcal{O}(1)$, δηλαδή παρατηρούμε επίδραση παρόμοια με αυτήν που παρατηρείται όταν συγκρίνονται μη φωλιασμένα μοντέλα. Παρατηρούμε ότι η στατιστική μας αυξάνεται με το $\tilde{n}$ με έναν σχεδόν γραμμικό ρυθμό για το εύρος των διαρκειών που εξετάζουμε, σε πλήρη αντίθεση με την συμπεριφορά της για Γκαουσιανά δεδομένα, όπου παρατηρείται μείωσή της με το $\tilde{n}$ και επομένως μπορεί να χαρακτηρισθεί ως $\mathcal{O}(1)$. Επομένως, ένας όρος ποινής με συμπεριφορά $\mathcal{O}(\log \tilde{n})$ είναι εμφανώς πολύ αργός ώστε να παρακολουθήσει την αύξηση αυτή.

Εστιάζοντας στην επίδραση των διαφορετικών αναλογιών $\{w_k, w_l\}$, μπορούμε επίσης να αιτιολογήσουμε την επιλογή μας για έναν όρο ποινής που αυξάνεται καθώς $w_k \to 0.5$. Παρατηρούμε ότι και για τις δύο υποθέσεις, η κλίση των καμπυλών αυξάνεται καθώς $w_k \in (0, 0.5]$. Επομένως, ο όρος ποινής που προτείνεται από το τοπικό κριτήριο δεν μπορεί να θεωρηθεί ως βέλτιστος. Ο βέλτιστος όρος ποινής πρέπει να εξαρτάται από τα $\{n_k, n_l\}$ και όχι απλώς από το άθροισμά τους. Χρησιμοποιούντας μόνο το $\tilde{n}$, το αποτέλεσμα είναι η εσφαλμένη συνένωση τμημάτων όπου $w_k \ll w_l$ με αυξανόμενη πιθανότητα καθώς $\frac{w_k}{w_l} \to 0$.

Ένα τελευταίο και ιδιαίτερα ενδιαφέρον συμπέρασμα μπορεί να εξαχθεί παρατηρώντας ότι όταν η υπόθεση $\mathcal{H}_0$ είναι ορθή, οι καμπύλες των πραγματικών δεδομένων ομιλίας διαφοροποιούνται ση-



Σχῆμα 13: *Η συμπεριφορά της στατιστικής* $\log GLR$. *Άνω γραμμή: πραγματική ομιλία, μεσαία γραμμή: Γκαουσιανά δεδομένα, κάτω γραμμή: συνθετική ομιλία. Αριστερά στήλη: ισχύει η* $\mathcal{H}_0$, *δεξιά στήλη: ισχύει η* $\mathcal{H}_1$. *Χρώμα και στυλ γραμμής αντιστοιχούν σε διαφορετικές αναλογίες* w_k, w_l: *Κόκκινη γραμμή με τελείες:* $w_k = 0.1$, *Πρασινη γραμμή με παύλες:* $w_k = 0.2$, *Μπλε γραμμή με παύλες και τελείες:* $w_k = 0.3$, *Μαύρη συμπαγής γραμμή:* $w_k = 0.4$ *και Κυανή γραμμή με '×':* $w_k = 0.5$.

μαντικά με αυτές των δεδομένων συνθετικής ομιλίας. Όπως και στην περίπτωση των Γκαουσιανών δεδομένων, οι καμπύλες των δεδομένων συνθετικής ομιλίας μπορούν να χαρακτηρισθούν ως $\mathcal{O}(1)$. Αυτό σημαίνει ότι η πολυτροπικότητα της ανθρώπινης ομιλίας δεν είναι ο βασικότερος παράγων που οδηγεί στην αλλοίωση της συμπεριφοράς της στατιστικής, αφού τα δεδομένα συνθετικής ομιλίας

είναι επίσης πολυτροπικά. Αντίθετα, η έντονη μεταβλητότητα της ανθρώπινης ομιλίας σε σχέση με την προσωδία, τον ρυθμό εκπομπής φωνημάτων και λοιπών χαρακτηριστικών φαίνεται να αποτελεί τον σημαντικότερο παράγοντα αλλοίωσης στη συμπεριφορά της στατιστικής μας.
Για να ανακεφαλαιώσουμε, ένας όρος ποινής που αυξάνεται με $c_{\tilde{n}} = \mathcal{O}(\tilde{n}^{1/2} \log \tilde{n})$ δείχνει να είναι ικανός να αντιμετωπίσει το πρόβλημα της χρήσης ελλειπώς ορισμένων μοντέλων. Σημειώνεται ότι ο ρυθμός συμπίπτει με αυτόν που προτείνεται στο [82], ώστε να αντιμετωπίσει το πρόβλημα της σύγκρισης μεταξύ μη φωλιασμένων μοντέλων. Επιλεόν, η αύξηση του όρου ποινής καθώς $\{w_k\}_{k=1}^{K} \to K^{-1}$ φαίνεται ότι συμβαδίζει με τα παραπάνω γραφήματα, κάτι που γίνεται φανερό για πραγματικά δεδομένα ανθώπινης ομιλίας.

5.7 Πειραματικά αποτελέσματα

5.7.1 Σύνολα αξιολόγησης και μετρικές αξιολόγησης

Για να αξιολογήσουμε την επίδοση των κριτηρίων χρησιμοποιούμε τη βάση ESTER, [83]. Ο αλγόριθμος στον οποίον βασιζόμαστε είναι η συγκολλητική ιεραρχική ομαδοποίηση, [27]. Έχει προηγηθεί η εφαρμογή του αλγορίθμου κατάτμησης και όπως έχουμε αναφέρει, το κατώφλι έχει εκτιμηθεί έτσι ώστε να ελαχιστοποιηθεί η πιθανότητα μη εντοπισμού σημείου αλλαγής ομιλητή. Τα αποτελέσματα του σταδίου κατάτμησης είναι ταυτόσημα για όλα τα κριτήρια. Επιπλέον, τα τμήματα που δεν αντιστοιχούν σε ομιλία έχουν απορριφθεί με τη χρήση ενός GMM - ταξινομητή. Για την εξαγωγή των αποτελεσμάτων χρησιμοποιείται το λογισμικό που παρέχεται από την NIST. Ως παραμετρικό χώρο χρησιμοποιούμε τους 18-διάστατους συντελεστές MFCC επαυξημένους με την λογαριθμική ενέργεια, με χρήση του λογισμικού *spro*-4.0. Η μοντελοποίηση των τμημάτων ομιλίας γίνεται με χρήση της απλής κανονικής κατανομής, με πλήρη πίνακα συμμεταβλητότητας. Η στατιστική logGLR έχει την παρακάτω κλειστή φόρμουλα υπολογισμού

$$\log GLR = \frac{n_k + n_l}{2} \log |\Sigma_{k \cup l}| - \frac{n_k}{2} \log |\Sigma_k| - \frac{n_l}{2} \log |\Sigma_l| \tag{112}$$

όπου $\Sigma_{k\cup l}$ ο πίνακας συμμεταβλητότητας των τμημάτων $\mathbf{y}_k \cup \mathbf{y}_l$. Σε κάθε αναδρομή, οι επαρκείς στατιστικές κάθε νέας ομάδας υπολογίζονται, όπως και οι εκατέρωθεν τιμές των ΔBIC. Ο αλγόριθμος τερματίζεται όταν όλα τα ΔBIC είναι μη-αρνητικά. Η αλγοριθμική υλοποίηση βασίζεται στο λογισμικό ανοικτού κώδικα που παρέχεται από το εργαστήριο LIUM, [79], [27].

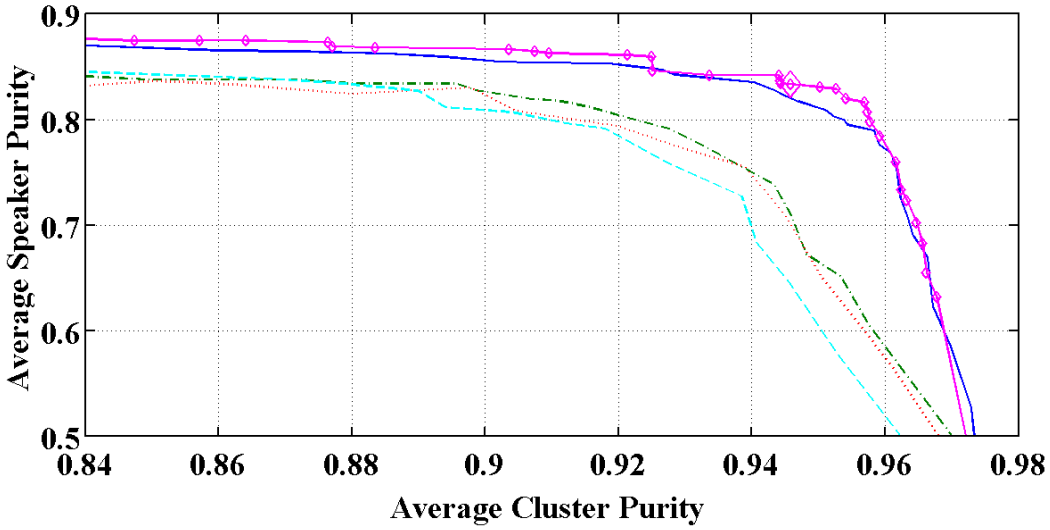
Εκτός του ολικού σφάλματος ομαδοποίησης (DER, %), χρησιμοποιούμε επίσης τα διαγράμματα μέσης καθαρότητας ομάδας (*acp*) και μέσης καθαρότητας ομιλητή (*asp*). Υψηλές τιμές *acp* αντιστοιχούν σε ομαδοποιήσεις στις οποίες κάθε ομάδα περιέχει δείγματα κυρίως από έναν ομιλητή αναφοράς. Δυικά, υψηλές τιμές *asp* αντιστοιχούν σε ομαδοποιήσεις στις οποίες το σύνολο των τμημάτων που ανήκουν σε έναν ομιλητή αναφοράς αποδίδεται σε μία μόνο ομάδα, [30]. Τέλος, όλα τα κριτήρια αξιολογούνται σε πολλαπλά σημεία λειτουργίας, μεταβάλλοντας την παράμετρο ρύθμισης λ.

5.7.2 Πειραματικά αποτελέσματα στη βάση ESTER

Η βάση ESTER αποτελείται από 32 εκπομπές λόγου από το Γαλλικό Ραδιάφωνο. Η βάση είναι χωρισμένη στα σύνολα ανάπτυξης (DEV) και αξιολόγησης (TEST). Το σύνολο ανάπτυξης αποτελείται από 14 εκπομπές, με διάρκειες από 8 εώς 60 λεπτά (7.4 ώρες συνολικά). Οι καμπύλες *acp*-*asp* σκιαγραφούνται στο Σχ. 14, ενώ οι αντίστοιχες για το σύνολο αξιολόγησης (18 εκπομπές, 9.3 ώρες συνολικά) στο Σχ. 15. Η επικράτηση των προτεινόμενων κριτηρίων έναντι των υπαρχόντων είναι εμφανής.

Στο Σχ. 16, παρουσιάζουμε τον συνολικό αριθμό ομιλητών (TNS) και τον σχετίζουμε με το ολικό σφάλμα ομαδοποίησης (DER). Το σχήμα αυτό είναι ιδιάιτερα κρίσιμο, καθώς βάσει αυτού μπορούμε να επιλέξουμε την κατάλληλη τιμή της παραμέτρου ρύθμισης για κάθε εφαρμογή. Παρατηρούμε ότι η τοπική εκδοχή επιτυγχάνει τη βέλτιστη επίδοση σε DER υποεκτιμώντας το TNS. Έτσι, η επίδοση αυτή δεν μπορεί να βελτιωθεί με χρήση περαιτέρω ιεραρχικής συγκέντρωσης χρησιμοποιώντας πιο σύνθετα μοντέλα (GMMs). Αντίθετα, η βέλτιστη επίδοση σε DER του τμηματικού BIC ρίζας επιτυγχάνεται με μιά ελαφρά υπέρβαση του TNS, που είναι αναμενόμενο καθώς πολλά τμήματα

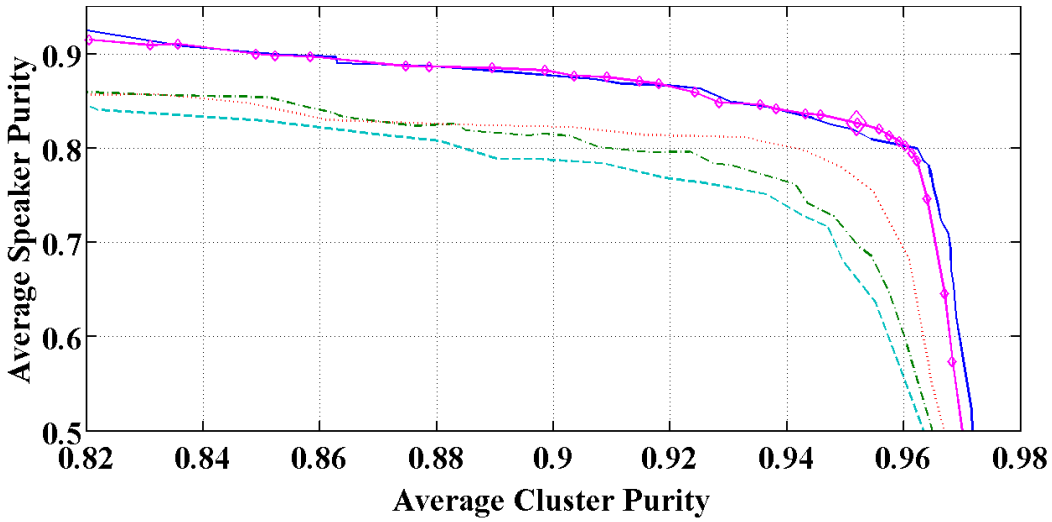
ομιλίας σε ειδησεογραφικές εκπομπές είναι αναμεμειγμένα με μουσική και ήχους περιβάλλοντος. Σημειώνουμε επίσης την πολύ καλή επίδοση του BIC^S_{SRM}, ιδιαίτερα για $\lambda = 1$.



Σχῆμα 14: *acp και asp στο σύνολο ανάπτυξης της βάσης ESTER. Κυανή, διακεκομμένη γραμμή: Ολικό-BIC, Κόκκινη γραμμή με τελείες: Τοπικό-BIC, Πράσινη, διακεκομμένη γραμμή με τελείες: Τμηματικό-BIC, Μπλέ, κανονική γραμμή: Τμηματικό-BIC ρίζας, Μωβ, κανονική γραμμή με ρόμβους: Τμηματικό-BIC ρίζας στις μέσες τιμές (Μεγάλος ρόμβος: $\lambda = 1$).*

Το εύρος τιμών της παραμέτρου λ που εξετάζεται είναι $[0.9, 11.0]$, εκτός από τα BIC^S_{SR} και BIC^S_{SRM}, για το οποία είναι $[0.015, 0.220]$ και $[0.3, 1.8]$, αντίστοιχα. Το BIC^S_{SR} επιτυγχάνει τη βέλτιστη επίδοση για $\lambda = 0.14$. Τα ολικά σφάλματα ομαδοποίησης κάθε κριτηρίου παρατίθενται στον Πίνακα 1. Στη στήλη TEST* παρουσιάζουμε το αποτέλεσμα στο ESTER-TEST με βάση τη βέλτιστη ρύθμιση που προέκυψε από το ESTER-TEST, ενώ στη στήλη TEST την καλύτερη επίδοση στο συγκεκριμένο σύνολο. Τονίζουμε ότι στον Πίνακα 1 η επίδοση του BIC^S_{SRM} επετεύχθη θέτοντας $\lambda = 1$. Σημειώνουμε τέλος ότι επιχειρήσαμε να εφαρμόσουμε την ιδέα της τετραγωνικής ρίζας και στο ολικό και τοπικό BIC (δηλαδή, πολλαπλασιάσαμε τους όρους ποινής με $\sqrt{n}$ και $\sqrt{n_k + n_l}$, αντίστοιχα), αλλά η απόδοσή τους μειώθηκε.

Ως τελευταίο πείραμα, επιχειρήσαμε να αξιολογήσουμε τα κριτήριά μας σε σχέση με το Τοπικό-BIC

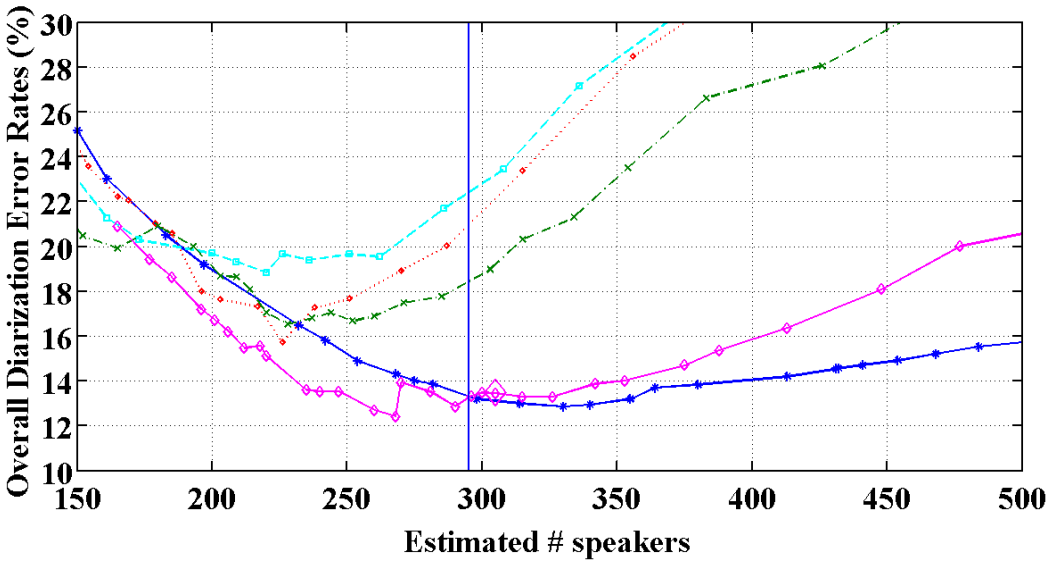


Σχῆμα 15: *acp και asp στο σύνολο αξιολόγησης της βάσης ESTER. Κυανή, διακεκομμένη γραμμή: Ολικό-BIC, Κόκκινη γραμμή με τελείες: Τοπικό-BIC, Πράσινη, διακεκομμένη γραμμή με τελείες: Τμηματικό-BIC, Μπλέ, κανονική γραμμή: Τμηματικό-BIC ρίζας, Μωβ, κανονική γραμμή με ρόμβους: Τμηματικό-BIC ρίζας στις μέσες τιμές (Μεγάλος ρόμβος: $\lambda = 1$).*

ανεξάρτητα από τον αλγόριθμο ιεραρχικής ομαδοποίησης. Ο συγκεκριμένος αλγόριθμος έχει την ικανότητα απόκριψης πολλών εκ των εσφαλμένων αποφάσεων των κριτηρίων, αφού σε κάθε αναδρομή μόνο το ζεύγος που παρουσιάζει το ελάχιστο ΔBIC συνενώνεται. Επομένως, πολλά από τα σφάλματα παραμένουν κρυφά, δηλαδή δεν επηρεάζουν την τελική λύση.

Για να προβούμε σε μια τέτοια εκτίμηση, υπολογίζουμε τις εκατέρωθεν αποστάσεις μεταξύ των αρχικών τμημάτων, βασισμένοι όμως στην πραγματική (ground truth) κατάτμηση, ώστε να αποφύγουμε τα όποια σφάλματα προκαλεί το στάδιο κατάτμησης. Τμήματα διάρκειας μικρότερης των 2 s αποκλείωναι από το πείραμα. Σφάλμα τύπου-I (τύπου-II) παρουσιάζεται όταν ΔBIC>0 (ΔBIC<0) ενώ τα δύο τμήματα ανήκουν στον ίδιο (διαφορετικό) ομιλητή. Ο Πίνακας 2 δείχνει τα ποσοστιαία σφάλματα των κριτηρίων στο σύνολο ανάπτυξης. Οι παράμετροι λ έχουν ως παραπάνω, δηλαδή προέκυψαν βάσει του ελάχιστου DER στο σύνολο ανάπτυξης, ενώ $\lambda = 1$ για το BIC^S_{SRM}.

Ο συνολικός αριθμός των εξεταζομένων ζευγών είναι 74552 και περιλαμβάνονται μόνο ζεύγη που ανήκουν στο ίδιο αρχείο ήχου. Σημειώνουμε ότι τα ζεύγη εκείνα που ανήκουν στον ίδιο ομιλητή

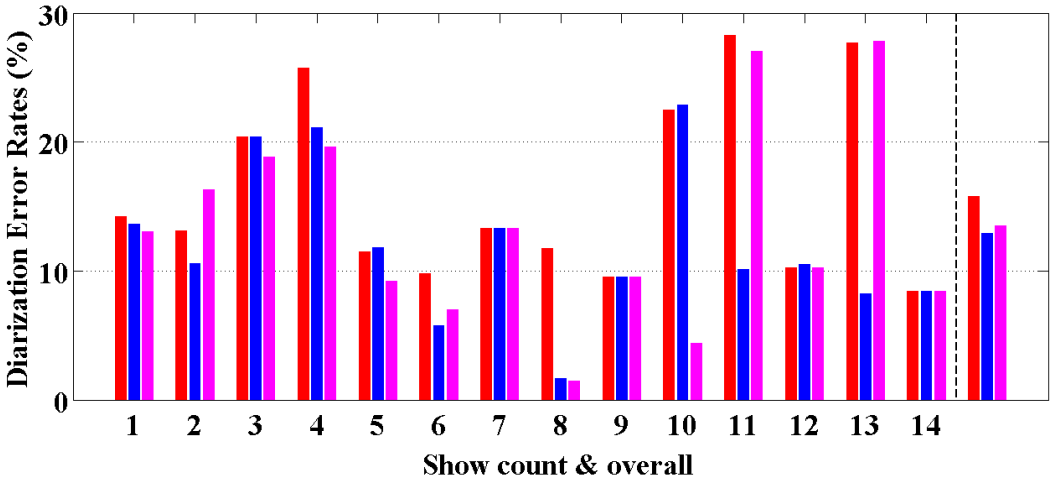


Σχῆμα 16: *Εκτιμώμενος αριθμος ομιλητών και DER (%) στη βάση ESTER (σύνολο ανάπτυξης). Κυανή, διακεκομμένη γραμμή με τετράγωνα: Ολικό-BIC, Κόκκινη γραμμή με τελείες και διαμάντια: Τοπικό-BIC, Πράσινη, διακεκομμένη γραμμή με τελείες και '×': Τμηματικό-BIC, Μπλέ, κανονική γραμμή με άστρα: Τμηματικό-BIC ρίζας, Μωβ, κανονική γραμμή με ρόμβους: Τμηματικό-BIC ρίζας στις μέσες τιμές (Μεγάλος ρόμβος: $\lambda = 1$). Ο πραγματικός αριθμός των ομιλητών σημειώνεται με την κάθετη γραμμή.*

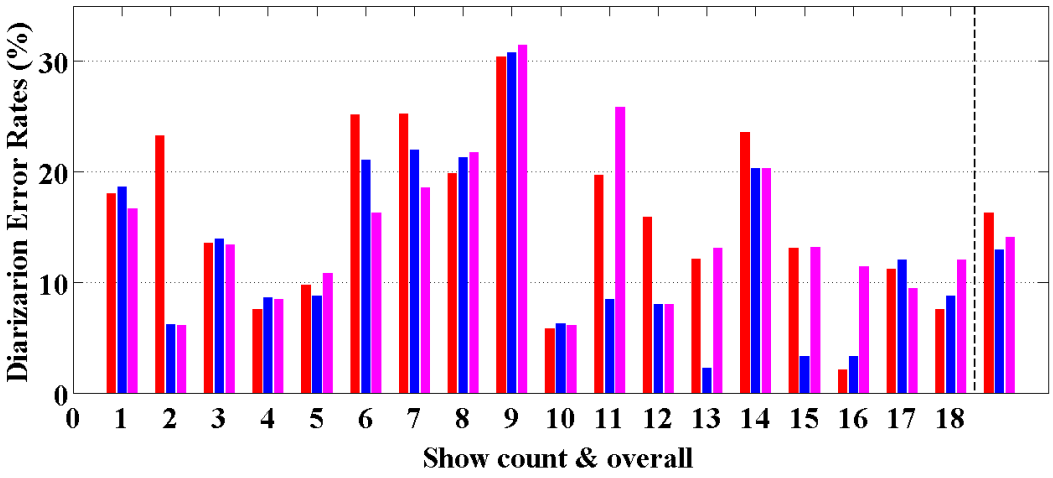
αντιστοιχούν στο 10.32% τους συνόλου των ζευγών, γεγονός που εξηγεί και τη μεγάλη διαφορά στον μέσο ρυθμό κρυφών λαθών (HER) μεταξύ των κριτηρίων. Ο κανονικοποιημένος μέσος ρυθμός HER (μέση τιμή του ρυθμού των δύο τύπων σφάλματος) είναι πιο αντιπροσωπευτικός δείκτης για μια τέτοια αξιολόγηση.

5.8 Επίλογος κεφαλαίου και μελλοντική έρευνα

Στο κεφάλαιο αυτό, παρουσιάσαμε μια νέα μορφή του BIC, κατάλληλη για το πρόβλημα της ομαδοποίησης ομιλητών. Συζητήσαμε αρκετά θέματα, όπως οι υπονοούμενες κατανομές των κριτηρίων καθώς και ο ρόλος της παραμέτρου ρύθμισης ως υπερπαραμέτρου των κατανομών. Κατόπιν, δε-



Σχῆμα 17: *Βέλτιστη επίδοση σε DER (%) για τις 14 εκπομπές της βάσης ESTER (σύνολο ανάπτυξης). Αριστερές, κόκκινες μπάρες: Τοπικό-BIC, Μέσες-μπλέ μπάρες: τμηματικό-BIC ρίζας, Δεξιές μπάρες με μωβ: τμηματικό-BIC ρίζας μέσων τιμών* ($\lambda = 1$)



Σχῆμα 18: *DER (%) για τις 18 εκπομπές της βάσης ESTER (σύνολο αξιολόγησης). Αριστερές, κόκκινες μπάρες: Τοπικό-BIC, Μέσες-μπλέ μπάρες: τμηματικό-BIC ρίζας, Δεξιές μπάρες με μωβ: τμηματικό-BIC ρίζας μέσων τιμών* ($\lambda = 1$). *πέραν του τελευταίου, η ρύθμιση βασίστηκε στο σύνολο ανάπτυξης.*

Πίνακας 1: Ελάχιστο ολικό σφάλμα ομαδοποίησης (%) στα δύο σύνολα της βάσης ESTER

	DEV	TEST	TEST*
Ολικό-BIC	18.84	19.63	21.55
Τοπικό-BIC	15.76	15.86	16.28
Τμηματικό-BIC	16.43	17.12	18.37
Τμηματικό-BIC ρίζας	12.91	12.94	12.94
Τμηματικό-BIC ρίζας (μέσων τιμών)	13.48	14.05	-
Σφάλμα εσφαλμένου εντοπισμού ομιλίας	0.3	0.6	0.6
Σφάλμα μή εντοπισμού ομιλίας	0.9	1.2	1.2

Πίνακας 2: Ρυθμός κρυφών σφαλμάτων (%) στο σύνολο ανάπτυξης της βάσης ESTER

	Τύπος-I	Τύπος-II	Συνολικά (κανον.)	Συνολικά
Τοπικό-BIC	2.00	68.59	35.79	62.62
Τμηματικό-BIC ρίζας	41.11	1.80	21.47	5.89
Τμηματικό-BIC ρίζας (μέσων τιμών)	22.05	6.92	14.49	8.48

ίξαμε πώς οι βασικές ιδιότητες των δύο εκδοχών (ολικής και τοπικής) μπορούν να συγκερασθούν και προτείναμε την τμηματική εκδοχή. Τέλος, βασισμένοι στα προβλήματα συνέπειας που παρουσιάζει η επιλογή μη φωλιασμένων μοντέλων, προτείναμε έναν αυστηρότερο όρο ποινής, ώστε να συνυπολογίσουμε το κόστος της χρήσης απλοϊκών μοντέλων. Τα πειραματικά αποτελέσματα δείχνουν ότι το προτεινόμενο κριτήριο υπερέχει αισθητά των υπαρχόντων και επιτυγχάνει επιδόσεις κοντά σε αυτές πολύ πιο σύνθετων και αργών αλγορίθμων (βλ. [84] και [79] για τις επιδόσεις των αλγορίθμων αυτών).

Ως μελλοντική έρευνα, θα εξετάσουμε την επίδοση του κριτηρίου σε άλλα πεδία εφαρμογής (όπως συσκέψεις), το ενδεχόμενο κέρδος από τη χρήση κλειστών τύπων υπολογισμού έναντι της προσέγγισης Laplace και θα επιδιώξουμε να βρούμε εναλλακτικούς τρόπους ερμηνείας του κριτηρίου, κυρίως μέσω της αναθεώρησης της εξάρτησης των εκ των προτέρων κατανομών από τον αριθμό

των δειγμάτων. Αναφέρουμε τέλος ότι πειραματιστήκαμε με τη χρήση εκ των προτέρων κατανομών επί της ομαδοποίησης $\mathbf{x} \in \mathcal{X}^n$, όπως αναφερθήκαμε στην παράγραφο 5.4.1, αλλά επιτύχαμε ανεπαίσθητη αλλαγή στην επίδοση του κριτηρίου. Παρ΄ όλα αυτά, απαιτείται μια πιο λεπτομερής μελέτη όσον αφορά στην ρύθμιση των αντίστοιχων υπερπαραμέτρων. Ο αλγόριθμος της ιεραρχικής συγκέντρωσης είναι επίσης πολύ περιοριστικός για την ανάδειξη της συνεισφοράς. Επομένως, συμπεριλαμβάνουμε και την εξέταση του κριτηρίου μαζί με τη μοντελοποίηση της χρονικής πληροφορίας με χρήση Γενετικών Αλγορίθμων ή Προσομοιωμένης Ανόπτησης (Simulated Annealing), [52]. Για πληρότητα, η προτεινόμενη Μπεϋσιανή μοντελοποίηση της χρονικής πληροφορίας παρατίθεται στην παράγραφο 8.3 του Παραρτήματος 1.

6 ΑΥΤΟΜΑΤΗ ΟΜΑΔΟΠΟΙΗΣΗ ΟΜΙΛΗΤΩΝ ΜΕ ΧΡΗΣΗ ΤΟΥ ΑΛΓΟΡΙΘΜΟΥ ΜΕΤΑΤΟΠΙΣΗΣ ΤΟΥ ΜΕΣΟΥ

Περίληψη

Στο κεφάλαιο αυτό εισάγουμε τον αλγόριθμο της μετατόπισης του μέσου (mean-shift) στο πρόβλημα της ομαδοποίησης ομιλητών. Ο αλγόριθμος αυτός επιχειρεί να ομαδοποιήσει ένα σύνολο από παρατηρήσεις σε έναν άγνωστο εκ των προτέρων αριθμό ομάδων μέσω του εντοπισμού των τρόπων (modes) της συνάρτησης πυκνότητας πιθανότητας. Χρησιμοποιείται εκτενώς στην όραση υπολογιστών και συγκεκριμένα στα προβλήματα της κατάτμησης εικόνας (segmentation), εξομάλυνσης με διατήρηση των ασυνεχειών (discontinuity-preserving smoothing) καθώς και παρακολούθησης αντικειμένων (object tracking) σε βίντεο.
Με τον αλγόριθμο αυτόν επιχειρούμε να προσφέρουμε μια εναλλακτική στρατηγική στο πρόβλημα της ομαδοποίησης τμημάτων ομιλίας, κυρίαρχη θέση στο οποίο κατέχει η συγκολλητική ιεραρχική ομαδοποίηση. Τονίζουμε ότι είναι η πρώτη απόπειρα εφαρμογής του συγκεκριμένου αλγορίθμου σε προβλήματα αναγνώρισης φωνής και ομιλητή. Η πρώτη εργασία μας σε αυτή τη μέθοδο παρουσιάζεται στο ([7]) ενώ η πλήρης εκδοχή του (την οποία και παρουσιάζουμε) έχει αποσταλεί για δημοσίευση σε διεθνές επιστημονικό περιοδικό της εκμάθησης μηχανών.

6.1 Εισαγωγή

6.1.1 Τα μειονεκτήματα της ιεραρχικής ομαδοποίησης

Πρωτού προχωρήσουμε στην ανάλυση της προτεινόμενης μεθόδου, χρήσιμο είναι να εντοπίσουμε τα μειονεκτήματα της ιεραρχικής συγκολλητικής ομαδοποίησης (Agglomerative Hierarchical Clustering), του αλγορίθμου δηλαδή που χρησιμοποιείται κατά κόρον στη βαθμίδα ομαδοποίησης ομιλητών,

δεδομένης μίας αρχικής κατάτμησης σε τμήματα ομιλίας. Βασική ιδέα πίσω από τον συγκεκριμένο αλγόριθμο είναι η αναδρομική ομαδοποίηση των πιο κοντινών στατιστικά ομάδων υπό κοινή ομάδα, σύμφωνα με ένα κριτήριο στατιστικής απόκλισης και ο τερματισμός της διαδικασίας έως ότου όλες οι εκατέρωθεν αποκλίσεις είναι πάνω από ένα προκαθορισμένο κατώφλι. Τρία είναι τα βασικότερα μειονεκτήματα αυτής της προσέγγισης.

1. Αδυναμία διόρθωσης μίας εσφαλμένης ένωσης δύο τμημάτων ομιλίας υπό κοινή ομάδα, σφάλμα το οποίο διαχέεται στις επόμενες αναδρομές. Βλέποντας τον αλγόριθμο ως μέσο βελτιστοποίησης μιας αντικειμενικής συνάρτησης, το πρόβλημα ομοιάζει με σύγκλιση του αλγορίθμου σε τοπικά ακρότατα. Λύση στο συγκεκριμένο πρόβλημα είναι η επανεξέταση του δεντρογράμματος που παράγει ο αλγόριθμος, το κλάδεμά του (dendrogram pruning) σε εναλλακτικούς κόμβους και η επιλογή της καταλληλότερης τελικής ομαδοποίησης, σύμφωνα με κάποιο κριτήριο. Άλλες λύσεις είναι η εφαρμογή τεχνικών εξομοιούμενης ανόπτησης (simulated annealing) ή και γενετικών αλγορίθμων [52], τεχνικές οι οποίες αντιμετωπίζουν το πρόβλημα της σύγκλισης σε τοπικά ελάχιστα (- μέγιστα) επιτρέποντας τις μεταβάσεις οι οποίες ελαττώνουν (- αυξάνουν) πρόσκαιρα την αντικειμενική συνάρτηση, με την πιθανότητα επιλογής τέτοιας μετάβασης να βαίνει μειούμενη στον χρόνο (δηλ. στις αναδρομές).

2. Η γενικότερη φιλοσοφία της βήμα-προς-βήμα ενοποίησης των πιο κοντινών στατιστικά τμημάτων είναι περιοριστική, καθώς επιβάλλει την αποκλειστική ανάθεση (hard-assignment) των τμημάτων σε ομάδες και όχι την πιθανοτική. Το ανάλογο σε τεχνικές μη-επιβλεπόμενης ομαδοποίησης με εκ των προτέρων γνώση του αριθμού των ομάδων είναι ο αλγόριθμος των K-μέσων (ή ο Τμηματικός αλγόριθμος K-μέσων (Segmental-K-means) για Υπονοούμενα Μαρκοβιανά Μοντέλα, HMMs) με την οικογένεια Expectation-Maximization (Baum-Welch για HMMs). Στη δεύτερη περίπτωση, αποκλειστική ανάθεση των παρατηρήσεων σε ομάδες λαμβάνει χώρα στον τερματισμό του αλγορίθμου και μόνο αν ο απώτερος στόχος είναι η αποκλειστική ανάθεση των παρατηρήσεων και όχι η εκτίμηση της συνάρτησης πυκνότητας πιθανότητας από το τυχαίο δείγμα.

3. Η ανάγκη ορισμού μίας αντικειμενικής συνάρτητης (π.χ. ελαχίστα τετράγωνα αποστάσεων

παρατηρήσεων - κέντρου ομάδας) είτε μίας μετρικής απόκλισης και καθορισμός κατάλληλου κατωφλίου απόφασης. Στις περιπτώσεις αυτές συγκαταλέγονται και τα ημιπαραμετρικά μοντέλα, όπως αυτά των μιγμάτων κανονικών κατανομών, όπου τη θέση της αντικειμενικής συνάρτησης καταλαμβάνει η πιθανοφάνεια ή η εκ των υστέρων πιθανότητα. Σε πολλές εφαρμογές, μία τέτοια συνάρτηση μπορεί να είναι αυθαίρετη και να μην οδηγεί στη φυσική ομαδοποίηση των παρατηρήσεων. Οι συναρτήσεις αυτές οδηγούν συνήθως σε ομάδες προκαθορισμένων σχημάτων (π.χ. υπερ-σφαιρών, υπερ-ελλείψεων, κ.ο.κ.) που δεν συμπίπτουν κατ΄ ανάγκη με τα σχήματα των ομάδων. Χρήσιμο είναι λοιπόν να εξετασθούν αλγόριθμοι ομαδοποίησης οι οποίοι θέτουν λιγότερους περιορισμούς γύρω από τα σχήματα των ομάδων.

Οι παραπάνω αδυναμίες του αλγορίθμου ιεραρχικής ομαδοποίησης έδωσε το έναυσμα διερεύνησης των δυνατοτήτων του αλγορίθμου της μετατόπισης του μέσου ως εναλλακτικής προσέγγισης στο πρόβλημα της ομαδοποίησης τμημάτων ομιλίας.

6.1.2 Σύντομη ιστορική αναδρομή και γενικά χαρακτηριστικά του αλγορίθμου

Η ιστορία του αλγορίθμου ξεκινάει από τη δημοσίευση των Fukunaga και Hostetler το 1975, [85]. Η δημοσίευση αυτή δεν έτυχε της προσοχής της επιστημονικής κοινότητας έως το 1995, όταν ο Cheng επανέφερε τον προβληματισμό, γενικεύοντας και κάποια από τα αποτελέσματα όσον αφορά στους πυρήνες των παραθύρων Parzen, [86]. Η μεγάλη ώθηση στον αλγόριθμο δόθηκε από τις δημοσιεύσεις του D. Comaniciu, [87], [88], όπου ο αλγόριθμος εφαρμόζεται σε πρόβλήματα όρασης υπολογιστών, κυρίως στην κατάτμηση εικόνας και εξομάλυνση με διατήρηση των ασυνεχειών. Στο [87], οι Comaniciu et al. επεκτείνουν τον αλγόριθμο έτσι ώστε το εύρος (bandwidth) των παραθύρων Parzen να είναι μεταβλητό.

Ας δώσουμε, αρχικά, μια γενική περιγραφή του αλγορίθμου μετατόπισης του μέσου. Οι αναγκαίες τροποποιήσεις ώστε ο αλγόριθμος να είναι σε θέση να αντιμετωπίσει το πρόβλημα της ομαδοπο-

ίησης ομιλητών θα εξετασθούν στη συνέχεια. Θεωρούμε έτσι ένα σύνολο ανεξάρτητων και όμοια κατανεμημένων (i.i.d.) παρατηρήσεων στον d-διάστατο χώρο. Το σύνολο αυτό είναι ένα τυχαίο δείγμα από μία άγνωστη συνάρτηση πυκνότητας πιθανότητας f για την οποία δεν κάνουμε καμία υπόθεση όσον αφορά στη δομή της, πέραν της ομαλότητάς της. Σε τέτοιες περιπτώσεις, η εκτίμηση της f αντιμετωπίζεται με μη-παραμετρικές μεθόδους, όπως οι μέθοδοι των παραθύρων Parzen. Ένα βασικό μειονέκτημα των μη-παραμετρικών μεθόδων εκτίμησης είναι η εκθετική ως προς το d πολυπλοκότητα. Ένα δεύτερο είναι η αραιότητα (sparsity) των παρατηρήσεων για μια στιβαρή εκτίμηση της f. Τα προβλήματα αυτά είναι γνωστά ως *κατάρα της διάστασης* (curse of dimensionality).

Σε πολλές εφαρμογές, ωστόσο, δεν ενδιαφερόμαστε τόσο για μία εκτίμηση της f, όσο για ορισμένα χαρακτηριστικά της. Στην εφαρμογή που εξετάζουμε τα χαρακτηριστικά αυτά είναι οι τρόποι της f (modes), δηλαδή τα σημεία στα οποία η f παρουσιάζει μέγιστο. Ο αλγόριθμος της μετατόπισης του μέσου απαντά σε αυτό ακριβώς το πρόβλημα. Επιτυγχάνει να εντοπίσει τα μέγιστα της f χωρίς την απόπειρα εκτίμησης της ίδιας της f. Επιπλέον, δεν υπολογίζει μόνο τα μέγιστα (τους τρόπους) της f αλλά οριοθετεί και τις περιοχές στον d-διάστατο χώρο σύμφωνα με τον τρόπο από τον οποίον κάθε σημείο του χώρου έλκεται. Δημιουργεί, δηλαδή, μία περιοχή αυθαίρετου (arbitrary) σχήματος γύρω από κάθε τρόπο η οποία αποτελείται από το σύνολο των σημείων εκείνων τα οποία έλκονται από τον συγκεκριμένο τρόπο. Έτσι, πέραν της εκτίμησης των τρόπων παρέχει και μια φυσική ομαδοποίηση των παρατηρήσεων σε ομάδες άγνωστου εκ των προτέρων αριθμού, όπου κάθε ομάδα ορίζεται ως η δεξαμενή έλξης του κάθε τρόπου (basin of attraction).

6.1.3 Η πολυτροπικότητα των κατανομών διανυσμάτων παρατήρησης ομιλητή

Όπως αναφέραμε και στην εισαγωγή και με δεδομένη τη χρήση των MFCC ή PLP ως διανυσμάτων χαρακτηριστικών, η χρήση του αλγορίθμου μετατόπισης του μέσου είναι πρακτικά αδύνατον να οδηγήσει στην επιθυμητή ομαδοποίηση. Αυτό γιατί οι συγκεκριμένοι παραμετρικοί χώροι έχουν

δημιουργηθεί για το πρόβλημα της αναγνώρισης φωνής και όχι ομιλητή. Ως εκ τούτου, οδηγούν σε πολυτροπικές κατανομές - με τρόπους που αντιστοιχούν στα φωνήματα - ακόμα και αν οι παρατηρήσεις προέρχονται από έναν και μονό ομιλητή και με τις ίδιες συνθήκες ηχογράφησης. Η εφαρμογή, λοιπόν, του αλγορίθμου σε αρχεία ήχου άνω του ενός ομιλητών οδηγεί σε ομαδοποίηση βασισμένη στα φωνήματα και όχι στους ομιλητές. Έτσι, οι αλγόριθμοι ομαδοποίησης αλλά και αναγνώρισης ομιλητών χρησιμοποιούν πάντοτε περιγραφικά (descriptive statistics) των παρατηρήσεων τμημάτων ομιλίας έναντι των ίδιων των παρατηρήσεων.

Ταυτόχρονα, η εμφάνιση των μεθόδων πυρήνα και άλλων συγγενών τεχνικών έστρεψε το ενδιαφέρον σε μεθόδους συνδυασμού των περιγραφικών στατιστικών με στατιστικές μεθόδους διαχωρισμού (discriminative statistics). Παράδειγμα από τον χώρο της αναγνώρισης ομιλητών αποτελεί η χρήση των μηχανών διανυσμάτων υποστήριξης (Support Vector Machines, SVM) σε παραμετρικούς χώρους που προέρχονται από περιγραφικά στατιστικά, όπως τα μίγματα κανονικών κατανομών [89], [90]. Τη θέση των παρατηρήσεων καταλαμβάνουν συναρτήσεις με όρισμα (ημι-)παραμετρικά μοντέλα.

Η προτεινόμενη μέθοδος αποτελεί μία πρώτη απόπειρα τροποποίησης ενός αλγορίθμου που δημιουργήθηκε για να ομαδοποιεί παρατηρήσεις με τρόπο ώστε να ομαδοποιεί τμήματα ομιλίας που περιγράφονται με παραμετρικά μοντέλα. Αν και στην παρούσα φάση θα εξετάσσουμε μόνο τη μοντελοποίηση των τμημάτων με μία κανονική κατανομή, δηλαδή $\theta_k = (\mu_k, \Sigma_k)$, ο αλγόριθμος ενδέχεται να αποδώσει και με πιο σύνθετες μοντελοποιήσεις, όπως μίγματα κανονικών κατανομών.

6.2 Αναλυτική περιγραφή του αλγορίθμου

6.2.1 Μη παραμετρική εκτίμηση συνάρτησης πυκνότητας πιθανότητας

Έστω ένα σύνολο n παρατηρήσεων $\{\mathbf{y}^{(i)}\}_{i=1}^{n}$, $\mathbf{y}^{(i)} \in \Re^d$, το οποίο θεωρούμε ως τυχαίο δείγμα από μία συνάρτηση πυκνότητας πιθανότητας f. Η μη παραμετρική εκτίμηση $\hat{f}$ της f μέσω παραθύρων

Parzen ορίζεται ως εξής

$$\hat{f}(\mathbf{y}) = \frac{1}{n}\sum_{i=1}^{n} K_{\mathbf{H}}(\mathbf{y} - \mathbf{y}^{(i)}) \tag{113}$$

όπου

$$K_{\mathbf{H}}(\mathbf{y}) = |\mathbf{H}|^{-1/2} K(\mathbf{H}^{-1/2}\mathbf{y}) \tag{114}$$

είναι ο πυρήνας της μεθόδου και $\mathbf{H}$ ένας συμμετρικός και θετικά ορισμένος $d \times d$ πίνακας. Μία συνήθης πρακτική είναι να θεωρούμε σφαιρικούς πυρήνες, δηλαδή $\mathbf{H} = h^2\mathbf{I}$, όπου $\mathbf{I}$ ο μοναδιαίος πίνακας. Έτσι, η μόνη ρυθμιζόμενη παράμετρος είναι το εύρος του πυρήνα h. Μέθοδοι εκτίμησης του βέλτιστου h αναλύονται στο [87] και θα παρουσιασθούν στη συνέχεια.

Εστιάζουμε τώρα σε μία συγκεκριμένη κλάση πυρήνων, αυτή των σφαιρικά συμμετρικών. Οι πυρήνες αυτοί ικανοποιούν την παρακάτω σχέση

$$K(\mathbf{y}) = c_{k,d}\, k(||\mathbf{y}||^2) \tag{115}$$

όπου $c_{k,d}$ τέτοιο ώστε $\int_{\Re^d} K(\mathbf{y})d\mathbf{y} = 1$, ενώ η συνάρτηση $k(y), y \geq 0$ ορίζεται ως το προφίλ του πυρήνα. Ένας δημοφιλής πυρήνας είναι ο κανονικός ή γκαουσσιανός πυρήνας

$$K_N(\mathbf{y}) = (2\pi)^{d/2} \exp\left(-\frac{1}{2}||\mathbf{y}||^2\right) \tag{116}$$

με προφίλ το ακόλουθο

$$k(y) = \exp\left(-\frac{1}{2}y\right) \tag{117}$$

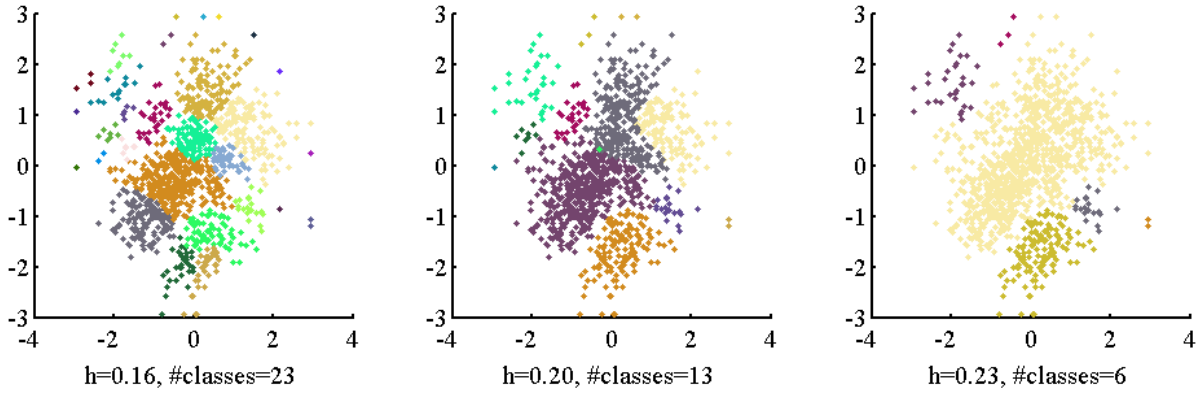
Ένας δεύτερος είναι ο πυρήνας του Epanechnikov

$$K_E(\mathbf{y}) = \begin{cases} \frac{1}{2}c_d^{-1}(d+2)(1-||\mathbf{y}||^2) & \text{για } ||\mathbf{y}|| < 1 \\ 0 & \text{για } ||\mathbf{y}|| \geq 1 \end{cases} \tag{118}$$

με αντίστοιχο προφίλ το ακόλουθο

$$k_E(y) = \begin{cases} 1-y & \text{για } y < 1 \\ 0 & \text{για } y \geq 1 \end{cases} \tag{119}$$

Το σχήμα 19 δείχνει ένα παράδειγμα της ευαισθησίας ως προς το εύρος πυρήνα h. Καθένα από τα τρία γραφήματα αντιστοιχεί σε διαφορετικές ομαδοποιήσεις που προκαλούνται αυξάνοντας το h.



Σχῆμα 19: *Παράδειγμα ομαδοποίησης διδιάστατων διανυσμάτων παρατήρησης (πρώτοι και δεύτεροι συντελεστές MFCC) από τμήμα φωνής διάρκειας 14 δευτερολέπτων. Η ευαισθησία της μεθόδου ως προς την επιλογή εύρους πυρήνα h φαίνεται από το πλήθος των ομάδων που δημιουργούνται. Τα χρώματα υποδηλώνουν τις δεξαμενές έλξης κάθε τρόπου, ενώ ο πυρήνας είναι Γκαουσσιανός.*

6.2.2 Εκτίμηση της κλήσης της πυκνότητας πιθανότητας

Όπως αναφέραμε παραπάνω, ο αλγόριθμος της μετατόπισης του μέσου δεν προβαίνει σε εκτίμηση της f, αλλά επιχειρεί τον εντοπισμό των τοπικών μεγίστων της. Δεδομένου ότι για τα σημεία αυτά θα ισχύει $\nabla f = \mathbf{0}$ θέτουμε

$$\hat{\nabla} f_{h,K}(\mathbf{y}) \equiv \nabla \hat{f}_{h,K}(\mathbf{y}) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{y} - \mathbf{y}^{(i)}) k' \left(\left\| \frac{\mathbf{y} - \mathbf{y}^{(i)}}{h} \right\|^2 \right) \tag{120}$$

όπου $k'(y)$ η παράγωγος της $k(y)$ ως προς y.

Ορίζουμε τώρα τη συνάρτηση $g(y) = -k'(y)$ και υποθέτουμε ότι η παράγωγος ορίζεται, εκτός ίσως από ένα πεπερασμένο σύνολο σημείων. Ένα τέτοιο σύνολο είναι τα $\mathbf{y}$ για τα οποία $\left\| \frac{\mathbf{y}-\mathbf{y}^{(i)}}{h} \right\| = 1$ για τον πυρήνα του Epanechnikov. Ο αρχικός πυρήνας $K(\mathbf{y})$ καλείται η σκιά (shadow) του πυρήνα $G(\mathbf{y})$, όπου

$$G(\mathbf{y}) = c_{g,d} g\left(\|\mathbf{y}\|^2 \right) \tag{121}$$

Για παράδειγμα, ο πυρήνας του Epanechnikov είναι η σκιά του ομοιόμορφου πυρήνα, δηλαδή του πυρήνα με προφίλ το ακόλουθο

$$g(y) = \begin{cases} 1 & \text{για } y < 1 \\ 0 & \text{για } y \geq 1 \end{cases} \tag{122}$$

ενώ ο κανονικός πυρήνας έχει το ίδιο προφίλ με τη σκιά του. Αντικαθιστώντας στη σχέση (120) τον πυρήνα $G(\mathbf{y})$ της (121) έχουμε

$$\hat{\nabla} f_{h,K}(\mathbf{y}) = \frac{2c_{k,d}}{nh^{d+2}} \sum_{i=1}^{n} (\mathbf{y}^{(i)} - \mathbf{y}) g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right) \tag{123}$$

η οποία ισοδυναμεί με

$$\hat{\nabla} f_{h,K}(\mathbf{y}) = \frac{2c_{k,d}}{nh^{d+2}} \left[\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right) \right] \left[\frac{\sum_{i=1}^{n} \mathbf{y}^{(i)} g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right)} - \mathbf{y} \right] \tag{124}$$

Και οι δύο όροι της παραπάνω έκφρασης παρουσιάζουν ενδιαφέρον. Ο πρώτος είναι ανάλογος της εκτίμησης πυκνότητας στο σημείο $\mathbf{y}$ υπολογισμένη με τον πυρήνα G, δηλαδή

$$\hat{f}_{h,G}(\mathbf{y}) = \frac{c_{g,d}}{nh^d} \sum_{i=1}^{n} (\mathbf{y} - \mathbf{y}^{(i)}) g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right). \tag{125}$$

Ο δεύτερος όρος είναι η *μετατόπιση του μέσου*

$$\mathbf{m}_{h,G}(\mathbf{y}) = \frac{\sum_{i=1}^{n} \mathbf{y}^{(i)} g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right)}{\sum_{i=1}^{n} g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right)} - \mathbf{y} \tag{126}$$

Ο όρος αυτός είναι η διαφορά μεταξύ του μέσου όρου των σημείων - με βάρος $g\left(\left\| \frac{\mathbf{y}^{(i)} - \mathbf{y}}{h} \right\|^2 \right)$ για το $\mathbf{y}^{(i)}$ - και του $\mathbf{y}$. Χρησιμοποιώντας τους ορισμούς (125) και (126), η μετατόπιση του μέσου γράφεται και ως εξής

$$\mathbf{m}_{h,G}(\mathbf{y}) = \frac{1}{2} h^2 c_{g,d} \frac{\hat{\nabla} f_{h,K}(\mathbf{y})}{\hat{f}_{h,G}(\mathbf{y})} \tag{127}$$

Η έκφραση (127) δείχνει τα εξής. Η κατεύθυνση της μετατόπισης του μέσου δίνεται από την κλίση της f στο $\mathbf{y}$, δηλαδή την κατεύθυνση της μέγιστης αύξησης της πυκνότητας. Επιπλέον, ο παρονομαστής κανονικοποιεί την μετατόπιση, έτσι ώστε σε σημεία με χαμηλή πυκνότητα (υπολογισμένης με πυρήνα G) το μέτρο της μετατόπισης να είναι μεγάλο και αντίστροφα.

Εναλλακτική εξαγωγή Ο κανονικός πυρήνας έχει την ιδιότητα να ταυτίζεται με τη σκιά του, δηλαδή $k_N(y) = g_N(y)$. Έτσι, από τα ιδιαίτερα χαρακτηριστικά του κανονικού πυρήνα είναι ότι η μετατόπιση του μέσου ορίζεται ως

$$\mathbf{m}_{h,G_N}(\mathbf{y}) = \frac{1}{2}h^2 \nabla \log \hat{f}_{h,G_N}(\mathbf{y}) \tag{128}$$

Παρατηρούμε λοιπόν ότι ο αλγόριθμος ταυτίζεται με αυτόν της απότομης ανόδου (gradient ascend) μετά από λογαρίθμιση, εφόσον ο πυρήνας είναι κανονικός. Τονίζουμε αυτή την παρατήρηση, καθώς όπως θα δείξουμε στη συνέχεια οδηγεί σε δύο διαφορετικές προσεγγίσεις αν θελήσουμε να τον επεκτείνουμε σε μη-ευκλείδιες γεωμετρίες. Υπό αυτή τη μορφή, η εξαγωγή της αναδρομικής σχέσης απαιτεί τη χρήση της φυσικής κλίσης (natural gradient), καθώς τα πιθανοτικά μοντέλα υπακούουν στη γεωμετρία του Riemann. Σημειώνουμε επίσης και η παραπάνω σχέση κάνει χρήση της φυσικής κλίσης. Η πολλαπλασιαστική σταθερά $\frac{1}{2}h^2$ δεν είναι άλλη από τον αντίστροφο του τανυστή μετρικής του χώρου, ο οποίος είναι

$$\mathcal{I}_{\mathbf{y}}(\mathbf{y}) = \frac{2}{h^2} I_d \tag{129}$$

όπου I_d ο d-διάστατος μοναδιαίος πίνακας. Αγνοώντας τη σταθερά 2 (που προήλθε από τη μητήρηση της συμβατικού τύπου της κανονικής κατανομής) βλέπουμε ότι ο $\mathcal{I}_{\mathbf{y}}(\mathbf{y})$ είναι ο πίνακας ακρίβειας (precision matrix) της κατανομής. Εντελώς αντίστοιχα, θα δούμε ότι στον χώρο των παραμέτρων ο πίνακας αυτός θα είναι η πληροφορία Fisher, δηλαδή ο τανυστής μετρικής των παραμετρικών μοντέλων.

Ένα δεύτερο πλεονέκτημα της συγκεκριμένης εκδοχής του αλγορίθμου έγκειται στο ότι μας απαλλάσσει από τον περιορισμό στη χρήση αποστάσεων, οι παράγωγοι των οποίων είναι γραμμικές ως προς $\mathbf{y}$, (σχ. 124). Αυτό θα μας επιτρέψει να χρησιμοποιήσουμε ένα σύνολο αποκλίσεων, οι οποίες δεν πληρούν κατ΄ ανάγκη αυτή την ιδιότητα, γενικεύοντας έτσι την ανάλυσή μας.

6.2.3 Ο αλγόριθμος και η σύγκλιση στους τρόπους της πυκνότητας

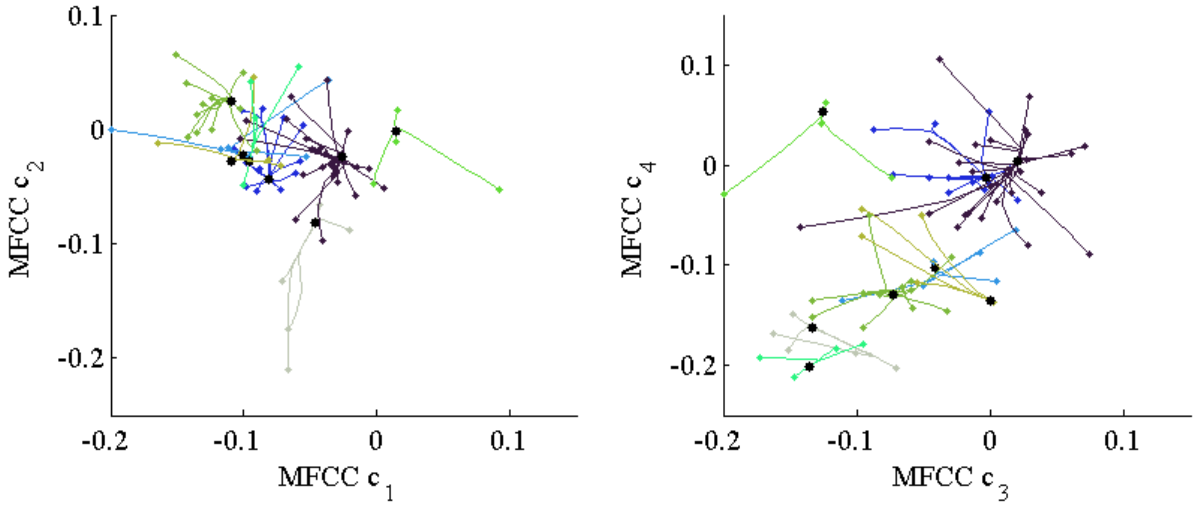
Από την παραπάνω ανάλυση, εύκολα προκύπτει η αναδρομική σχέση του αλγορίθμου.
Για κάθε $\mathbf{y}^{(i)}, i = 1, \ldots, n$ θέτουμε $\mathbf{y} = \mathbf{y}^{(i)}$ και αναδρομικά

- υπολογίζουμε τη μετατόπιση $\mathbf{m}_{h,G}(\mathbf{y})$,
- ενημέρωνουμε $\mathbf{y} \leftarrow \mathbf{y} + \mathbf{m}_{h,G}(\mathbf{y})$.

έως ότου επιτευχθεί σύγκλιση.

Ας συμβολίσουμε με $\{\mathbf{y}_j\}_{j=1,2\ldots}$ την ακολουθία που δημιουργείται από την παραπάνω αναδρομή. Ο παραπάνω αλγόριθμος διασφαλίζει ότι η ακολουθία συγκλίνει σε ένα τοπικό μέγιστο της f. Επιπλέον, η ακολουθία $\hat{f}_{h,G}(\mathbf{y}_j), j = 1, 2, \ldots$ είναι μονοτονικά αυξανόμενη, έως ότου συγκλίνει στο τοπικό μέγιστο.

Ο αλγόριθμος της ομαδοποίησης ολοκληρώνεται ομαδοποιώντας τα $\{\mathbf{y}^{(i)}\}_{i=1}^{n}$ σύμφωνα με τον



Σχῆμα 20: *Παράδειγμα ομαδοποίησης τμημάτων ομιλίας σε ομιλητές. Τα χρώματα υποδηλώνουν τον τρόπο στον οποίο συγκλίνει κάθε τμήμα.*

τρόπο στον οποίο συνέκλινε η αντίστοιχη ακολουθία. Στο Σχ. 20 σκιαγραφείται ένα παράδειγμα

ομαδοποίησης σε ομιλητές, σύμφωνα με τη μέθοδο που αναπτύξαμε παραπάνω. Τα χρώματα υποδηλώνουν τον τρόπο στον οποίο συγκλίνει κάθε τμήμα, δηλαδή κάθε χρώμα αποτελεί και μία ομάδα (ιδανικά τα τμήματα που ανήκουν σε κάθε ομιλητή). Όπως θα αναλύσουμε στη συνέχεια, τα σημεία $\mathbf{y}^{(i)}, i = 1, \ldots, n$ είναι οι τιμές των παραμέτρων με τις οποίες παραμετροποιούμε την κατανομή κάθε τμήματος ομιλίας. Όπως θα δούμε επίσης, οι πυρήνες έχουν το αντίστοιχό τους στις εκ των υστέρων κατανομές των παραμέτρων αυτών.

6.3 Η εφαρμογή των ιδεών αυτών στο πρόβλημα της ομαδοποίησης ομιλητών

6.3.1 Εισαγωγή

Όπως ήδη αναφέραμε, το πρόβλημα της ομαδοποίησης ομιλητών δεν ανήκει στην κλάση των προβλημάτων ομαδοποίησης παρατηρήσεων (επιπέδου συνάρτησης εκπομπής), καθώς η οντότητα με την οποία αναπαρισθούμε το κάθε τμήμα ομιλίας είναι ένα στατιστικό παραμετρικό μοντέλο. Έτσι, αντί των παρατηρήσεων $\mathbf{y}^{(i)}, i = 1, \ldots, n$ θα πρέπει να θεωρήσουμε ένα σύνολο από κατανομές, τις οποίες παραμετροποιούμε ως $\boldsymbol{\theta}^k, k = 1, \ldots, N$. Η διαφορά αυτή είναι θεμελιώδης για πολλούς και διαφόρους λόγους, τους οποίους και απαριθμίζουμε.

(α) Αμφιβολία ως προς την εκτίμηση

Η πρώτη θεμελιώδης διαφορά έγκειται στη μοντελοποίηση της αμφιβολίας μας ως προς την εκτίμηση των παραμέτρων, αφού κάθε τμήμα ομιλίας αποτελείται από ένα σύνολο πεπερασμένων δειγμάτων. Ως εκ τούτου, η αναπαράσταση της εμπειρικής κατανομής ως ένα άθροισμα παλμών Dirac κεντραρισμένους στην ML ή MAP εκτίμηση είναι υποβέλτιστη. Αντίθετα, και με δεδομένο ότι η μοντελοποίηση που επιλέγουμε είναι Μπεϋσιανή, θα πρέπει να αντικαταστήσουμε τους παλμούς Dirac με κατανομές πιθανότητας. Όσο το πλήθος των παρατηρήσεων για ένα τμήμα ομιλίας τείνει στο άπειρο, η κατανομή αυτή θα τείνει σε έναν παλμό Dirac στην ML εκτίμηση. Στο

επίπεδο πιθανοφάνειας, μπορούμε να παρομοιάσουμε την μοντελοποίηση των $\boldsymbol{\theta}_k, k = 1, \ldots, N$ με τον παραπάνω αλγόριθμο, όπου στην εικόνα έχει προστεθεί λευκός κανονικός θόρυβος με μηδενική μέση τιμή, μεταβλητής ισχύος, όπου η τιμή της ισχύς του είναι γνωστή για κάθε εικονοστοιχείο. Στην προκείμενη περίπτωση, η προς εξομάλυνση κατανομή δεν θα ήταν πλέον η εμπειρική κατανομή (παλμοί Dirac), αλλά κανονικές κατανομές με κέντρο τις παρατηρήσεις και συμμεταβλητότητα ίση με την ισχύ του θορύβου κάθε εικονοστοιχείου. Δεδομένου μάλιστα του ότι η συνέλιξη δύο κανονικών κατανομών $\boldsymbol{\theta}^k = (\mu^k, \Sigma^k)$ και $\boldsymbol{\theta}^0 = (0, \Sigma^0)$, οδηγεί και πάλι σε κανονική κατανομή $\boldsymbol{\theta}^{k'} = (\mu^k, \Sigma^k + \Sigma^0)$, παρατηρούμε ότι για την περίπτωση του κανονικού προσθετικού θορύβου μηδενικής μεσης τιμής, η λύση ισοδυναμεί με εξομάλυνση των παλμών Dirac με χρήση κανονικού πυρήνα μεταβλητού εύρους.

(β) Τετραγωνικές αποστάσεις και στατιστικές αποκλίσεις

Όπως είδαμε παραπάνω, η εξαγωγή της αναδρομικής σχέσης του αλγορίθμου απαιτεί τη χρήση τετραγωνικών ευκλείδιων αποστάσεων. Δεδομένου όμως ότι οι οντότητές μας αναπαριστούν κατανομές πιθανότητας, η χρήση τετραγωγικών ευκλείδιων αποστάσεων είναι μη αποδεκτή λύση. Αντιθέτως, η στατιστική έχει αναπτύξει μία πληθώρα οικογενειών αποκλίσεων μεταξύ κατανομών. Τα δύο πιο διαδεδομένα παραδείγματα τέτοιων αποκλίσεων είναι η απόκλιση Kullback-Leibler και η απόσταση Hellinger. Αν περιοριστούμε μάλιστα στις οικογένειες των εκθετικών κατανομών, οι αποκλίσεις αυτές - όπως άλλωστε και η τετραγωνική ευκλείδια απόσταση, η Mahalanobis, η Itakura–Saito, κ.ά. - αποτελούν μέλη των αποκλίσεων του Bregman, [91]. Τονίζουμε επίσης ότι η Μπεϋσιανή στατιστική είναι στενά συνδεδεμένη με τις αποκλίσεις αυτές, καθώς πολλές από τις οικογένειες των εκ των προτέρων κατανομών (συζυγείς, εντροπικές, κ,ά.) βασίζονται ακριβώς σε αυτές τις αποκλίσεις.

(γ) Ανάγκη χρήσης μεθόδων γεωμετρίας της πληροφορίας

Ένα βασικό χαρακτηριστικό των κατανομών πιθανότητας είναι η μη-ευκλείδια γεωμετρία τους. Εν αντιθέσει με τον αρχικό αλγόριθμο, η γεωμετρία των κατανομών ακολουθεί τη Γεωμετρία του Riemann, η οποία χαρακτηρίζεται από θετικά ορισμένο μεν, μη σταθερό δε τανυστή μετρικής (metric tensor), που ταυτίζεται με τον πίνακα πληροφορίας του Fisher, $\mathcal{I}(\boldsymbol{\theta})$. Το γεγονός αυτό, αν δε ληφθεί υπ' όψη, οδηγεί σε εσφαλμένες μεθόδους διαφόρισης των στατιστικών μεγεθών και σε υποβέλτιστους κανόνες κλίσης (gradient rules). Ως εκ τούτου, για να καταλήξουμε στον απαιτούμενο αναδρομικό τύπο, θα απαιτηθεί η χρήση της θεωρίας τανυστών, της συμμεταβλητής παραγώγισης (covariant derivative) καθώς και της φυσικής κλίσης (natural gradient).

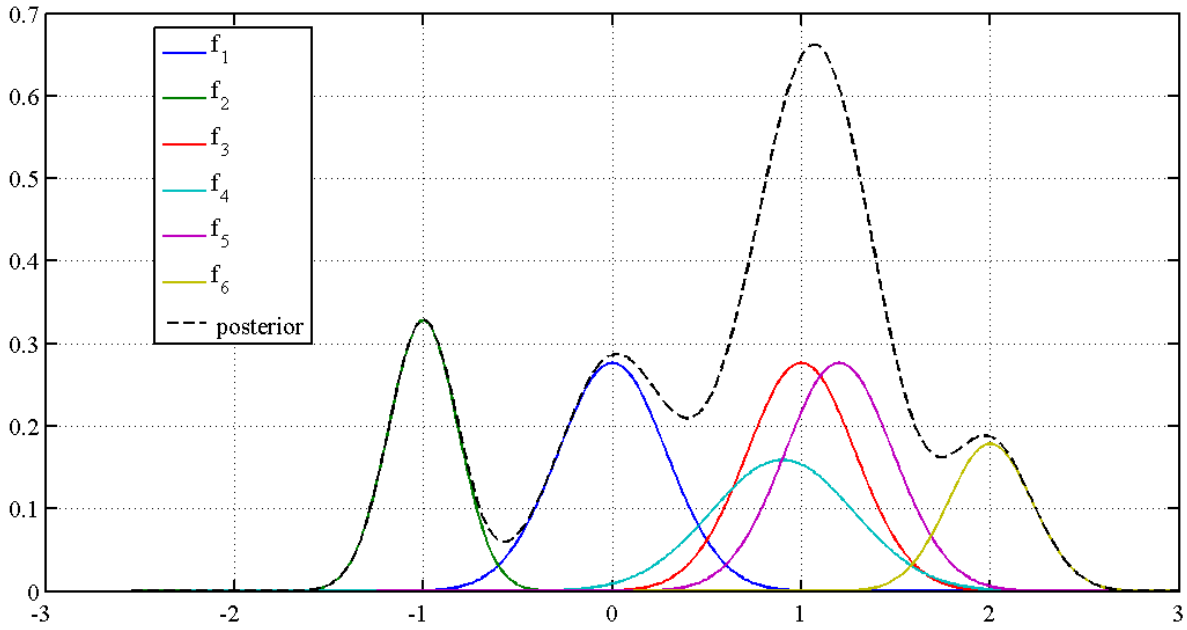
6.3.2 Μπεϋσιανή μοντελοποίηση του προβλήματος

Χαρακτηριστικό της Μπεϋσιανής στατιστικής είναι η πιθανοτική μοντελοποίηση της αμφιβολίας μας όσον αφορά τις παραμέτρους. Έτσι, ο πυρήνας μας έχει ως φυσική προέκταση την εκ των υστέρων κατανομή των $\boldsymbol{\theta}^k, k = 1, \ldots, N$. Η εκ των υστέρων κατανομής έχει ως έννοια τη δέσμευση της κατανομής των $\boldsymbol{\theta}^k, k = 1, \ldots, N$ ως προς τα δεδομένα $(\mathbf{y}, \mathbf{z})$, όπου $\mathbf{z} = [z^{(1)}, z^{(2)} \ldots, z^{(n)}]$, $z^{(i)} \in \{1, 2, \ldots, N\}$, οι διακριτές μεταβλητές που υποδηλώνουν την αρχική κατάτμηση ενός συνόλου παρατηρήσεων $\mathbf{y}$ σε τμήματα $\{\mathbf{y}_k\}_{k=1}^N$. Έτσι, η εκ των υστέρων κατανομή της $\boldsymbol{\theta}^k$ εξαρτάται μόνο από τις παρατηρήσεις $\mathbf{y}_k = \{\mathbf{y}^{(i)} : z^{(i)} = k\}$, αν θεωρήσουμε εκ των προτέρων κατανομή των $\boldsymbol{\theta}^k, k = 1, \ldots, N$ ανεξάρτητη της συγκεκριμένης πραγμάτωσης $\mathbf{y}$. Οι εκ των υστέρων κατανομές όμως είναι περιοριστικές καθώς έχουν κλειστή φόρμα υπολογισμού μόνο για την περίπτωση συζυγών στην πιθανοφάνεια κατανομών. Έτσι, δεν θα χρησιμοποιήσουμε τον όρο αυτόν. Αντίθετα, θα τις θεωρούμε εκ των προτέρων κατανομές, όπου η εκ των προτέρων γνώση μας είναι τα δεδομένα $(\mathbf{y}, \mathbf{z})$.

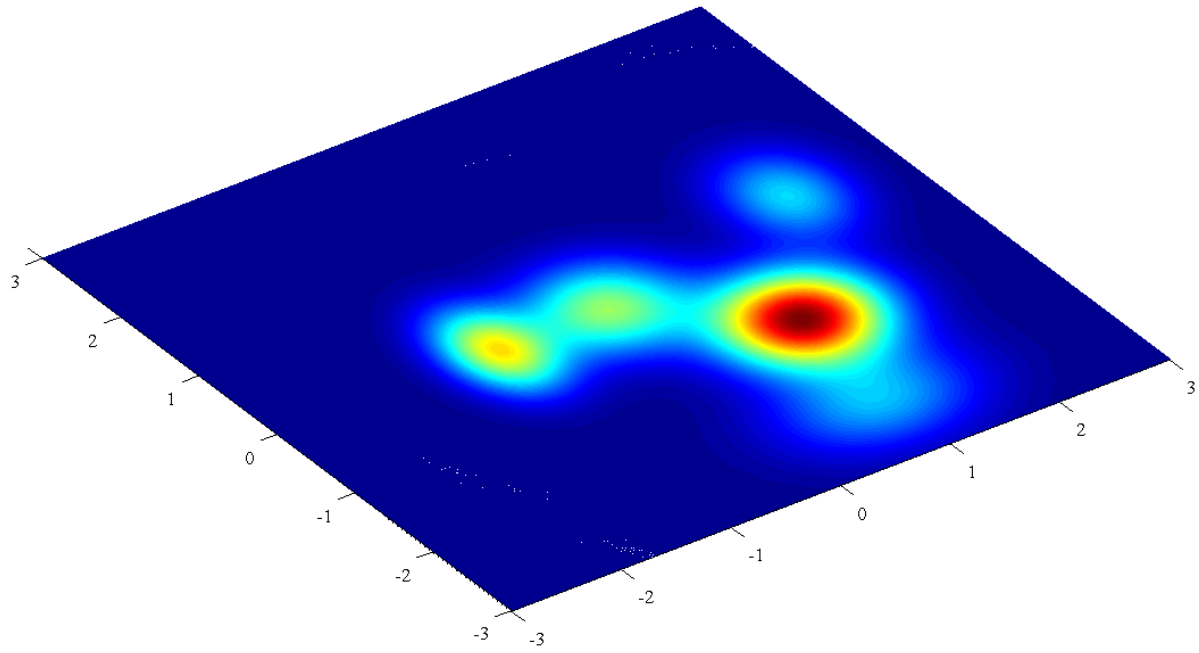
Η μοντελοποίηση λοιπόν είναι μία ευθεία αναγωγή αυτής του αρχικού αλγορίθμου, από τον χώρο των παρατηρήσεων (επίπεδο συνάρτησης εκπομπής) στον χώρο των παραμετρικών μοντέλων (επίπεδο

εκ των υστέρων πιθανότητας). Όπως θα δείξουμε στη συνέχεια, η αναγωγή αυτή έχει αναλυτική λύση χωρίς την ανάγκη προσεγγίσεων ή ευρετικών κανόνων όταν η συνάρτηση πιθανοφάνειας των $\boldsymbol{\theta}^k, k = 1, \ldots, N$ ανήκει στις εκθετικές οικογένειες. Θα χρησιμοποιηθούν επίσης στοιχεία του κλάδου της γεωμετρίας της πληροφορίας (Information Geometry) με βάση την οποία θα προταθούν οι εκ των υστέρων κατανομές που θα αντικαταστήσουν τους πυρήνες στο $\mathbf{y} \in \Re^d$.

Ο προσαρμοσμένος αλγόριθμος θα πρέπει λοιπόν να εκτιμήσει τις κορυφές (τους τρόπους) της εκ των προτέρων κατανομής, καθώς και να αποδώσει τα $\boldsymbol{\theta}^k, k = 1, \ldots, N$ στους κατάλληλους τρόπους. Τα $\boldsymbol{\theta}^k$ τα οποία ελκύονται από τον ίδιο τρόπο ομαδοποιούνται υπό κοινή ομάδα. Ταυτόχρονα, η τιμή του τρόπου αυτού είναι η αναμενόμενη τιμή των παραμέτρων και ως εκ τούτου - όπως και στην ιεραρχική ομαδοποίηση - μπορεί να χρησιμοποιηθεί περαιτέρω, αν ανδιαφερόμαστε για μια εκτίμηση σημείου των παραμέτρων του μοντέλου ομιλητή. Η κεντρική ιδέα αναπαρίσταται στο Σχ. 21 για την απλή περίπτωση μονοδιάστατων παραμέτρων και στο Σχ. 22 για δισδιάστατες παραμέτρους.



Σχῆμα 21: *Παράδειγμα ομαδοποίησης κατανομών, όπου η εκ των υστέρων κατανομή της μονοδιάστατης παραμέτρου είναι κανονική. Παρατηρούμε 6 τμήματα να ομαδοποιούνται σε 4 κλάσεις.*



Σχῆμα 22: *Παράδειγμα ομαδοποίησης κατανομών, όπου η εκ των υστέρων κατανομή των δισδιάστατων παραμέτρων είναι κανονική. Παρατηρούμε 6 τμήματα να ομαδοποιούνται σε 4 κλάσεις. Παρατηρήστε ότι η κάτω δεξιά κατανομή θα ενοποιηθεί με τη μεγαλύτερη καθώς υπάρχει μονοπάτι που εκκινεί από το κέντρο της πρώτης και καταλήγει στο κέντρο της δεύτερης, με μονοτονικά αυξανόμενη εκ των υστέρων πιθανότητα.*

6.3.3 Σχεδιάζοντας τον πυρήνα στον χώρο των παραμέτρων

Ερχόμαστε τώρα να αναλύσουμε τα χαρακτηριστικά που θα πρέπει να έχει ένας πυρήνας στον χώρο των παραμέτρων. Ο πυρήνας - που όπως αναφέραμε δεν είναι τίποτε άλλο από την εκ των υστέρων πιθανότητα των $\boldsymbol{\theta}^k$ - ορίζεται από την απόσταση (πιο σωστά την απόκλιση) από την αναμενόμενη τιμή $\tilde{\boldsymbol{\theta}}^k$, $D_\delta(\boldsymbol{\theta}||\tilde{\boldsymbol{\theta}}^k)$, το σχήμα του και την ανοχή του (strength). Η ανοχή είναι η αντίστοιχη ποσότητα με το (αντίστροφο) εύρος του πυρήνα.

Η οικογένεια των δ-αποκλίσεων

Ας επαναλάβουμε τις βασικές εκφράσεις που θα χρησιμοποιήσουμε στο κεφάλαιο. Η οικογένεια των δ-αποκλίσεων ορίζεται ως εξής

$$D_\delta(p||q) = \begin{cases} \frac{\int p d\mathbf{y}}{1-\delta} + \frac{\int q d\mathbf{y}}{\delta} - \frac{\int p^\delta q^{1-\delta} d\mathbf{y}}{\delta(1-\delta)}, \text{ εάν } \delta \in (0,1) \\ \int q - p + p\log(\frac{p}{q}) d\mathbf{y}, \text{ εάν } \delta = 1 \\ \int p - q + q\log(\frac{q}{p}) d\mathbf{y}, \text{ εάν } \delta = 0 \end{cases} \tag{130}$$

όπου για την εξαγωγή των εκφράσεων για $\delta \in \{0,1\}$ χρησιμοποιούμε το όριο $\frac{p^\delta}{\delta} \to \log p$, $\delta \to 0$. Για $\delta = 1$ έχουμε την Kullback-Leibler απόκλιση, ενώ για $\delta = 0$ έχουμε και αντίστροφη Kullback-Leibler απόκλιση, δηλαδή με αντίστροφα ορίσματα. Η μόνη συμμετρική απόκλιση είναι η περίπτωση $\delta = \frac{1}{2}$, όπου

$$D_{\frac{1}{2}}(p||q) = 2\int \left(\sqrt{p} - \sqrt{q}\right)^2 d\mathbf{y} \tag{131}$$

και είναι η τετραγωνική απόσταση Hellinger επί δύο. Από τις παραπάνω αποκλίσεις θα εξετάσσουμε τις Kullback-Leibler ($\delta = \{0,1\}$) καθώς και την Hellinger ($\delta = \frac{1}{2}$).

Οι κατανομές στον χώρο των παραμέτρων

Έχοντας ορίσει την οικογένεια των στατιστικών αποκλίσεων, εξετάζουμε τώρα το σχήμα του στον παραμετρικό χώρο. Υπευθυμίζουμε εδώ ότι με όρους Μπεϋσιανής στατιστικής, το σχήμα του πυρήνα δεν είναι άλλο από την εκ των υστέρων σ.π.π., δεδομένης της απόκλισης που χρησιμοποιούμε, αφού η τελευταία εξετάσθηκε στην προηγούμενη παράγραφο. Για το σκοπό αυτό, θα χρησιμοποιήσουμε και πάλι στοιχεία από τη θεωρία της Γεωμετρίας της Πληροφορίας. Θα δείξουμε ότι η οικογένεια των σ.π.π. προκύπτει από ελαχιστοποίηση μίας συνάρτησης κόστους, ενώ οι παράμετροι που ρυθμίζονται από τον χρήστη έχουν και αυτές σαφή γεωμετρική ερμηνεία.
Παραθέτουμε τη μορφή της σ.π.π., η οποία έχει ως ακολούθως

$$\Pi^\alpha_{\delta,\nu}(p_\theta; t_0) \propto \begin{cases} |\mathcal{I}(\boldsymbol{\theta})|^{1/2}[1 + \alpha\nu D_\delta(p_\theta||t_0)]^{-\frac{1}{\nu}}, \text{ εάν } \nu \in (0,1] \\ |\mathcal{I}(\boldsymbol{\theta})|^{1/2} e^{-\alpha D_\delta(p_\theta||t_0)}, \text{ εάν } \nu = 0 \end{cases} \tag{132}$$

οι οποίες θα δίνουν και το σχήμα των πυρήνων στον παραμετρικό πλέον χώρο ($\boldsymbol{\theta}$ ή $\boldsymbol{\eta}$). Η εξαγωγή των παραπάνω κατανομών έχει δοθεί στην παράγραφο 3.5. Υπενθυμίζεται ότι προκύπτει από την ελαχιστοποίηση μίας συνάρτησης κόστους παρόμοιας με αυτή των μοντέλων μέγιστης εντροπίας και παραμετροποιείται από τις σταθερές (δ, ν).

6.3.4 Εξομάλυνση των κατανομών - Συνέλιξη και MAP εκτίμηση

Ας εξετάσσουμε λοιπόν την έως τώρα εκτίμηση της κατανομής. Θεωρούμε $\{\hat{\boldsymbol{\theta}}^k\}_{k=1}^N$, τις ML-εκτιμήσεις των παραμέτρων με τις οποίες εκφράζουμε τα τμήματα ομιλίας. Η έως τώρα κατανομή είναι λοιπόν ένα μίγμα κατανομών στον χώρο Θ, που δίνεται από την ακόλουθη σχέση

$$\hat{f}_{\delta,0}(\boldsymbol{\theta}) \propto |\mathcal{I}_\theta(\boldsymbol{\theta})|^{1/2} \sum_{k=1}^{N} w_k \exp\left(-\alpha n_k D_\delta(\boldsymbol{\theta}||\hat{\boldsymbol{\theta}}^k)\right) \tag{133}$$

όπου $w_k = \frac{n_k}{n}$. Σημειώνουμε ότι η κανονικοποίηση της (133) δεν απαιτείται, καθώς στη λογαριθμική κλίμακα μετατρέπεται σε μία προσθετική σταθερά η οποία θα μηδενισθεί όταν υπολογίσουμε την παράγωγο.

Πρέπει να τονισθεί, ωστόσο, ότι το παραπάνω μίγμα αποτελεί απλώς το αντίστοιχο της εμπειρικής κατανομής (άθροισμα παλμών Dirac) του αρχικού αλγορίθμου και όχι της εκτίμησης $\hat{\nabla} f_{h,K}$, (120). Το γεγονός ότι οι παλμοί του αρχικού αλγορίθμου αντικαταστάθηκαν από κατανομές δεν ήταν το αποτέλεσμα της εξομάλυνσης, αλλά σχετίζεται με την αμφιβολία μας σε σχέση με την εκτίμηση των παραμέτρων. Άλλωστε, καθώς το πλήθος των δειγμάτων ανά τμήμα απειρίζεται, οι εώς τώρα κατανομές μας τείνουν προς παλμούς Dirac, αφού η αμφιβολία της εκτίμησης μηδενίζεται ασυμπτωτικά. Είδαμε λοιπόν ότι ο αρχικός αλγόριθμος εξομαλύνει την εμπειρική κατανομή (συνελίσσοντάς τη με έναν πυρήνα) έτσι ώστε να εξάγει την εκτίμηση της πραγματικής κατανομής, $\hat{\nabla} f_{h,K}$. Θα πρέπει λοιπόν και στη μέθοδό μας να προβούμε σε μία αντίστοιχη εξομάλυνση, παρά το γεγονός ότι η κατανομή μας δεν αποτελείται από παλμούς Dirac, δηλαδή από διακριτές μάζες πιθανότητας στον χώρο Θ.

Η εξομάλυνση που προτείνουμε έχει διττό σκοπό. Ας εστιάζουμε αρχικά στο σκέλος που σχετίζεται με το εύρος (bandwidth) των κατανομών. Δείξαμε ότι στις προτεινόμενες κατανομές, το

εύρος είναι αντιστρόφως ανάλογο του αριθμού των δειγμάτων για κάθε τμήμα ομιλίας, μέσω της μεταβλητής γ_e, η οποία εκφράζει την εμπιστοσύνη μας σε σχέση με την εκτίμηση των παραμέτρων της κατανομής. Επιθυμούμε τώρα να αυξήσουμε το εύρος αυτό (τη μεταβλητότητα της κατανομής δηλαδή) με μία προσθετική ποσότητα, κοινή σε κάθε κατανομή και ανεξάρτητη του αριθμού των δειγμάτων κάθε τμήματος. Η πράξη αυτή είναι σε αντιστοιχία με τον αρχικό αλγόριθμο, στον οποίο εξομαλύνουμε την εμπειρική κατανομή, αυξάνοντας το εύρος της με μία σταθερή ποσότητα. Η διαφορά έγκειται στο ότι η εμπειρική κατανομή αποτελείται από παλμούς Dirac και έτσι η μεταβλητότητα είναι μηδενική, κάτι που δεν συμβαίνει στην περίπτωση μας. Ας παραθέσουμε λοιπόν την εκτίμηση της κατανομής μετά την εξομάλυνση.

$$\tilde{f}_{\delta,0}(\boldsymbol{\theta}) \propto |\mathcal{I}_\theta(\boldsymbol{\theta})|^{1/2} \sum_{k=1}^{N} w_k \exp\left(-\alpha n_k^* D_\delta(\boldsymbol{\theta}||\tilde{\boldsymbol{\theta}}_\delta^k)\right) \tag{134}$$

Συγκρίνοντας με τις σχέσεις (133) και (134) βλέπουμε τις δύο μεθόδους εξομάλυνσης. Η πρώτη, στην οποία αναφερθήκαμε ήδη, είναι η αντικατάσταση του πλήθους των δειγμάτων n_k με το $\tilde{n}_k$, το οποίο δίνεται από τον ακόλουθο τύπο

$$(\alpha n_k^*)^{-1} = (\alpha n_k)^{-1} + \sigma_\theta^2 \Rightarrow n_k^* = \frac{n_k \times 1/\sigma_\theta^2}{\alpha n_k + 1/\sigma_\theta^2} \tag{135}$$

Έτσι, ισχύει ότι $\alpha n_k^* < 1/\sigma_\theta^2$, γεγονός που σημαίνει ότι η ακρίβεια κάθε κατανομής ουδέποτε απειρίζεται, αλλά έχει ως άνω όριο την ακρίβεια της εξομάλυνσης $1/\sigma_\theta^2$.

Η δεύτερη μέθοδος εξομάλυνσης που προτείνουμε είναι η αντικατάσταση της ML-εκτίμησης $\hat{\boldsymbol{\theta}}^k$ με την $\tilde{\boldsymbol{\theta}}_\delta^k$. Η εκτίμηση αυτή είναι η γενικευμένη εκ των υστέρων εκτίμηση (δ- MAP), η οποία ταυτίζεται με τη MAP για $\delta = 0$, δηλαδή για συζυγείς εκ των προτέρων κατανομές. Ως εκ τούτου, η $\tilde{\boldsymbol{\theta}}_0^k$ προκύπτει με την προσθήκη στις πραγματικές παρατηρήσεις $\mathbf{y}_k$ ενός πλήθους εικονικών παρατηρήσεων, οι οποίες αντιστοιχούν στις υπερπαραμέτρους της εκ των προτέρων κατανομής. Αυτό σημαίνει ότι για $\delta = 0$ η άθροιση πραγματοποιείται στον χώρο των αναμενόμενων παραμέτρων, δηλαδή

$$\tilde{\boldsymbol{\eta}}_0^k = \frac{n_k}{n_k + n_0}\hat{\boldsymbol{\eta}}^k + \frac{n_0}{n_k + n_0}\boldsymbol{\eta}^0 \tag{136}$$

Αντίστοιχα, για $\delta = 1$, η άθροιση θα πραγματοποιείται στον χώρο των φυσικών παραμέτρων, δηλαδή

$$\tilde{\boldsymbol{\theta}}_1^k = \frac{n_k}{n_k + n_0}\hat{\boldsymbol{\theta}}^k + \frac{n_0}{n_k + n_0}\boldsymbol{\theta}^0 \tag{137}$$

Η αξία της χρήσης της δ- MAP εκτίμησης $\tilde{\boldsymbol{\theta}}_\delta^k$ ωστόσο θα φανεί αν διαχωρίσουμε την εκτίμηση της $\tilde{f}_{\delta,\nu}(\boldsymbol{\theta})$ με την εκτίμηση των παραμέτρων κάθε τμήματος ομιλίας που χρησιμοποιείται ως σημείο εκκίνησης κάθε αναδρομής. Στη δεύτερη περίπτωση, προτείνουμε τη χρήση της εκτίμησης ML, $\hat{\boldsymbol{\theta}}^k$. Ο διαχωρισμός αυτός μας επιτρέπει να αποτυπώσουμε την αμφιβολία απέναντι στην εκτίμηση των παραμέτρων στον αλγόριθμό μας. Πιο συγκεκριμένα, ας θεωρήσουμε δύο τμήματα ομιλίας k και l, με αριθμό παρατηρήσεων n_k και n_l, όπου $n_k > n_l$ και έστω ότι ισχύει $\hat{\boldsymbol{\theta}}^k = \hat{\boldsymbol{\theta}}^l$. Με τον διαχωρισμό των δύο εκτιμήσεων, η αναδρομή στην περίπτωση του k-οστού τμήματος θα αρχίζει πιο κοντά (με όρους δ-γεωμετρίας) στην $\tilde{\boldsymbol{\theta}}_\delta^k$ σε σχέση με την αντίστοιχη απόσταση της $\hat{\boldsymbol{\theta}}^l$ από την $\tilde{\boldsymbol{\theta}}_\delta^l$. Ως εκ τούτου, μικρά τμήματα ομιλίας θα έχουν μεγαλύτερη δυνατότητα διαφυγής από τη δεξαμενή έλξης που δημιουργούν (self-attraction) οδηγώντας σε πιο συμπαγείς ομαδοποιήσεις. Αντίθετα, για μεγάλα τμήματα ομιλίας θα έχουμε $\tilde{\boldsymbol{\theta}}_\delta^k \approx \hat{\boldsymbol{\theta}}^k$, $\forall \delta \in [0,1]$, και ως εκ τούτου θα έχουν μικρότερες πιθανότητες διαφυγής από τη δεξαμενή έλξης $\tilde{\boldsymbol{\theta}}_\delta^k$. Αναφέρουμε, τέλος, ότι η διάκριση αυτή είναι απολύτως συμβατή με την απαίτησή μας για ταύτιση του αλγορίθμου μας με τον αρχικό αλγόριθμο όταν $\{n_k\}_{k=1}^N \to \infty$, αφού $\lim_{n_k\to\infty} \tilde{\boldsymbol{\theta}}_\delta^k = \hat{\boldsymbol{\theta}}^k$. Σημειώνουμε ότι η χρήση της MAP εκτίμησης θα πρέπει να συνοδευθεί και με την τροποποίηση του $\tilde{n_k}$ που δίνεται από τη σχέση (135) ως εξής

$$n_k^* = \frac{(n_k + n_0) \times 1/\sigma_\theta^2}{\alpha(n_k + n_0) + 1/\sigma_\theta^2} \tag{138}$$

ώστε να συμπεριλάβουμε και το πλήθος των εικονικών δειγμάτων n_0. Επιπλέον, θέτουμε την παράμετρο α ίση με τη μονάδα καθώς η χρήση της εξομάλυνσης την καθιστά περιττή.
Τέλος, τα βάρη w_k θα πρέπει να υποστούν και αυτά αναθεώρηση, αρχικά λόγω της MAP εκτίμησης ως

$$\tilde{w}_k = \frac{n_k + n_0}{n + N n_0} \tag{139}$$

δηλαδή την MAP εκτίμηση με χρήση Dirichlet εκ των προτέρων κατανομών. Μία δεύτερη δυνατότητα είναι η χρήση των εξομαλυσμένων βαρών, που τα ορίζουμε ως

$$w_k^* = \frac{n_k^*}{\sum_{k'=1}^N n_{k'}^*} \tag{140}$$

Όπως θα δείξουμε στη συνέχεια, η χρήση των w_k^* οδήγησε σε καλύτερα αποτελέσματα σε σχέση με τα $\tilde{w}_k$.

Πίνακας 3: Απλή και φυσική κλίση της απόκλισης Kullback-Leibler.

Τύπος κλίσης	Απλή κλίση $\nabla_{\boldsymbol{\varphi}} D_\delta(p\|\|q)\big\|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_p}$		Φυσική κλίση $\tilde{\nabla}_{\boldsymbol{\varphi}} D_\delta(p\|\|q)\big\|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_p}$	
Τύπος παραμέτρων	$\boldsymbol{\varphi}=\boldsymbol{\theta}$	$\boldsymbol{\varphi}=\boldsymbol{\eta}$	$\boldsymbol{\varphi}=\boldsymbol{\theta}$	$\boldsymbol{\varphi}=\boldsymbol{\eta}$
$\delta=1$	$\mathcal{I}(\boldsymbol{\theta}_p)(\boldsymbol{\theta}_p-\boldsymbol{\theta}_q)$	$\boldsymbol{\theta}_p-\boldsymbol{\theta}_q$	$\boldsymbol{\theta}_p-\boldsymbol{\theta}_q$	$\mathcal{I}(\boldsymbol{\theta}_p)(\boldsymbol{\theta}_p-\boldsymbol{\theta}_q)$
$\delta=0$	$\boldsymbol{\eta}_p-\boldsymbol{\eta}_q$	$\mathcal{I}(\boldsymbol{\eta}_p)(\boldsymbol{\eta}_p-\boldsymbol{\eta}_q)$	$\mathcal{I}(\boldsymbol{\eta}_p)(\boldsymbol{\eta}_p-\boldsymbol{\eta}_q)$	$\boldsymbol{\eta}_p-\boldsymbol{\eta}_q$

6.3.5 Διαφόριση της απόκλισης και η φυσική κλίση

Ερχόμαστε τώρα να δούμε που οδηγεί η διαφόριση ως προς $\boldsymbol{\theta}$ του πυρήνα. Χρήσιμο είναι εδώ να θυμίσουμε το πώς προκύπτει το διάνυσμα μετατόπισης του μέσου στον αρχικό αλγόριθμο. Όπως δείχνει η (124), η αναδρομική σχέση προκύπτει από τη διαφόριση ενός πυρήνα που περιέχει τετραγωνικές αποστάσεις και σύμφωνα με την (245), οι δ-αποκλίσεις αντιστοιχούν σε τετραγωνικές αποστάσεις, επίσης. Επομένως, με την κατάλληλη διαφόριση περιμένουμε να καταλήξουμε σε μία ανάλογη σχέση στον παραμετρικό χώρο που εξετάζουμε.

Δεδομένου ότι βρισκόμαστε σε έναν μη-ευκλείδιο χώρο του οποίου γνωρίζουμε τη δομή του, η απλή (κατευθυντική) κλίση $\nabla_{\boldsymbol{\theta}}$ είναι υποβέλτιστη επιλογή, καθώς υπονοεί Ευκλείδια γεωμετρία. Αντίθετα, η διαφόριση σε περιπτώσεις όπου ο τανυστής μετρικής είναι γνωστός, θα πρέπει να πραγματοποιείται με χρήση της φυσικής κλίσης (Natural Gradient) $\tilde{\nabla}_{\boldsymbol{\theta}} = \mathcal{I}(\boldsymbol{\theta})^{-1}\nabla_{\boldsymbol{\theta}}$. Το ακόλουθο παράδειγμα επαληθεύει τον ισχυρισμό αυτόν.

Έστω ότι επιθυμούμε να δημιουργήσουμε έναν αλγόριθμο απότομης καθόδου (steepest descend) για την ελαχιστοποίηση της απόκλισης Kullback-Leibler, δηλαδή για $\delta \in \{0,1\}$ μεταξύ της $p \in \mathcal{P}$ και της $q \in \mathcal{P}$. Χωρίς βλάβη της γενικότητας θεωρούμε την q σαν δεδομένη σταθερά και την p ως τη μεταβλητή μας. Ο παρακάτω πίνακας περιέχει τους συνδυασμούς όσον αφορά στην κλίση που λαμβάνουμε (απλή ή φυσική), τη φορά της απόκλισης Kullback-Leibler ($\delta = 1$ για την ευθεία, $\delta = 0$ για την αντίστροφη) καθώς και τις παραμέτρους ως προς τις οποίες διαφορίζουμε ($\boldsymbol{\theta}$ για τις φυσικές παραμέτρους, $\boldsymbol{\eta}$ για τις αναμενόμενες παραμέτρους). Ο πίνακας 3 είναι ιδιαίτερα χρήσιμος για την κατανόηση της φυσικής κλίσης. Έστω, λοιπόν ο αλγόριθμος ελαχιστοποίησης

της απόκλισης Kullback-Leibler, η αναδρομή του οποίου ορίζεται ως εξής

$$\boldsymbol{\varphi}_p^{t+1} \leftarrow \boldsymbol{\varphi}_p^t - \mu^t \tilde{\nabla}_{\boldsymbol{\varphi}} D_\delta(p||q)\big|_{\boldsymbol{\varphi}=\boldsymbol{\varphi}_p} \tag{141}$$

για τη φυσική κλίση - ενώ για την απλή κλίση αντικαθιστούμε το $\tilde{\nabla}_{\boldsymbol{\varphi}}$ με $\nabla_{\boldsymbol{\varphi}}$. Το $\boldsymbol{\varphi}$ συμβολίζει $\boldsymbol{\theta}$ ή $\boldsymbol{\eta}$ ενώ το μ^t είναι το βήμα του αλγορίθμου, το οποίο είναι γενικά μια φθίνουσα συνάρτηση του αύξοντα αριθμού της αναδρομής, t. Υπευθυμίζουμε επίσης ότι $\mathcal{I}(\boldsymbol{\theta}_p)(\boldsymbol{\theta}_p - \boldsymbol{\theta}_q) \approx \boldsymbol{\eta}_p - \boldsymbol{\eta}_q$ για $\boldsymbol{\theta}_q$ κοντά στο $\boldsymbol{\theta}_p$, αφού ο $\mathcal{I}(\boldsymbol{\theta})$ είναι η Ιακωβιανή (Jacobian) του μετασχηματισμού $\boldsymbol{\theta} \mapsto \boldsymbol{\eta}$.

Όπως φαίνεται από τον πίνακα, η χρήση της απλής κλίσης είναι υποβέλτιστη, καθώς αντιστρέφει την παραμετροποίηση από $\boldsymbol{\theta}$ σε $\boldsymbol{\eta}$ και αντίστροφα. Αντίθετα, η φυσική κλίση οδηγεί σε έναν αλγόριθμο όπου μία και μόνη αναδρομή αρκεί για να συγκλίνει, αν χρησιμοποιηθεί η κατάλληλη απόκλιση Kullback-Leibler.

Στην περίπτωσή μας, λοιπόν, όπως έχουμε ένα μίγμα N κατανομών στον χώρο Θ, η αντικειμενική συνάρτηση θα πρέπει να ισούται με τον λογάριθμο της κατανομής. Αυτό σε πλήρη αντιστοιχία με τη χρήση κανονικού πυρήνα στον αρχικό αλγόριθμο. Για $N = 1$ και χρήση εκθετικού πυρήνα ($\nu = 0$), ο αλγόριθμος θα εκφυλίζεται σε παραγώγιση της απόκλισης, όπως δηλαδή στις παραπάνω σχέσεις, η οποία και συγκλίνει σε ένα μόνο βήμα για $\delta \in \{0, 1\}$. Μία πιο λεπτομερής ανάλυση της φυσικής κλίσης παρουσιάζεται στο Παράρτημα 9.5.

6.3.6 Το μέτρο αναφοράς των κατανομών

Ένα δεύτερο χαρακτηριστικό που χρήζει προσοχής είναι το μέτρο αναφοράς των κατανομών. Ο ορισμός της εκ των προτέρων κατανομής, στον τυχαίο παραμετρικό χώρο $\boldsymbol{\theta} \in \Theta$ δόθηκε ως

$$\hat{f}_{\delta,0}(\boldsymbol{\theta}) \propto |\mathcal{I}_\theta(\boldsymbol{\theta})|^{1/2} \sum_{k=1}^{N} w_k^* \exp\left(-\alpha n_k^* D_\delta(\boldsymbol{\theta}||\hat{\boldsymbol{\theta}}^k)\right) \tag{142}$$

Θα πρέπει να τονίσουμε ότι η παραπάνω έκφραση είναι η σ.π.π. με μέτρο αναφοράς το μέτρο Lebesgue, $d\boldsymbol{\theta} = d\theta_1 d\theta_2, \ldots, d\theta_P$. Ζητούμε λοιπόν να βρούμε σημεία εκείνα στον χώρο Θ που η $\hat{f}_{\delta,\nu}(\boldsymbol{\theta})$ παρουσιάζει μέγιστο. Το μέτρο Lebesgue όμως είναι ακατάλληλο για μια τέτοια διαδικασία, αφού (α) εξαρτάται από την παραμετροποίηση και (β) αγνοεί τον τανυστή μετρικής που εισάγει η

χρήση των δ-αποκλίσεων. Η μεγιστοποίηση λοιπόν της μάζας πιθανότητας που βρίσκεται στον απειροστό όγκο $d\boldsymbol{\theta}$ δεν μπορεί να είναι το κριτήριό μας.

Το πρόβλημα θα πρέπει λοιπόν να επαναδιατυπωθεί ως ακολούθως. Επιθυμούμε να βρούμε τα σημεία εκείνα στον χώρο Θ, όπου η απειροστή μάζα πιθανότητας $\hat{f}_{\delta,\nu}(\boldsymbol{\theta})d\boldsymbol{\theta} = \tilde{f}_{\delta,\nu}(\boldsymbol{\theta})d\mathcal{V}$ μεγιστοποιείται, όπου $\tilde{f}_{\delta,\nu}(\boldsymbol{\theta})$ η σ.π.π. ως προς το κατάληλλο μέτρο αναφοράς. Το μέτρο αυτό δεν είναι άλλο από την κατανομή του Jeffreys, έχουμε δηλαδή

$$d\mathcal{V} = |\mathcal{I}_{\theta}(\boldsymbol{\theta})|^{1/2} d\theta_1 d\theta_2, \ldots, d\theta_P \tag{143}$$

και

$$\tilde{f}_{\delta,\nu}(\boldsymbol{\theta}) = \frac{dF_{\delta,\nu}(\boldsymbol{\theta})}{d\mathcal{V}} \tag{144}$$

Μία διαισθητική εξήγηση για τη χρήση του συγκεκριμένου μέτρου αναφοράς προκύπτει με βάση τη Μπεϋσιανή στατιστική. Όπως αναφέραμε, ο όρος $|\mathcal{I}_{\boldsymbol{\theta}}(\boldsymbol{\theta})|^{1/2}$ είναι ο μη-κανονικοποιημένος Jeffreys prior και εκφράζει την ομοιόμορφη κατανομή στον χώρο των παραμετρικών μοντέλων. Η όποια παραμετροποίηση επιλέγεται, χρησιμεύει στο να εκφράσουμε μαθηματικά τα μοντέλα αυτά, να υπολογίσουμε αποκλίσεις μεταξύ τους κ.ο.κ.. Η ομοιομορφία αυτή ωστόσο δεν διατηρείται στην τυχαία επιλεγείσα παραμετροποίηση Θ, εάν δεν αναθεωρήσουμε τον τρόπο με τον οποίον παραγωγίζουμε τα μεγέθη μας. Έτσι, εάν ορίζαμε ως αρχική θέση ενός αναδρομικού αλγορίθμου το $p_0 \in \mathcal{P}$ και δεν διαθέταμε καμία εκ των προτέρων γνώση, θα επιθυμούσαμε ο αλγόριθμος να τερμάτιζε στην ίδια θέση p_0, απλούστατα γιατί η παράγωγος στον αφηρημένο χώρο των παραμετρικών μοντέλων πρέπει να είναι μηδενική, ώστε να ανταναικλά την ομοιόμορφη κατανομή στο χώρο αυτόν. Είτε λοιπόν θα πρέπει να αφήσουμε την παραμετροποίηση ως έχει και εργαστούμε με τη συμμεταβλητή παράγωγο, είτε θα πρέπει ορίσουμε την σ.π.π. με βάση το πραγματικό μέτρο ομοιομορφίας, για το οποίο θα ισχύει ότι $\tilde{f}_{\delta,\nu}(\boldsymbol{\theta}) \propto 1$, απουσία εκ των προτέρων γνώσης.

Τροποποιούμε λοιπόν την έκφραση της σ.π.π. που αντιστοιχεί στην κατανομή

$$\tilde{f}_{\delta,1}(\boldsymbol{\theta}) \propto \sum_{k=1}^{N} w_k^* \exp\left(-\alpha n_k^* D_{\delta}(\boldsymbol{\theta}||\tilde{\boldsymbol{\theta}}^k)\right) \tag{145}$$

δηλαδή με το κατάλληλο μέτρο αναφοράς, τον Jeffreys prior.

Τονίζουμε μόνο ότι η σχέση (145) είναι πλέον σχέση μεταξύ τανυστών και συγκεκριμένα βαθμωτών

πεδίων (scalar fields). Πράγματι, το δεξί σκέλος της (145) είναι ανεξάρτητο της παραμετροποίησης και ο κανόνας μετασχηματισμού του είναι απλή αντικατάσταση χωρίς τη χρήση της Ιακωβιανής. Για βαθμωτά, όμως, πεδία, η συμμεταβλητή παράγωγος ταυτίζεται με τη μερική παράγωγο, γεγονός που οδηγεί στους παρακάτω αναδρομικούς τύπους.

6.4 Οι εκφράσεις των αναδρομικών τύπων

Για να εξάγουμε λοιπόν τον αναδρομικό τύπο θα πρέπει να διαφορίσουμε - με χρήση της φυσικής κλίσης - τον πυρήνα της (132) και να θέσουμε την ποσότητα ίση με μηδέν.

6.4.1 Γεωμετρίες $(\delta, \nu) = (\{0, 1\}, 0)$

Θα προχωρήσουμε αρχικά με την ειδική περίπτωση όπου $(\delta, \nu) = (0, 0)$ και στη συνέχεια θα εξετάσουμε την πιο γενική περίπτωση. Θέτοντας $\alpha = 1$, έχουμε ότι η διαφόριση ως προς τις αναμενόμενες παραμέτρους

$$\tilde{\nabla}_{\boldsymbol{\eta}} \sum_{k=1}^{N} w_k^* \exp\left(-n_k^* D_0(p_\eta || p_{\eta^k})\right) = \mathbf{0} \tag{146}$$

οδηγεί στην παρακάτω αναδρομή

$$\hat{\boldsymbol{\eta}}^{t+1} \leftarrow \frac{\sum_{k=1}^{N} w_k^* n_k^* \exp\left(-n_k^* D_0(p_{\hat{\eta}^t} || p_{\tilde{\eta}^k})\right) \tilde{\boldsymbol{\eta}}^k}{\sum_{k=1}^{N} w_k^* n_k^* \exp\left(-n_k^* D_0(p_{\hat{\eta}^t} || p_{\tilde{\eta}^k})\right)} \tag{147}$$

Τονίζουμε ότι ταυτόσημα αποτελέσματα θα λαμβάναμε αντί της χρήσης της φυσικής απόκλισης, διαφορίζαμε ως προς $\boldsymbol{\theta}$ και θέταμε το διάνυσμα ίσο με $\mathbf{0}$.

Βλέπουμε επομένως ότι η περίπτωση της συζυγούς εκ των προτέρων κατανομής οδηγεί σε κλειστό τύπο υπολογισμού στο σύστημα συντεταγμένων των αναμενόμενων παραμέτρων, όπως δείχνει και ο Πίνακας 6.3.5.

Εντελώς συμμετρικά, η επιλογή $(\delta, \nu) = (1, 0)$, οδηγεί στον ακόλουθο αναδρομικό τύπο

$$\hat{\boldsymbol{\theta}}^{t+1} \leftarrow \frac{\sum_{k=1}^{N} w_k^* n_k^* \exp\left(-n_k^* D_1(p_{\hat{\boldsymbol{\theta}}^t} || p_{\tilde{\boldsymbol{\theta}}^k})\right) \tilde{\boldsymbol{\theta}}^k}{\sum_{k=1}^{N} w_k^* n_k^* \exp\left(-n_k^* D_1(p_{\hat{\boldsymbol{\theta}}^t} || p_{\tilde{\boldsymbol{\theta}}^k})\right)} \tag{148}$$

είναι επομένως ένας μέσος όρος στον χώρο Θ, δηλαδή τον χώρο των φυσικών παραμέτρων. Τονίζουμε ότι η διαφόριση ως προς $\boldsymbol{\theta}$ ή $\boldsymbol{\eta}$ οδηγεί στον ίδιο αναδρομικό τύπο, κάνοντας χρήση της φυσικής κλίσης όπου χρειάζεται.

6.4.2 Γεωμετρίες $(\delta, \nu) = (\{0, 1\}, (0, 1])$

Όπως δείξαμε παραπάνω, η παράμετρος $\nu \in [0, 1]$ ορίζει το σχήμα της κατανομής, με $\nu = 0$ να αντιστοιχεί στην εκθετική κατανομή. Εάν αντίθετα θέσουμε $\nu \in (0, 1]$ θα λάβουμε ένα μίγμα κατανομών πολυωνυμικού βαθμού

$$\tilde{f}_{\delta,\nu}(\boldsymbol{\theta}) \propto \sum_{k=1}^{N} w_k^* [1 + n_k^* \nu D_1(p_\theta || p_{\theta^k})]^{-\frac{1}{\nu}} \tag{149}$$

Για να εξάγουμε τον αναδρομικό τύπο, θέτουμε και πάλι την παράγωγο ίση με το μηδενικό διάνυσμα, και θεωρώντας $\delta = 1$

$$\tilde{\nabla}_{\boldsymbol{\theta}} \sum_{k=1}^{N} w_k [1 + n_k^* \nu D_1(p_\theta || p_{\theta^k})]^{-\frac{1}{\nu}} = \mathbf{0} \tag{150}$$

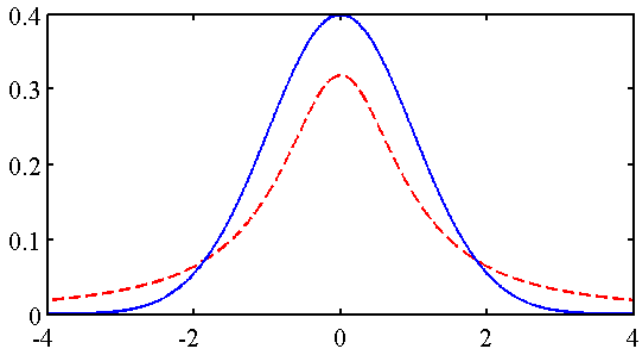
οδηγούμαστε στην παρακάτω αναδρομή

$$\hat{\boldsymbol{\theta}}^{t+1} \leftarrow \frac{\sum_{k=1}^{N} w_k^* n_k^* \nu [1 + n_k^* \nu D_1(p_{\hat{\theta}^t} || p_{\theta^k})]^{-\frac{\nu+1}{\nu}} \tilde{\boldsymbol{\theta}}^k}{\sum_{k=1}^{N} w_k^* n_k^* \nu [1 + n_k^* \nu D_1(p_{\hat{\theta}^t} || p_{\theta^k})]^{-\frac{\nu+1}{\nu}}} \tag{151}$$

Ομοίως, για $\delta = 0$ θα λάβουμε

$$\hat{\boldsymbol{\eta}}^{t+1} \leftarrow \frac{\sum_{k=1}^{N} w_k^* n_k^* \nu [1 + n_k^* \nu D_0(p_{\hat{\eta}^t} || p_{\hat{\eta}^t})]^{-\frac{\nu+1}{\nu}} \tilde{\boldsymbol{\eta}}^k}{\sum_{k=1}^{N} w_k^* n_k^* \nu [1 + n_k^* \nu D_0(p_{\hat{\eta}^t} || p_{\eta^k})]^{-\frac{\nu+1}{\nu}}} \tag{152}$$

Σε σχέση με τις εκθετικές κατανομές, οι πολυωνυμικές διαθέτουν πιο μεγάλες ουρές (heavy-tailed priors) και επομένως είναι πιο στιβαρές (robust) σε ασυνήθιστες τιμές (outliers). Για το πρόβλημα της εκτίμησης θέσης (location) γνωστής μεταβλητότητας, οι αντίστοιχες κατανομές (κανονική και Cauchy) σκιαγραφούνται στο Σχ. 23.



Σχῆμα 23: *Κατανομή Cauchy (κόκκινη διακεκομμένη γραμμή) και κανονική κατανομή (μπλέ κανονική γραμμή). Παρατηρείσθε τις έντονες ουρές της κατανομής Cauchy*

6.5 Παρατηρήσεις πάνω στις αναδρομικές σχέσεις

Στο σημείο αυτό θα εστιάσουμε στις αναδρομικές σχέσεις και τις διαφορές τους με τον αρχικό αλγόριθμο. Οι ομοιότητες αρχικά είναι μεταξύ τους εμφανείς. Αν εστιάσσουμε στις αποκλίσεις Kullback-Leibler $\delta = \{0, 1\}$, παρατηρούμε ότι και οι δύο αναδρομικές σχέσεις ορίζουν ως νέα τιμή έναν μέσο όρο του συνόλου των παρατηρήσεων, με βάρη που προσδιορίζονται από τον πυρήνα και την απόσταση της τρέχουσας θέσης από τις παρατηρήσεις. Η διαφορά τους έγκειται στη διαφορετική αφηνική παραμετροποίηση που προκύπτει από για $\delta = \{0, 1\}$. Για $\delta = 0$ - πυρήνας που αντιστοιχεί στις συζυγείς εκ των προτέρων κατανομές - η γραμμική (affine) παραμετροποίηση είναι η $\boldsymbol{\eta}$, δηλαδή οι αναμενόμενες παράμετροι. Αντίθετα, για $\delta = 1$ - πυρήνας που αντιστοιχεί στις εντροπικές εκ των προτέρων κατανομές - η γραμμική παραμετροποίηση θα είναι οι $\boldsymbol{\theta}$, δηλαδή οι φυσικές παράμετροι.

Αξίζει να παρουσιάσουμε επίσης τις σχέσεις που προέκυπταν αν δεν είχαμε ενσωματώσει στον αλγόριθμό μας την απαραίτητη αλλαγή του μέτρο αναφοράς σε Jeffreys. Η αναδρομική σχέση θα πρόεκυπτε ως

$$\hat{\boldsymbol{\theta}}^{t+1} \leftarrow \frac{\sum_{k=1}^{N} w_k^* \exp\left(-n_k^* \bar{D}_1(p_{\hat{\boldsymbol{\theta}}^t} || p_{\tilde{\boldsymbol{\theta}}^k})\right)(n_k^* \tilde{\boldsymbol{\theta}}^k + \nabla_\theta |\mathcal{I}_\theta(\hat{\boldsymbol{\theta}}^t)|^{1/2})}{\sum_{k=1}^{N} w_k^* \alpha_k \exp\left(-n_k^* \bar{D}_1(p_{\hat{\boldsymbol{\theta}}^t} || p_{\tilde{\boldsymbol{\theta}}^k})\right)} \tag{153}$$

Για χώρους μεταβλητής μετρικής ισχύει όμως ότι $\nabla_\theta |\mathcal{I}_\theta(\hat{\boldsymbol{\theta}})|^{1/2} \neq \mathbf{0}$ γενικά και έτσι η αναδρομική σχέση δεν έχει εξασφαλισμένη σύγκλιση $\forall \boldsymbol{\theta} \in \Theta$. Απουσία γνώσης, θα έχουμε $\{\alpha_k\}_k \to 0$ και ως εκ τούτου θα είχαμε

$$\hat{\boldsymbol{\theta}}^{t+1} \propto \nabla_\theta |\mathcal{I}_\theta(\hat{\boldsymbol{\theta}}^t)|^{1/2} \tag{154}$$

το οποίο είτε δεν θα συνέκλινε, είτε θα συνέκλινε στο σύνορο του χώρου $\partial\Theta$.

6.5.1 Η γεωμετρία της Hellinger απόκλισης

Ερχόμαστε τώρα να μελετήσουμε πιο γενικές περιπτώσεις, όπου $(\delta, \nu) = (\{0,1\}, 0)$. Ας ξεκινήσουμε με την περίπτωση $(\delta, \nu) = (\frac{1}{2}, 0)$. Διατηρούμε δηλαδή τη $(\nu = 0)$-γεωμετρία στον χώρο των παραμέτρων (εκθετικές εκ των προτέρων κατανομές) και χρησιμοποιούμε την $(\delta = \frac{1}{2})$- γεωμετρία στον χώρο των παρατηρήσεων.

Η απόκλιση Hellinger είναι ιδιαίτερης σημασίας, καθώς είναι η μόνη συμμετρική $\delta-$απόκλιση. Ορίζεται ως

$$D_{\frac{1}{2}}(p(\mathbf{y})||q(\mathbf{y})) = 4\left(1 - \int_{\mathcal{Y}} \sqrt{p(\mathbf{y})q(\mathbf{y})} d\mathbf{y}\right) \tag{155}$$

και όπως κάθε μέλος των δ-αποκλίσεων, αντιστοιχεί σε τετραγωγική απόσταση στον ευκλείδιο χώρο. Σε αντίθεση με τις $\{0,1\}$-αποκλίσεις, οι $(0,1)$-αποκλίσεις δεν διαθέτουν γραμμικές συντεταγμένες. Για εκθετικές οικογένειες κατανομών, η απόκλιση λαμβάνει την παρακάτω συμπαγή έκφραση

$$D_{\frac{1}{2}}(p(\mathbf{y})||q(\mathbf{y})) = 4\left[1 - \exp\left(\psi(\bar{\boldsymbol{\theta}}) - \frac{\psi(\boldsymbol{\theta}_p) + \psi(\boldsymbol{\theta}_q)}{2}\right)\right] \tag{156}$$

όπου $\bar{\boldsymbol{\theta}} = \frac{\boldsymbol{\theta}_p + \boldsymbol{\theta}_q}{2}$. Λόγω της κυρτότητας της $\psi(\boldsymbol{\theta})$ ως προς $\boldsymbol{\theta}$ ισχύει ότι

$$\psi(\bar{\boldsymbol{\theta}}) - \frac{\psi(\boldsymbol{\theta}_p) + \psi(\boldsymbol{\theta}_q)}{2} \leq 0 \tag{157}$$

με ισότητα μόνο όταν $\boldsymbol{\theta}_p = \boldsymbol{\theta}_q$. Συμπεραίνουμε επίσης ότι $0 \leq D_{\frac{1}{2}}(p(\mathbf{y})||q(\mathbf{y})) \leq 4$, με μέγιστο όταν η ποσότητα στον εκθέτη τείνει στο $-\infty$. Σε αντίθεση λοιπόν με τις αποκλίσεις Kullback-Leibler, η απόκλιση Hellinger έχει άνω φράγμα και ως εκ τούτου δεν μπορεί να οδηγήσει σε κανονικοποιήσιμες κατανομές στον χώρο των παραμέτρων. Παρόλα αυτά, μπορούν να θεωρήσουμε

ένα οσοδήποτε μεγάλο κυρτό (ως προς τις $\frac{1}{2}$-γεωδαιτικές) σύνολο στον χώρο $\Theta^{tr} \subset \Theta$, όπου αναμένουμε τις τιμές των παραμέτρων και να αποδώσουμε μηδενική πυκνότητα πιθανότητας Θ^{tr}. Έτσι, θεωρούμε την τεμαχισμένη (truncated) εκδοχή της απόκλισης Hellinger

$$D^{tr}_{\frac{1}{2}}(p(\mathbf{y})||q(\mathbf{y})) = \begin{cases} D_{\frac{1}{2}}(p(\mathbf{y})||q(\mathbf{y})), & \boldsymbol{\theta}_p, \boldsymbol{\theta}_q \in \Theta^{tr} \\ +\infty, & \text{αλλού} \end{cases} \tag{158}$$

Παραγωγίζοντας την $D_{\frac{1}{2}}(p(\mathbf{y})||q(\mathbf{y}))$ ως προς $\boldsymbol{\theta}_p$ λαμβάνουμε

$$\frac{\partial D_{\frac{1}{2}}(p(\mathbf{y})||q(\mathbf{y}))}{\partial \boldsymbol{\theta}_p} = 2\left(\boldsymbol{\eta}_p - \langle \boldsymbol{\eta} \rangle_{\bar{\boldsymbol{\theta}}}\right) \exp\left(\psi(\bar{\boldsymbol{\theta}}) - \frac{\psi(\boldsymbol{\theta}_p) + \psi(\boldsymbol{\theta}_q)}{2}\right) \tag{159}$$

όπου

$$\langle \boldsymbol{\eta} \rangle_{\bar{\boldsymbol{\theta}}} = \int_{\mathcal{Y}} \mathbf{t}(\mathbf{y}) p(\mathbf{y}|\bar{\boldsymbol{\theta}}) d\mathbf{y} \tag{160}$$

οι αναμενώμενες παράμετροι της κατανομής $\bar{\boldsymbol{\theta}}$. Μία επιπρόσθετη διαφορά της $\delta \in (0,1)$ - γεωμετριών σε σχέση με τις $\delta \in \{0,1\}$ είναι η μη-προσθετική ιδιότητα ως προς τον αριθμό των δειγμάτων. Έτσι, η σ.π.π. για N τμήματα ομιλίας θα έχει την ακόλουθη μορφή

$$\tilde{f}_{\frac{1}{2},0}(\boldsymbol{\theta}) \propto \sum_{k=1}^{N} w_k^* \exp\left(-\alpha D_{\frac{1}{2},\tilde{n}_k}(\boldsymbol{\theta}||\tilde{\boldsymbol{\theta}}^k)\right) \tag{161}$$

Θέτουμε

$$H(\boldsymbol{\theta}, \tilde{\boldsymbol{\theta}}^k) = \exp\left(-\alpha D_{\frac{1}{2},\tilde{n}_k}(\boldsymbol{\theta}||\tilde{\boldsymbol{\theta}}^k)\right) \tag{162}$$

ώστε

$$D_{\frac{1}{2},\tilde{n}_k}(\boldsymbol{\theta}||\tilde{\boldsymbol{\theta}}^k) = 4\left[1 - H(\boldsymbol{\theta}_p, \tilde{\boldsymbol{\theta}}_q)^{\tilde{n}_k}\right] \tag{163}$$

Στις παραπάνω σχέσεις σημειώνουμε $\tilde{n}_k$ το εξομαλυσμένο αριθμό δειγμάτων, δηλαδή $\tilde{n}_k = \frac{n_k n^0}{n_k + n^0}$, όπου n^0 η μέγιστη αντοχή (strength) (για $n_k \to \infty$) που αντιστοιχεί στο αντίστροφο εύρος του πυρήνα του αρχικού αλγορίθμου.

Παραγωγίζοντας λοιπόν το μίγμα της (161) και εξισώνοντάς τη με το μηδενικό διάνυσμα, θα λάβουμε την ακόλουθη αναδρομική σχέση

$$\hat{\boldsymbol{\eta}}^{t+1} \leftarrow \frac{\sum_{k=1}^{N} w_k \exp\left(-\alpha D_{\frac{1}{2},\tilde{n}_k}(\hat{\boldsymbol{\theta}}^t||\tilde{\boldsymbol{\theta}}^k)\right) \frac{\alpha \tilde{n}_k}{2} H(\hat{\boldsymbol{\theta}}^t||\tilde{\boldsymbol{\theta}}^k)^{\tilde{n}_k - 1} \langle \boldsymbol{\eta} \rangle^t_{\tilde{\boldsymbol{\theta}}_k}}{\sum_{k=1}^{N} w_k \exp\left(-\alpha D_{\frac{1}{2},\tilde{n}_k}(\hat{\boldsymbol{\theta}}^t||\tilde{\boldsymbol{\theta}}^k)\right) \frac{\alpha \tilde{n}_k}{2} H(\hat{\boldsymbol{\theta}}^t||\tilde{\boldsymbol{\theta}}^k)^{\tilde{n}_k - 1}} \tag{164}$$

Σημειώνουμε ότι η διαφορική σ.π.π

$$\tilde{g}_{\frac{1}{2},0}(\boldsymbol{\theta}|\tilde{\boldsymbol{\theta}}) \propto \exp\left(-\alpha D_{\frac{1}{2},\tilde{n}_k}(\hat{\boldsymbol{\theta}}^t||\tilde{\boldsymbol{\theta}})\right) \frac{\alpha\tilde{n}_k}{2} H(\hat{\boldsymbol{\theta}}^t||\tilde{\boldsymbol{\theta}})^{\tilde{n}_k-1} \tag{165}$$

τείνει στο 0 για μεγάλες τιμές της $D_{\frac{1}{2},\tilde{n}_k}(\boldsymbol{\theta}||\tilde{\boldsymbol{\theta}})$ και ως εκ τούτου είναι κανονικοποιήσιμη.
Από την αναδρομική σχέση παρατηρούμε την ενδιαφέρουσα ιδιότητα της $\frac{1}{2}$-απόκλισης σε σχέση με τις τροχιές $\{\hat{\boldsymbol{\eta}}^t\}_{t=1,2,\ldots}$. Η νέα θέση $\hat{\boldsymbol{\eta}}^{t+1}$ θα είναι μέσος όρος με βάρη όχι των $\tilde{\boldsymbol{\eta}}^k$, $k = 1, \ldots, N$ αλλά των $\langle\boldsymbol{\eta}\rangle_{\tilde{\boldsymbol{\theta}}_k}$, $k = 1, \ldots, N$, όπου

$$\langle\boldsymbol{\eta}\rangle_{\tilde{\boldsymbol{\theta}}_k^t} = \boldsymbol{\eta}\left(\frac{\tilde{\boldsymbol{\theta}}_k + \hat{\boldsymbol{\theta}}^t}{2}\right) \tag{166}$$

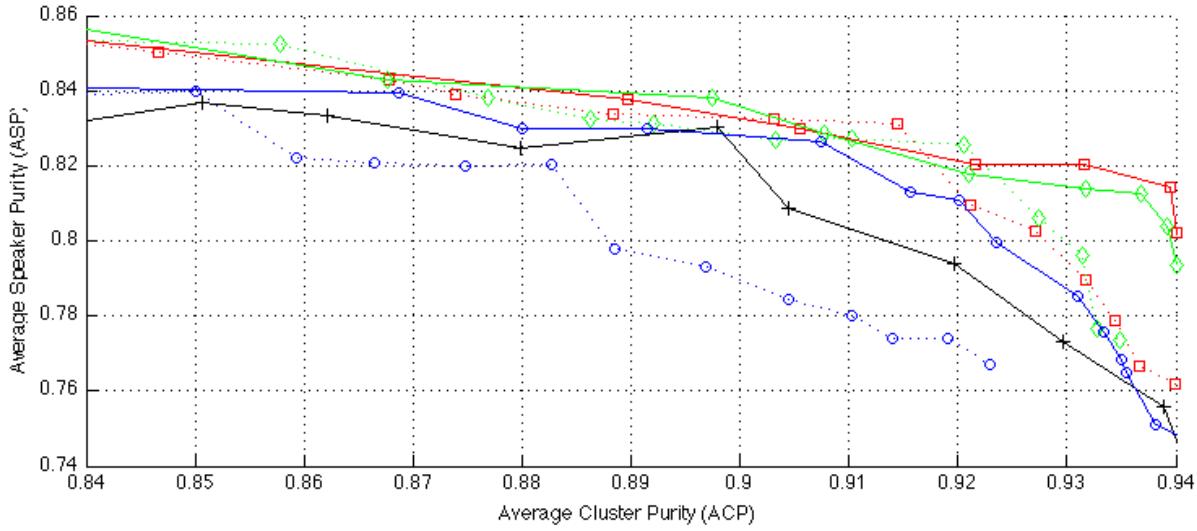
Αν θεωρήσουμε και πάλι την τετριμμένη περίπτωση όπου $N = 1$, παρατηρούμε ότι η νέα θέση $\hat{\boldsymbol{\eta}}^{t+1}$ στις $\boldsymbol{\eta}$-συντεταγμένες θα είναι ο μέσος όρος της θέσης του (μοναδικού) κέντρου έλξης (ή τρόπο, mode) $\tilde{\boldsymbol{\eta}}^1$ και της προηγούμενης θέσης $\hat{\boldsymbol{\eta}}^t$, με τη διαφορά ότι ο μέσος όρος υπολογίζεται στις δυικές συντεταγμένες $\boldsymbol{\theta}$. Σε Ευκλείδια γεωμετρία θα επρόκειτο λοιπόν για ένα πρώτης τάξης IIR βαθυπερατό φίλτρο, που σημαίνει ότι η ακολουθία συγκλίνει μόνο ασυμπτωτικά. Αυτό δεν αποτελεί πρόβλημα ωστόσο, καθώς πάντοτε αφήνουμε ένα περιθώριο απόκλισης κάτω από το οποίο θεωρούμε ότι δύο τροχιές συγκλίνουν σε κοινό τρόπο. Τα πειραματικά αποτελέσματα αποδεικνύουν ότι οι τροχιές συγκλίνουν πιο αργά σε σχέση με τις $\{0, 1\}$-γεωμετρίες, και ταυτόχρονα είναι πιο ομαλές.

6.6 Πειραματικά αποτελέσματα

Όπως και στα προηγούμενα κεφάλαια, εξετάζουμε τον προτεινόμενο αλγόριθμο χρησιμοποιώντας της βάση ESTER. Η σύγκριση πραγματοποιείται με τον αλγόριθμο ιεραρχικής ομαδοποίησης, με χρήση του τοπικού-BIC, τον πλέον διαδεδομένο αλγόριθμο στην ομαδοποίηση ομιλητών. Θα εξετάσσουμε τους εκθετικούς πυρήνες που προκύπτουν από τις $\delta = \{0, \frac{1}{2}, 1\}$ γεωμετρίες σε πολλαπλά σημεία λειτουργίας, τα οποία προκύπτουν από την επιλογή του συντελεστή εξομάλυνσης των κατανομών. Θα εξετάσσουμε επίσης και την επίπτωση της χρήσης βαρών $\tilde{w}_k$ έναντι των w_k^* στις κατανομές καθώς και του αριθμού εικονικών δειγμάτων που χρησιμοποιούνται στην εκτίμηση

MAP.

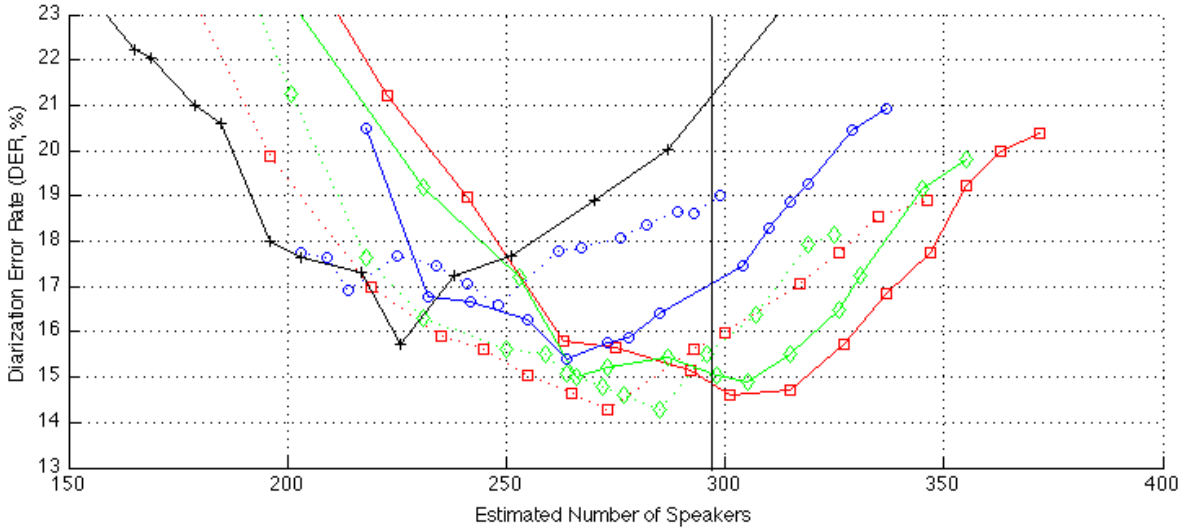
Το πρώτο διάγραμμα (Σχ. 24) εξετάζουμε την επίδοση των πυρήνων στο σύνολο ανάπτυξης (ESTER-DEV) με βάση τη μέση καθαρότητα και κάλυψη ομάδας (Average Cluster Purity vs. Average Speaker Purity, *acp-asp*), ενώ στο δεύτερο διάγραμμα (Σχ. 25) εξετάζουμε τον εκτιμώμενο αριθμό των ομιλητών σε σχέση με το ολικό σφάλμα ομαδοποίησης (DER,%).



Σχῆμα 24: *Average Cluster Purity vs. Average Speaker Purity, Γραμμές με τελείες: με βάρη $\tilde{w}_k$, Συνεχείς γραμμές: με βάρη w_k^*, Κόκκινη με τετράγωνα: $\delta = 1$, Μπλε με κύκλους: $\delta = \frac{1}{2}$, Πράσινη με ρόμβους: $\delta = 0$, Συνεχής γραμμή με σταυρούς: τοπικό-BIC.*

Η επίδοση του αλγορίθμου μας για $\delta = 1$ ανά εκπομπή συγκρίνεται με αυτή του τοπικού BIC στο Σχ. 26. Ο αριθμός των εικονικών δειγμάτων αντιστοιχεί σε διάρκεια 1.5 δευτερολέπτων ($N_v = 150$ εικονικές παρατηρήσεις). Οι μέσες τιμές των υπερπαραμέτρων είναι $(\mu_0, \Sigma_0) = (\mathbf{0}_d, \frac{0.75}{n}\sum_{i=1}^{n}\mathbf{y}_i\mathbf{y}_i^T)$, δηλαδή πίνακα συμμεταβλητότητας ίσο με τον 75% αυτής του αρχείου. Υπενθυμίζουμε ότι η ($\delta = 1$)-γεωμετρία σημαίνει αναδρομική σχέση στη θ παραμετροποίηση, η ($\delta = 0$)-γεωμετρία στην η παραμετροποίηση, ενώ η $\delta = \frac{1}{2}$ σε πυρήνα με τη απόσταση Hellinger.

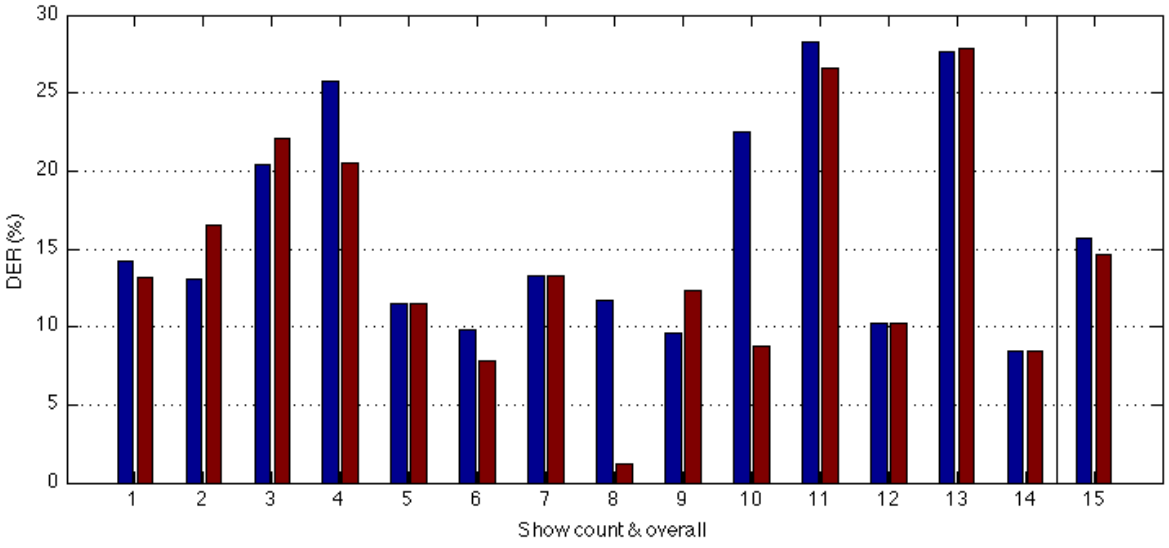
Από τα διαγράμματα παρατηρούμε τη σαφή υπεροχή της μεθόδου μας σε σχέση με το τοπικό-BIC, με εξαίρεση την απόσταση Hellinger. Πιο συγκεκριμένα, το Σχήμα 25 δείχνει την υπεροχή της



Σχῆμα 25: *Εκτιμώμενος αριθμός ομιλητών και DER, (%), Γραμμές με τελείες: με βάρη $\tilde{w}_k$, Συνεχείς γραμμές: με βάρη w_k^*, Κόκκινη με τετράγωνα: $\delta = 1$, Μπλε με κύκλους: $\delta = \frac{1}{2}$, Πράσινη με ρόμβους: $\delta = 0$, Συνεχής γραμμή με σταυρούς: τοπικό-BIC. Ο πραγματικός αριθμός των ομιλητών δίνεται από την κάθετη γραμμή.*

μεθόδου μας στην από κοινού εκτίμηση του αριθμού των ομιλητών (ENS) και της ελαχιστοποίησης του DER.

Οι κατανομές με χρήση βαρών $\tilde{w}_k$ - παρότι ελαχιστοποιούν το DER - έχουν γενικά χαμηλότερη επίδοση στο από κοινού διάγραμμα ENS-DER. Αυτό συμβαίνει επειδή η χρήση βαρών, δρώντας ως πολλαπλασιαστής της μάζας πιθανότητας κάθε συντελεστή της κατανομής, επιτρέπει στα τμήματα μικρής διάρκειας την εύκολη μεταπήδησή τους στη δεξαμενή έλξης των μεγαλύτερων. Υπενθυμίζεται ότι η χρήση της εκτίμησης MAP είναι ένας μηχανισμός που από μόνος του επιτρέπει στα μικρής διάρκειας τμήματα να αποκολληθούν πιο εύκολα από τη δική τους δεξαμενή έλξης. Ως εκ τούτου, η χρήση της εκτίμησης MAP επαρκεί για την ενσωμάτωση της διάρκειας των τμημάτων στον αλγόριθμο, καθιστώντας τη χρήση βαρών περιττή εώς επιζήμια. Η επιδείνωση αυτή φαίνεται και στο διάγραμμα *acp-asp* του Σχ. 24. Παρατηρήστε την πολύ καλή κάλυψη ομάδας (asp) που παρουσιάζουν οι κατανομές με βάρη w_k^* (συνεχείς γραμμές) για υψηλές τιμές καθαρότητας ομάδας (acp).

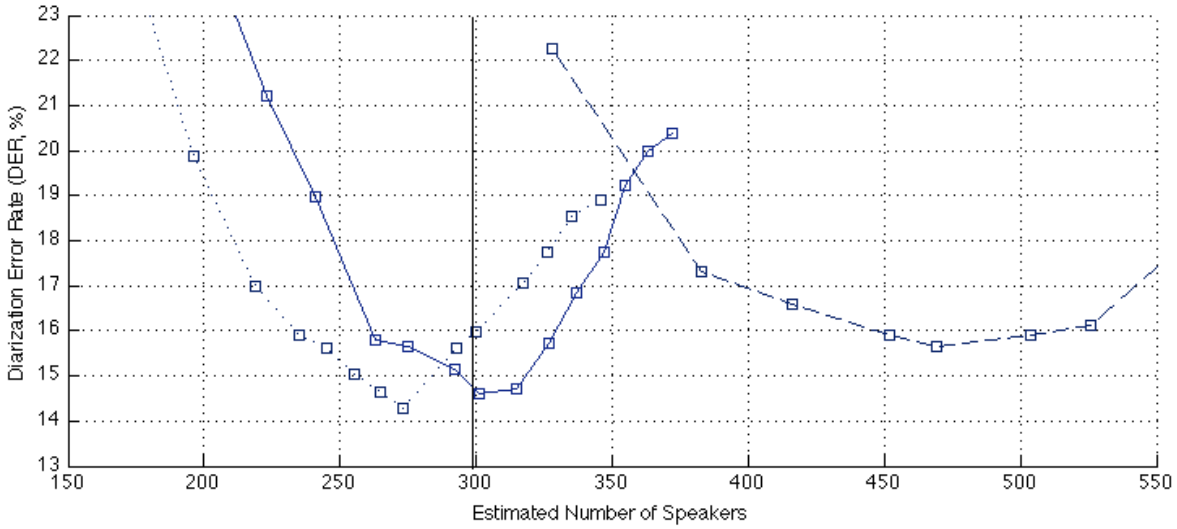


Σχῆμα 26: *Σύγκριση του DER, (%) ανά εκπομπή του συνόλου ανάπτυξης. Αριστερές μπάρες: Τοπικό-BIC, Δεξιές μπάρες: προτεινόμενος αλγόριθμος για* $\delta = 1$. *Τα συνολικά DER, (%) στην τελευταία στήλη.*

Στο Σχ. 28 εξετάζουμε της επίπτωση της εκτίμησης MAP, έναντι της χρήσης εκτίμησης ML καθώς και της χρήσης βαρών w_k^* ή $\tilde{w}_k$. Παρατηρήστε τη συμπεριφορά του αλγορίθμου με τη χρήση εκτίμησης ML και το βαθμό υπερεκτίμησης του αριθμού των ομιλητών. Κάθε αναδρομή εκκινεί από το κέντρο της δεξαμενής έλξης του, ανεξαρτήτως της διάρκειας του τμήματος. Ως εκ τούτου, τα μικρά τμήματα δεν διαφεύγουν εύκολα από την δεξαμενή έλξης τους, με αποτέλεσμα τη μεγάλη υπερεκτίμηση του αριθμού των ομιλητών. Επιπλέον, η τοποθέτηση βαρών δεν επαρκεί ώστε να διαφύγουν από τη δεξαμενή έλξης τους. Συμπεραίνουμε λοιπόν ότι η εκτίμηση MAP είναι μία βασική υπεύθυνη για την πολύ καλή συμπεριφορά του αλγορίθμου μας.

Στο Σχ. 28, η επίδραση του αριθμού των εικονικών παρατηρήσεων (strength, N_v) σκιαγραφείται για $N_v = \{100, 150, 200\}$, για χρήση εκτίμηση MAP έναντι ML καθώς και του είδους βαρών. Παρατηρούμε ότι ένα πλήθος εικονικών δειγμάτων κοντά στο 1.5 δευτερόλεπτο ($N_v = 150$) έχει την καλύτερη επίδοση στο από κοινού διάγραμμα ENS-DER.

Έχοντας ρυθμίσει τις παραμέτρους των μεθόδων με βάση το ESTER-DEV προχωρούμε στο σύνολο αξιολόγησης (ESTER-TEST). Τα συγκεντρωτικά αποτελέσματα παρατίθενται στον Πίνακα



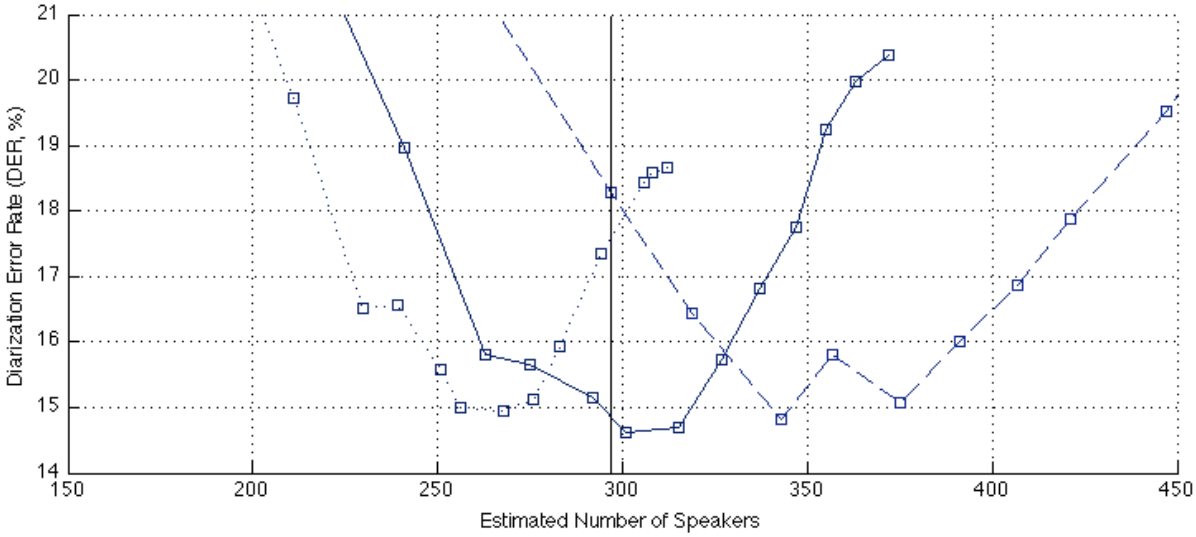
Σχῆμα 27: *Εκτιμώμενος αριθμός ομιλητών και DER, (%) για* $\delta = 1$. *Γραμμή με τελείες: με βάρη* $\tilde{w}_k$, *MAP εκτίμηση, Συνεχής γραμμή: με βάρη* w_k^*, *MAP εκτίμηση, Διακεκομμένη γραμμή: με βάρη* $\tilde{w}_k$, *ML εκτίμηση. Ο πραγματικός αριθμός των ομιλητών δίδεται από την κάθετη γραμμή.*

4. Στη στήλη TEST δίνουμε τα βέλτιστα αποτελέσματα επί του ESTER-TEST, χωρίς να λάβουμε υπόψη μας τις βέλτιστες ρυθμίσεις που εξαγάγαμε βάσει του ESTER-DEV. Οι επιδόσεις με βάση το ESTER-DEV δίδονται στη στήλη TEST*. Στην παρένθεση αναφέρουμε και το ποσοστό σφάλματος εκτίμησης του αριθμού ομιλητών, το οποίο ορίζουμε ως

$$SE = \frac{TNS - ENS}{TNS} \times 100\% \tag{167}$$

όπου TNS ο συνολικός αριθμός ομιλητών της βάσης και ENS ο εκτιμώμενος από τη μέθοδο αριθμός. Η επίδοση του αλγορίθμου μας για $\delta = 1$ ανά εκπομπή συγκρίνεται με αυτή του τοπικού BIC στο Σχ. 29. Η παράμετροι έχουν ρυθμισθεί βάσει του συνόλου ανάπτυξης.

Τα αποτελέσματα δείχνουν τη σαφή υπεροχή της μεθόδου μας σε σχέση με το τοπικό - BIC. Η υπεροχή αυτή είναι μάλιστα πιο εμφανής όταν η αξιολόγηση γίνεται με βάση την από κοινού DER-SE μετρική αξιολόγησης. Η $\delta = 1$ γεωμετρία, η οποία αντιστοιχεί στη χρήση εντροπικών εκ των προτέρων κατανομών και αναδρομικό τύπο με βάση της φυσικής παραμέτρους υπερέχει σε σχέση με τη $\delta = 0$ γεωμετρία (συζυγείς εκ των προτέρων κατανομών και αναμενόμενες παράμετροι στον αναδρομικό τύπο). Τέλος, η απόσταση Hellinger, παρά τη συμμετρία ως προς τα ορίσματά της,

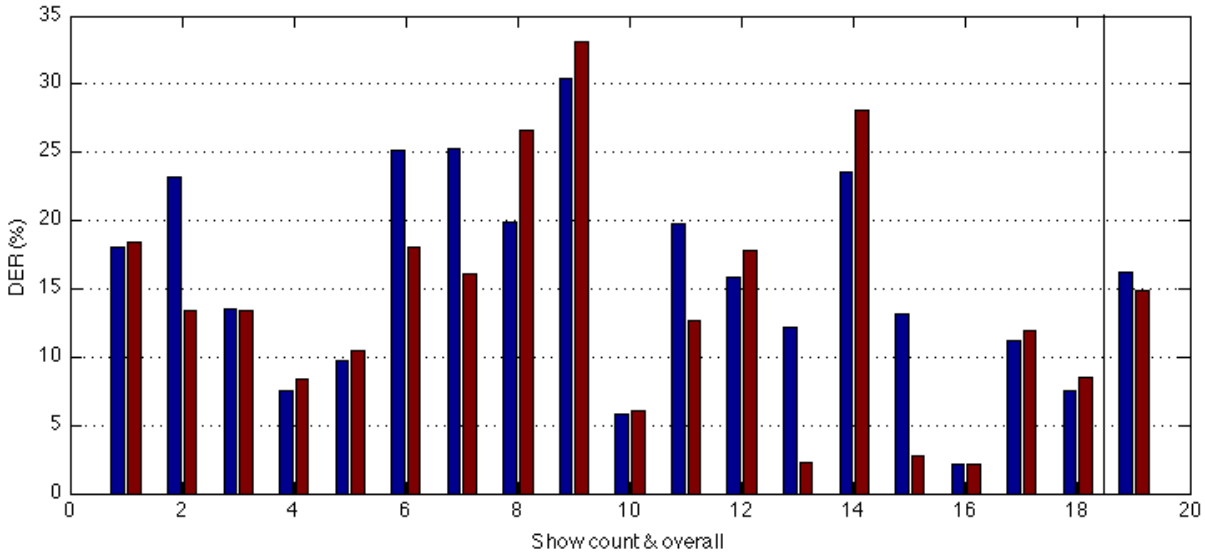


Σχῆμα 28: *Εκτιμώμενος αριθμός ομιλητών και DER, (%) για* $\delta = 1$. *Εξέταση της επίδρασης του πλήθους των εικονικών παρατηρήσεων (strength,* N_v*). Διακεκομμένη γραμμή:* $N_v = 100$*, Συνεχής γραμμή:* $N_v = 150$*, Γραμμή με τελείες:* $N_v = 200$. *Ο πραγματικός αριθμός των ομιλητών δίδεται από την κάθετη γραμμή.*

οδήγησε σε εμφανώς υποδεέστερα αποτελέσματα.

Πίνακας 4: Ελάχιστο ολικό σφάλμα ομαδοποίησης (%) και σε παρένθεση το ποσοστό σφάλματος εκτίμησης αριθμού ομιλητών (%). Στη στήλη TEST* η ρύθμιση παραμέτρων γίνεται βάσει του συνόλου DEV.

	DEV	TEST	TEST*
Τοπικό - BIC	15.76 (-23)	15.86 (-9.4)	16.28 (-20)
Mean-Shift, $\delta = 1$	14.65 (+2.3)	14.54 (-1.8)	14.91 (+2.8)
Mean-Shift, $\delta = \frac{1}{2}$	15.77 (-7.8)	15.60 (+1.0)	15.88 (-9.3)
Mean-Shift, $\delta = 0$	14.93 (+3.6)	15.74 (+6.3)	15.74 (+6.3)
Σφάλμα εσφαλμένου εντοπισμού ομιλίας	0.3	0.6	0.6
Σφάλμα μη εντοπισμού ομιλίας	0.9	1.2	1.2



Σχῆμα 29: *Σύγκριση του DER, (%) ανά εκπομπή του συνόλου αξιολόγησης. Αριστερές μπάρες: Τοπικό-BIC, Δεξιές μπάρες: προτεινόμενος αλγόριθμος για* $\delta = 1$. *Τα συνολικά DER, (%) στην τελευταία στήλη.*

6.7 Επίλογος Κεφαλαίου

6.7.1 Σύνοψη της προτεινόμενης μεθόδου

Στο κεφάλαιο αυτό προσαρμόσαμε τον αλγόριθμο μετατόπισης του μέσου στο πρόβλημα της ομαδοποίησης ομιλητών. Η προσαρμογή αυτή ήταν αναγκαία, καθώς ο συγκεκριμένος αλγόριθμος δρα στο πεδίο των παρατηρήσεων, σε αντίθεση με το πρόβλημά μας, στο οποίο οι αλγόριθμοι δρουν στο πεδίο των τμημάτων ομιλίας. Δείξαμε ότι οι τετραγωνικές αποστάσεις έχουν ως φυσικό ανάλογο την οικογένεια των δ-αποκλίσεων καθως και το ότι οι πυρήνες μπορούν να αντικατασταθούν από τις εκ των προτέρων κατανομές που προτείναμε. Επιπλέον, χρησιμοποιώντας στοιχεία της γεωμετρίας της πληροφορίας, δείξαμε ότι για την οικογένεια των εκθετικών κατανομών, ο αλγόριθμος έχει αναλυτική λύση και ως εκ τούτου δεν υπεισέρχονται ευρετικοί κανόνες. Προτείναμε, επίσης, την ενσωμάτωση της πληροφορίας που φέρει η διάρκεια των τμημάτων μέσω της εκτίμησης MAP. Η τελευταία αυτή προσαρμογή βελτίωσε θεαματικά τα αποτελέσματα, καθιστώντας τα ανώτερα από αυτά που επιτυγχάνουμε με χρήση του τοπικού-BIC στη βάση ESTER. Τονίζουμε ότι ο συγκεκριμένος αλγόριθμος είναι εφαρμόσιμος σε οποιαδήποτε ανάλογη εφαρμογή ομαδοποίησης. Η μόνη απαίτηση είναι το να ανήκει η κατανομή με την οποία περιγράφουμε τα τμήματα σε εκθετική οικογένεια.

6.7.2 Κατευθύνσεις για μελλοντική έρευνα

Για μελλοντική έρευνα προτείνονται δύο κατευθύνσης. Αρχικά, η εφαρμογή του αλγορίθμου σε άλλα προβλήματα ομαδοποίησης θα έχει ιδιαίτερο ενδιαφέρον. Ιδιαίτερα στο πεδίο της ομαδοποίησης ομιλητών, θα επιχειρήσουμε να επεκτείνουμε τον αλγόριθμο ώστε να υποστηρίζει μίγματα κανονικών κατανομών (GMMs). Σημειώνουμε ότι τα GMMs σχηματίζουν οικογένεια κατανομών εφόσον χρησιμοποιηθούν τα ολοκληρωμένα δεδομένα (complete-data) $\mathbf{z}^{(i)} = (\mathbf{x}^{(i)}, \mathbf{y}^{(i)})$, όπου

$\mathbf{x}^{(i)}$ η (πιθανοτική) εκτίμηση του συντελεστή του GMM στον οποίον ανήκει η i-οστη παρατήρηση. Τα κρυφά δεδομένα, όμως, μπορούν να εκτιμηθούν από το E-βήμα της τελευταίας αναδρομής του αλγορίθμου EM. Αντικαθιστώντας $\mathbf{x}^{(i)}$ με την εκτίμησή της, $\tilde{\mathbf{x}}^{(i)}$ και θεωρώντας τη συνάρτηση πιθανοφάνειας των $\mathbf{z}^{(i)}$, το μοντέλο ανήκει πλέον σε εκθετική κατανομή. Οι απαραίτητες παραμετροποιήσεις δίδονται στα [92], [93]. Τέλος, η εκτίμηση MAP - απαραίτητη για τη συγκεκριμένη εφαρμογή, αποτελεί ήδη τμήμα της εκπαίδευσης των συγκεκριμένων μοντέλων ([94]). Ως εκ τούτου, η εφαρμογή της μεθόδου μας δεν απαιτεί νέο κώδικα για τη στατιστική περιγραφή των τμημάτων και ενδέχεται να αποτελέσει εναλλακτική λύση σε πλήρως Μπεϋσιανές μεθόδους, [80], [95].

Μία δεύτερη κατεύθυνση μελλοντικής έρευνας θα είναι η εξέταση μεθόδων εκτίμησης του βέλτιστου συντελεστή εξομάλυνσης από τα ίδια τα δεδομένα. Με τον τρόπο αυτόν, η διαδικασία της εκμάθησης είναι δυνατόν να αποφευχθεί και να έχουμε έναν πλήρως αυτοματοποιημένο αλγόριθμο. Στα πλαίσια αυτά, υπάρχουσες τεχνικές μπορούν να εξετασθούν, με την κατάλληλη ωστόσο προσαρμογή που απαιτείται για τη μετάβαση από το πεδίο των παρατηρήσεων σε αυτό των παραμέτρων. Ταυτόχρονα, η χρήση διαφορετικού συντελεστή εξομάλυνσης για κάθε τμήμα ομιλίας (variable bandwidth) θα πρέπει να εξετασθεί επίσης. Η βιβλιογραφία της μη-παραμετρικής εκτίμησης κατανομών είναι πλούσια σε αυτόν τον τομέα και πολλές από αυτές είναι εύκολα προσαρμόσιμες στην περίπτωσή μας, [87], [96]. Το γεγονός αυτό, σε συνδυασμό με την πολύ καλή επίδοση του αλγορίθμου με σταθερό και ενιαίο για κάθε αρχείο συντελεστή εξομάλυνσης, μας κάνει να πιστεύουμε ότι ο συγκεκριμένος αλγόριθμος μπορεί και βελτιωθεί σημαντικά.

7 ΕΠΙΛΟΓΟΣ

Η διατριβή αυτή είχε ως σκοπό:

1. να παρουσιάσει το πρόβλημα της αυτόματης ομαδοποίησης αρχείων ομιλίας σε ομιλητές και τις εφαρμογές του,

2. να καλύψει με επάρκεια τις υπάρχουσες προσεγγίσεις με έμφαση στις βήμα-προς-βήμα αλγοριθμικές προσεγγίσεις,

3. να προτείνει νέες μεθόδους, αλγοριθμικές προσεγγίσεις και κριτήρια για ομαδοποίηση, σε περιπτώσεις όπου η πραγματική κατανομή των δειγμάτων διαφέρει σημαντικά από τη μοντελοποίησή μας (model's misspecification),

4. να αναδείξει τη μεθοδολογία της Μπεϋσιανής στατιστικής και της Γεωμετρίας της Πληροφορίας στην εκμάθηση μηχανών και ιδιαίτερα στις εφαρμογές ομαδοποίησης.

Για τον πρώτο σκοπό, αναπτύξαμε τα πεδία εφαρμογής των αλγορίθμων, με έμφαση σε αυτό των εκπομπών ειδησεογραφικού χαρακτήρα, συνδέσαμε τους αλγορίθμους και τη ρύθμιση των παραμέτρων με τον απώτερο στόχο της ομαδοποίησης (αναγνώριση ομιλητών, αναγνώριση φωνής, εμπλουτισμός κειμένου αυτόματης απομαγνητοφώνησης, κ.ά.) καθώς και τις μετρικές αξιολόγησης των επιδόσεων των αλγορίθμων.
Όσον αφορά τον δεύτερο σκοπό, κάναμε μία περιεκτική αναφορά στις διάφορες μεθόδους που έχουν προταθεί, αναφερθήκαμε στις μεθόδους κατάτμησης και κατηγοριοποίησης των παρατηρήσεων σε ευρείες κλάσεις, παρουσιάσαμε τις τεχνικές και τα κριτήρια που χρησιμοποιούνται κατά κόρον στο στάδιο της ομαδοποίησης των τμημάτων ομιλίας και τέλος, αναφερθήκαμε στο στάδιο μετεπεξεργασίας και συγκεκριμένα σε αλγορίθμους αναθεώρησης των σημείων εναλλαγής ομιλητών.
Ερχόμαστε, τώρα, στον τρίτο και τέταρτο σκοπό της διατριβής, αυτόν της πρότασης νέων μεθόδων, αλγοριθμικών προσεγγίσεων και κριτηρίων με βάση τη Μπεϋσιανή στατιστική και τη γεωμετρία της

Πληροφορίας. Οι προτάσεις μας αναπτύχθηκαν στα Κεφάλαια 4, 5 και 6 και κάλυψαν το πρόβλημα της ομαδοποίησης δεδομένης μίας αρχικής κατάτμησης. Η πρώτη μας πρόταση αναπτύχθηκε στο Κεφάλαιο 4 και αφορά στον αποτελεσματικό συνδυασμό παραμετρικών χώρων, μετρικών απόκλισης και κατωφλιώσεων μέσω εκθετικών μοντέλων που προκύπτουν με εφαρμογή της αρχής της Μέγιστης Εντροπίας. Με τα μοντέλα αυτά επιχειρείται η προσέγγιση της εκ των υστέρων πιθανότητας δύο τμήματα ομιλίας να προέρχονται από τον ίδιο ομιλητή. Επιπλέον, δείξαμε ότι τα μοντέλα αυτά μπορούν να εξειδικευθούν περαιτέρω μέσα από μία πολυεπίπεδη κατάτμηση του χώρου εισόδου και κατόπιν να συνδυασθούν μεταξύ τους δημιουργώντας έτσι ένα υπερ-μοντέλο. Παρουσιάσαμε το υπόβαθρο της μεθόδου και συγκεκριμένα την εξαγωγή του ως λύσης μέγιστης εντροπίας με περιορισμούς πάνω στις αναμενόμενες τιμές των ταξινομητών, όπως επίσης και μεθόδους εκτίμησης των παραμέτρων και επιλογής των καταλληλότερων ταξινομητών.

Η δεύτερη συνεισφορά της διατριβής είναι η επαναπροσέγγιση ενός από τα θεμελιώδη κριτήρια ομαδοποίησης ομιλητών, του Μπεϋσιανού Κριτηρίου Πληροφορίας (BIC). Η κεντρική ιδέα του προτεινόμενου κριτηρίου είναι η σύνδεση των βασικών χαρακτηριστικών των δύο υπαρχουσών εκδοχών του BIC (τοπικό και ολικό) σε ένα νέο κριτήριο, το οποίο ονομάσαμε τμηματικό-BIC. Δείξαμε ότι η ιδιότητα της αυτόνομης μετρικής απόκλισης που παρέχει το τοπικό-BIC μπορεί να συνδυαστεί με την απαίτηση να προσεγγίζουμε τη λογαριθμική πιθανοφάνεια ταξινόμησης συνολικών ομαδοποιήσεων, ιδιότητα που έχει το ολικό BIC. Η μεθοδολογία μας βασίστηκε στην ανάλυση των Kass & Wasserman όσον αφορά στις εκ των προτέρων πιθανότητες που υπονοούνται από το BIC. Δείξαμε ότι η πιθανοφάνεια κατάταξης επιτρέπει την τροποποίηση των εκ των προτέρων πιθανοτήτων των παραμέτρων, έτσι ώστε η πληροφορία που φέρουν να αντιστοιχεί από μία παρατήρηση ανά αρχείο σε μία παρατήρηση ανά ομάδα. Δείξαμε ότι η επιλογή αυτή είναι συμβατή με την ομαδοποίηση σε ομιλητές, καθώς η τελευταία δε αποτελεί τη φυσική κατάτμηση του χώρου αλλά μπορεί να επιτευχθεί μόνο μέσω ενός ισχυρού μηχανισμού παραμονής στην παρούσα κατάσταση (state persistence). Ως εκ τούτου, οι προτεινόμενες εκ των προτέρων κατανομές συνεχίζουν να φέρουν ασαφή (vague) πληροφορία συγκρινόμενες με τη φυσική ομαδοποίηση του χώρου. Κατόπιν, με βάση την παρατήρηση ότι τα μοντέλα απλής κανονικής κατανομής είναι ελλειπή (misspecified) για την περιγραφή της διαδικασίας δημιουργίας των παρατηρήσεων, τροποποιήσαμε τον όρο ποινής

ώστε να ποινικοποιεί πιο δραστικά την πολυπλοκότητα. Τα πειραματικά αποτελέσματα έδειξαν σημαντικότατη αύξηση στην ακρίβεια ομαδοποίησης στο σύνολο των σημείων λειτουργίας.

Η τελευταία συνεισφορά της διατριβής παρουσιάσθηκε στο Κεφάλαιο 6, όπου προτείναμε μία τροποποιημένη μορφή του αλγορίθμου της μετατόπισης του μέσου (Mean Shift) ως εναλλακτικής προσέγγισης στο πρόβλημα της ομαδοποίησης τμημάτων ομιλίας. Ο συγκεκριμένος αλγόριθμος, παρότι έχει βρει εφαρμογή σε πολλά προβλήματα που σχετίζονται με την επεξεργασία εικόνας, δεν έχει εξετασθεί σε προβλήματα ομαδοποίησης ηχητικών οντοτήτων. Ως εκ τούτου, αποκτά ενδιαφέρον να διερευνηθεί η ικανότητά του στην ομαδοποίηση ομιλητών, αλλά και σε άλλες εφαρμογές της επεξεργασίας φωνής και ομιλητή. Ο αλγόριθμος αυτός υπέστη σημαντική προσαρμογή, καθώς δρά στο πεδίο των παρατηρήσεων. Χρησιμοποιώντας Μπεϋσιανές τεχνικές στην εκτίμηση παραμέτρων και Γεωμετρία της Πληροφορίας, δείξαμε ότι ο αλγόριθμος είναι ικανός να δρα και στο πεδίο των στατιστικών μοντέλων εκθετικών κατανομών. Τα πειραματικά αποτελέσματα έδειξαν ότι η προτεινόμενη μέθοδος είναι πιο ακριβείς σε σχέση με την καθιερωμένη αλγοριθμική αντιμετώπιση του προβλήματος, τον αλγόριθμο ιεραρχικής συγκέντρωσης, με χρήση του τοπικού-BIC.

Θα πρέπει να τονισθεί ότι οι αλγόριθμοι και τα κριτήρια που προτείνονται στα Κεφάλαια 5 και 6 αφορούν σε γενικές μεθόδους ομαδοποίησης οντοτήτων που μπορούν να περιγραφούν στατιστικά, και ως εκ τούτου θα είχε ενδιαφέρον να εξετασθούν και σε πλειάδα άλλων εφαρμογών ομαδοποίησης, πολύ διαφορετικών από αυτήν της ομαδοποίησης ομιλητών.

8 ΠΑΡΑΡΤΗΜΑ 1: BIC, ΚΛΕΙΣΤΟΙ ΤΥΠΟΙ ΚΑΙ ΧΡΟΝΙΚΗ ΠΛΗΡΟΦΟΡΙΑ

8.1 Κλειστοί τύποι υπολογισμού και BIC

Στην παράγραφο αυτή, παρουσιάζουμε εν συντομία της εργασία μας ([6]) που αφορά στις κλειστές φόρμες του παραπάνω κριτηρίου. Υπενθυμίζουμε ότι το BIC αναπτύχθηκε από τον G. Schwarz ως προσέγγιση της λογαριθμικής ολοκληρωμένης πιθανοφάνειας ενός μοντέλου (προσέγγιση $\mathcal{O}(1)$) για την οικογένεια των εκθετικών κατανομών, [68], ενώ η ισχύς του για παραμετρικά μοντέλα πέραν της συγκεκριμένης οικογένειας έχει επιδειχθεί μόνο πειραματικά.
Στην συγκεκριμένη μοντελοποίηση που υιοθετούμε ωστόσο, το BIC έχει κλειστή μορφή υπολογισμού. Το γεγονός αυτό δεν έχει επισημανθεί από την κοινότητα που ασχολείται με την ομαδοποίηση ομιλητών. Η πλειονότητα των εργασιών πάνω στο συγκεκριμένο κριτήριο, είτε θεωρεί το κριτηρίο προσέγγιση ενός μη-υπολογίσιμου ολοκληρώματος, είτε το ερμηνεύει μέσω της θεωρίας πληροφορίας (Ελάχιστο Μήκος Περιγραφής, Minimum Description Length), [97], [71], [69], [79]. Στην παράγραφο αυτή παρουσιάζουμε τους κλειστούς τύπους υπολογισμού του κριτηρίου, μαζί με τη μέθοδο εξαγωγής τους.

8.2 Συζυγείς εκ των προτέρων κατανομές και ολοκληρωμένη πιθανοφάνεια

Θεωρούμε αρχικά την περίπτωση όπου το σύνολο των παρατηρήσεων $\mathbf{y} \in \mathcal{Y}^n$ σχηματίζει μία και μόνη ομάδα. Για να εξάγουμε την ολοκληρωμένη πιθανοφάνεια, θα πρέπει να χρησιμοποιήσουμε την Κανονική - Αντίστροφη Wishart εκ των προτέρων κατανομή για τις παραμέτρους $\varphi = (\mu, \Sigma)$.

Προχωρούμε λοιπόν ως εξής. Αρχικά δεσμεύουμε την κατανομή της μ ως προς Σ, και έχουμε

$$\mu|\Sigma \sim \mathcal{N}(\mu_0, \frac{1}{\nu}\Sigma) \tag{168}$$

που σημαίνει ότι θεωρούμε κέντρο της κατανομής την τιμή μ_0 με εκ των προτέρων αμφιβολία (prior uncertainty) ίση με $\frac{1}{\nu}\Sigma$. Κατόπιν, ολοκληρώνουμε ως προς μ και θέτοντας $\bar{\mathbf{y}} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{y}^{(i)}$, έχουμε

$$p(\mathbf{y}|\Sigma) = \left(\frac{\nu}{n+\nu}\right)^{d/2}(2\pi)^{-nd/2}|\Sigma|^{-n/2}\exp\left(-\frac{1}{2}\mathrm{tr}(\Sigma^{-1}P)\right) \tag{169}$$

όπου

$$P = \sum_{i=1}^{n}(\mathbf{y}^{(i)} - \bar{\mathbf{y}})(\mathbf{y}^{(i)} - \bar{\mathbf{y}})^T + \frac{n\nu}{n+\nu}(\mu_0 - \bar{\mathbf{y}})(\mu_0 - \bar{\mathbf{y}})^T \tag{170}$$

και

$$m = \frac{1}{n+\nu}\left(n\bar{\mathbf{y}} + \nu\mu_0\right) \tag{171}$$

Η ποσότητα (171) είναι η αναμενόμενη εκ των υστέρων (posterior mean) τιμή της μ. Όπως παρατηρούμε, ισούται με τον μέσο όρο της αναμενόμενης τιμής των παρατηρήσεων μας και της εκ των προτέρων κατανομής με βάρη n και ν αντίστοιχα. Σημειώνουμε ότι το αποτέλεσμα αυτό δικαιολογεί τον όρο *εικονικές* ή *φανταστικές* παρατηρήσεις που αποδίδεται στις υπερπαραμέτρους (μ_0, ν).

Για να εξάγουμε της ποσότητα $p(\mathbf{y}|K=1)$ θα πρέπει να χρησιμοποιήσουμε τη συζυγή κατανομή στην πιθανοφάνεια της παραμέτρου Σ, δηλαδή την αντίστροφη-Wishart

$$\Sigma \sim \mathcal{IW}(\Psi, p) \tag{172}$$

όπου $\Psi = p\Sigma_0$, ενώ Σ_0 η εκ των προτέρων αναμενόμενη τιμή της Σ. Θέτωντας $p >= d-1$, επιτυγχάνουμε την πλέον ασαφή (vague) κατανομή, διασφαλίζοντας ταυτόχρονα το να είναι και ολοκληρώσιμη. Ταυτόχρονα, επαληθεύουμε και τον λόγο για τον οποίον επιλέξαμε να μην εφαρμόσουμε την παράμετρο ρύθμισης λ στο σύνολο των παραμέτρων, παρα μόνο στο υποσύνολο που αντιστοιχεί στις μέσες τιμές $\{\mu_k\}_{k=1}^{K}$. Εν αντιθέσει με την κανονική κατανομή, η συζυγής κατανομή του Σ_k απαιτεί έναν ελάχιστο αριθμό εικονικών δειγμάτων για να είναι ολοκληρώσιμη. Δεδομένης λοιπόν της ερμηνείας που αποδώσαμε στην παράμετρο ρύθμισης, οφείλουμε να μην την

εφαρμόσουμε στο σύνολο των παραμέτρων των $\{\Sigma_k\}_{k=1}^K$.

Ολοκληρώνοντας ως προς Σ και με χρήση βασικών τεχνικών άλγεβρας πινάκων, λαμβάνουμε

$$p(\mathbf{y}|K) = \left(\frac{\nu}{n+\nu}\right)^{\frac{d}{2}} \pi^{-\frac{nd}{2}} \frac{|\Psi|^{\frac{p}{2}}}{|\Psi+P|^{\frac{n+p}{2}}} \frac{\Gamma_d(\frac{p+n}{2})}{\Gamma_d(\frac{p}{2})} \tag{173}$$

όπου

$$\Gamma_d(x) = \pi^{d(d-1)/4} \prod_{j=1}^{d} \Gamma\left(x + \frac{1-j}{2}\right) \tag{174}$$

και $\Gamma(\cdot)$ η συνάρτηση Γάμμα.

Έχοντας εξαγάγει τον κλειστό τύπο υπολογισμού για την περίπτωση της μίας κλάσης ($K = 1$), μπορούμε τώρα να προχωρήσουμε και στον υπολογισμό της ολοκληρωμένης πιθανοφάνειας για τη γενική περίπτωση $K \geq 1$. Υπενθυμίζουμε ότι η στατιστική ποσότητα που επιθυμούμε να εξαγάγουμε είναι η ακόλουθη

$$p(\mathbf{y}|\mathbf{x}, K) = \prod_{k=1}^{K} p(\mathbf{y}_k) \tag{175}$$

όπου η δέσμευση της ανωτέρω ποσότητας ως προς τις υπερπαραμέτρους υπονοείται χάριν απλότητας.

Ως εκ τούτου, έχουμε να υπολογίσουμε το ακόλουθο ολοκλήρωμα

$$p(\mathbf{y}|\mathbf{x}, K) = \int_{\Phi} p(\mathbf{y}|\mathbf{x}, K)\pi(\varphi)d\varphi \tag{176}$$

όπου $\varphi = \{\varphi_k\}_{k=1}^K$.

Η δέσμευση της $\pi(\varphi)$ ως προς τις υπερπαραμέτρους καθιστά τις φ_k *i.i.d.* και ως εκ τούτου

$$\pi(\varphi) = \prod_{k=1}^{K} \pi(\varphi_k) \tag{177}$$

Η ίδια παραγοντοποίηση ισχύει και για τη συνάρτηση πιθανοφάνειας του μοντέλου, δηλαδή

$$p(\mathbf{y}|\varphi, \mathbf{x}, K) = \prod_{k=1}^{K} p(\mathbf{y}_k|\varphi_k),\ \varphi \in \Phi,\ \varphi_k \in \Phi_k \tag{178}$$

Ως εκ τούτου, η ζητούμενη ποσότητα στην (176) γράφεται ως

$$p(\mathbf{y}|\mathbf{x}, K) = \int_{\Phi_1} \dots \int_{\Phi_K} p(\mathbf{y}|\varphi, \mathbf{x}, K)\pi(\varphi_1)\dots\pi(\varphi_K)d\varphi_1\dots d\varphi_K \tag{179}$$

ή αλλιώς

$$p(\mathbf{y}|\mathbf{x}, K) = \prod_{k=1}^{K} \int_{\Phi_k} p(\mathbf{y}_k|\varphi_k)\pi(\varphi_k)d\varphi_k \tag{180}$$

Έτσι, ο ζητούμενος κλειστός τύπος υπολογισμού της ολοκληρωμένης πιθανοφάνειας είναι η ακόλουθη

$$p(\mathbf{y}|\mathbf{x}, K) = \prod_{k=1}^{K} \left(\frac{\nu_k}{n_k + \nu_k}\right)^{\frac{d}{2}} \pi^{-\frac{n_k d}{2}} \frac{|\Psi|^{\frac{p}{2}}}{|\Psi + P|^{\frac{n_k+p}{2}}} \frac{\Gamma_d(\frac{p+n_k}{2})}{\Gamma_d(\frac{p}{2})} \tag{181}$$

Ταυτόχρονα, είναι και η μοναδική, καθώς μπορεί να εξαχθεί μόνο με χρήση των συζυγών εκ των προτέρων κατανομών, οι οποίες είναι μοναδικές για κάθε συνάρτηση πιθανοφάνειας.

8.3 Χρονική πληροφορία και BIC

8.3.1 Εισαγωγή

Ερχόμαστε τώρα να εμπλουτίσουμε το παραπάνω κρητήριο, κάνοντας χρήση της χρονικής (ή δυναμικής) πληροφορίας. Οι εκδοχές του BIC που έχουν επικρατήσει στην ομαδοποίηση ομιλητών βασίζονται στην οριακή πιθανοφάνεια κατάταξης

$$p(\mathbf{y}|\mathbf{x}) = \int_{\Phi} f(\mathbf{y}|\varphi, \mathbf{x})\pi(\varphi|\mathbf{x})d\varphi. \tag{182}$$

Η Μπεϋσιανή διαδικασία σύμφωνα με την οποία το βέλτιστο $\mathbf{x}$ (βέλτιστο με την έννοια του εκ των υστέρων πιο πιθανού) προκύπτει ως εκείνο το οποίο μεγιστοποιεί την οριακή πιθανοφάνεια κατάταξης υποθέτει ομοιόμορφη κατανομή για $\mathbf{x} \in \mathcal{X}$. Η χρήση ομοιόμορφης κατανομής, όμως, αντιβαίνει με τη φύση της διαδικασίας δημιουργίας των δεδομένων η οποία έχει έντονα Μαρκοβιανά χαρακτηριστικά και συγκεκριμένα παρουσιάζει σημαντική εμμονή στην ίδια κατάσταση (state persistence).

Επιπλέον, λόγω της ανεπαρκούς μοντελοποίησης των δεδομένων ομιλητή, η διαδικασία μεγιστοποίησης της (182) είναι αδύνατον να οδηγήσει σε ομαδοποίηση των $\mathbf{y}$ με βάση τους ομιλητές.

Αντίθετα, ομαδοποιεί τις παρατηρήσεις γύρω από φωνήματα, τη φυσική ομαδοποίηση των $\mathbf{y}$ δεδομένου του παραμετρικού χώρου (MFCC, PLP κ.ο.κ.). Δείξαμε ότι η συνήθης αντιμετώπιση του προβλήματος αυτού είναι η επιβολή περιορισμών στα $\mathbf{x}$, έτσι ώστε το ελάχιστο τμήμα μη αλλαγής ομάδας που εμφανίζεται να υπερβαίνει ένα εμπειρικό κατώφλι (π.χ. 1 έως 2 sec). Στους βήμα-προς-βήμα (ή μη-συζευγμένους) αλγορίθμους, ο περιορισμός αυτός τίθεται από το στάδιο κατάτμησης. Έτσι, το στάδιο της ομαδοποίησης έχει στη διάθεσή του ένα σύνολο τμημάτων με διάρκειες άνω του κατωφλίου. Ο περιορισμός αυτός είναι ισοδύναμος με επιβολή μηδενικής εκ των προτέρων πιθανότητας στο υποσύνολο των ομαδοποίησων $\mathbf{x} \in \mathcal{X}^0 \subset \mathcal{X}$ που παραβιάζουν τον περιορισμό και χρήση ομοιόμορφης κατανομής στα $\mathbf{x} \in \mathcal{X}^1 \subset \mathcal{X}$, όπου $\mathcal{X}^1$ το υποσύνολο που υπακούει στο περιορισμό.

8.3.2 Υπερβαίνοντας την ομοιόμορφη κατανομή

Μελετούμε τώρα μεθόδους που εντάσσονται πιο ομαλά στο πνεύμα των Μπεϋσιανών στατιστικών. Δείξαμε ότι τα κριτήρια που κάνουν χρήση της πιθανοφάνειας κατάταξης εκφράζουν την *εκ των υστέρων* πιθανότητα του $\mathbf{x}$ ως $P(\mathbf{x}|\mathbf{y}) \propto P(\mathbf{y}|\mathbf{x})\pi(\mathbf{x})$. Τόσο το ολικό όσο και το τμηματικό BIC προχωράει στην ανάλυση της οριακής πιθανοφάνειας ως

$$P(\mathbf{y}|\mathbf{x}) = \int_{\Theta} P(\mathbf{y}|\varphi, \mathbf{x})\pi(\varphi|\mathbf{x})d\varphi \tag{183}$$

υποθέτοντας ομοιόμορφη κατανομή $\pi(\mathbf{x})$ για $\mathbf{x} \in \mathcal{X}$.

Για τη μοντελοποίηση της διαδικασίας δημιουργίας των $\mathbf{x}$ σε σχέση με τον αριθμό των ομιλητών K μπορούν να ακολουθήσουμε δύο στρατηγικές. Οι στρατηγικές αυτές μπορούν σχηματικά να διατυπωθούν ως εξής:

1. Όλοι οι συμμετέχοντες ομιλητές θα λάβουν τον λόγο

2. Ένα υποσύνολο των συμμετεχόντων ομιλητές θα λάβει τον λόγο

Έστω $\hat{K} \equiv \max(\hat{\mathbf{x}})$ και $\mathcal{K}(n) \subset \mathbb{N}$ το υποσύνολο των φυσικών αριθμών τέτοιο ώστε η πιθάνοτητα $P(k \notin \mathcal{K}(n)|n)$ να είναι αμελητέα. Ας θεωρήσουμε αρχικά την πρώτη στρατηγική. Σύμφωνα με αυτήν, ισχύει ότι $\pi(\mathbf{x}) \equiv p_{\mathbf{x}}(\hat{\mathbf{x}}) = p_{\mathbf{x}|k}(\hat{\mathbf{x}}|\hat{K})p(\hat{K}|n)$, αφού $p_{\mathbf{x}|k}(\hat{\mathbf{x}}|k) = 0$ για $k \neq \hat{K}$ και το $\hat{K}$ προκύπτει μονοσήμαντα από την $\hat{\mathbf{x}}$. Αν θεωρήσουμε τη δεύτερη στρατηγική, θα έχουμε

$$\pi(\mathbf{x}) = \sum_{k \in \mathcal{K}(n)} p_{\mathbf{x}|K}(\mathbf{x}|k)p_{K|n}(k|n) \tag{184}$$

Σε αντίθεση με την πρώτη στρατηγική, εδώ ισχύει ότι $p_{\mathbf{x}|K}(\hat{\mathbf{x}}|k) = 0$ μόνο για $k < \hat{K}$. Έτσι, η πιθανότητα $p_{\mathbf{x}}(\hat{\mathbf{x}})$ που προκύπτει θα ενισχύει πιο φειδωλές μοντελοποιήσεις σε σχέση με την πρώτη στρατηγική. Η $p_{K|n}(K|n)$ μπορεί να θεωρηθεί ομοιόμορφη στο $\mathcal{K}(n)$. Μία εναλλακτική επιλογή είναι η Poisson δεδομένης της διάρκειας του αρχείου, n.

Σε πρώτη φάση θα επιχειρήσουμε να αναπτύξουμε τη δεσμευμένη πιθανότητα ως προς k $p_{\mathbf{x}|K}(\mathbf{x}|k)$, η οποία εμφανίζεται και στις δύο στρατηγικές. Έστω $\mathbf{a} \in \mathcal{A}$ το διάνυσμα των εξωτερικών παραμέτρων του μοντέλου (π.χ. οι πιθανότητες μετάβασης στο HMMs). Η δεσμευμένη πιθανότητα $p_{\mathbf{x}|K}(\mathbf{x}|k)$ μπορεί να ολοκληρωθεί ως προς τις $\mathbf{a} \in \mathcal{A}$

$$\pi_{\mathbf{x}|K}(\mathbf{x}|K) = \int_{\mathcal{A}} p_{\mathbf{x}|\mathbf{a},K}(\mathbf{x}|\mathbf{a}, K)\pi_{\mathbf{a}|K}(\mathbf{a}|K)d\mathbf{a} \tag{185}$$

Το παραπάνω ολοκλήρωμα είναι χρήσιμο γιατί εμπεριέχει τη δυναμική πληροφορία, η οποία χάνεται πλήρως θεωρώντας ομοιόμορφη κατανομή στα $\mathbf{x} \in \mathcal{X}$. Έτσι, στους αλγορίθμους αποσυζευγμένης κατάτμησης ομαδοποίησης, η δυναμική πληροφορία εξαντλείται στην κατάτμηση. Χωρίς την προσέγγιση του παραπάνω ολοκληρώματος, η ομαδοποίηση των τμημάτων ομιλίας ισοδυναμεί με αυτήν μεμονωμένων αρχείων ήχου που δεν έχουν χρονική συνάφεια μεταξύ τους. Η ανάπτυξη του παραπάνω ολοκληρώματος θα μπορούσε να δράσει συμπληρωματικά, ενθαρρύνοντας ομαδοποιήσεις οι οποίες έχουν αραιή διασύνδεση μεταξύ των καταστάσεων (sparce interconnectivity) καθώς και μη απότομες εναλλαγές μεταξύ των καταστάσεων. Με τον τρόπο αυτόν, θα ήταν δυνατή η αποφυγή ευρετικών κανόνων (όπως η ελάχιστη διάρκεια παραμονής σε μια κατάσταση) καθώς και η εξερεύνηση στρατηγικών για συγκεκριμένα πεδία (εκπομπές λόγου, συσκέψεις, κ.ο.κ.).

8.3.3 Ενσωμάτωση της χρονικής πληροφορίας με χρήση κατανομών Dirichlet

Μελετούμε τώρα τρόπους ενσωμάτωσης της χρονικής πληροφορίας. Η εξίσωση (185) είναι η εκ των προτέρων κατανομή της ομαδοποίησης $\mathbf{x}$, δεδομένου του K. Στην κατανομή αυτή προκύπτει από την ολοκλήρωση τής από κοινού πιθανότητας του ζεύγους $(\mathbf{x}, \mathbf{a})$ ως προς $\mathbf{a}$, δηλαδή τις εξωτερικές παραμέτρους του μοντέλου. Υποθέτοντας Κρυφό Μαρκοβιανό Μοντέλο ανεξάρτητο του χρόνου (time-independent HMM), το $\mathbf{a}$ αποτελείται από τον $K \times K$ πίνακα μετάβασεων μεταξύ των K καταστάσεων $a_{ij}, i = 1, \ldots, K, j = 1, \ldots, K$, όπου a_{ij} η πιθανότητα μετάβασης στην j-οστή κατάσταση δεδομένου ότι το σύστημα βρίσκεται στην i-οστή και από τις K το πλήθος πιθανότητες εκκίνησης $\{\pi_i\}_{i=1}^K$. Προς το παρόν, θεωρούμε το K δεδομένο και δεν το συμπεριλαμβάνουμε στους συμβολισμούς των ποσοτήτων.

Ας συμβολίσουμε με $\mathbf{a}_i$ την i-οστή γραμμή του πίνακα μετάβασης, δηλαδή το διάνυσμα $\mathbf{a}_i = [a_{i1}, \ldots, a_{iK}]$, όπου ισχύει ότι $\sum_{j=1}^K a_{ij} = 1$. Επιπλέον, έστω $\mathbf{x}_i$ το σύνολο των κρυφών μεταβλητών για τις οποίες η αμέσως προηγούμενη κατάσταση του συστήματος είναι η i-οστή, δηλαδή $\mathbf{x}_i = \{x^{\{(t)\}} : x^{\{(t-1)\}} = i\}$ και έστω $\mathbf{t}_i$ οι χρονικές στιγμές (time indices) που αντιστοιχούν στα $\mathbf{x}_i$. Η μεταβλητές $\mathbf{x}_i$ ακολουθούν μια πολυωνυμική (multinomial) κατανομή, δεδομένων των παραμέτρων $\mathbf{a}_i$, δηλαδή $p(\mathbf{x}_i|\mathbf{a}_i) = \prod_{t \in \mathbf{t}_i} a_{i,x^{(t)}}$. Η ολική πιθανότητα του $\mathbf{x}$ δεδομένων των $\mathbf{a}$, δηλαδή του συνόλου των εξωτερικών παραμέτρων του μοντέλου, δίδεται από τον παρακάτω τύπο

$$p(\mathbf{x}|\mathbf{a}) = \pi_{x^{\{1\}}} \prod_{i=1}^{K} p(\mathbf{x}_i|\mathbf{a}_i) = \pi_{x^{\{1\}}} \prod_{i=1}^{K} \prod_{t \in \mathbf{t}_i} a_{i,x^{(t)}} \tag{186}$$

Το πρόβλημα έγκειται στο ότι η παραπάνω ποσότητα θα πρέπει να ολοκληρωθεί ως προς $d\pi_{\mathbf{a}}(\mathbf{a})$, ώστε να οδηγήσει στην (185).

Μία λύση είναι θεωρήσουμε ότι η $\pi_{\mathbf{a}}(\mathbf{a})$ ως αυθαίρετη και να εφαρμόσουμε τεχνικές δειγματοληψίας. Αντίθετα, στην παρούσα διατριβή εστιάζουμε σε λύσεις οι οποίες οδηγούν σε κλειστούς τύπους υπολογισμού και ως εκ τούτου επιλέγουμε τη χρήση των συζυγών στη συνάρτηση πιθανοφάνειας εκ των προτέρων κατανομών της $\pi_{\mathbf{a}}(\mathbf{a})$. Δεδομένου ότι η συνάρτηση πιθανοφάνειας των $\mathbf{a}_i$ στα $\mathbf{x}_i$ είναι η κατηγοριακή, η συζυγής της είναι η κατανομή Dirichlet, η οποία δίδεται από τον παρακάτω

τύπο

$$\pi(\mathbf{a}_i|\mathbf{q}_i) = \frac{1}{\mathcal{B}(\mathbf{q}_i)} \prod_{j=1}^{K} a_{ij}^{q_{ij}-1} \delta\left(\sum_{j=1}^{K} a_{ij}, 1\right) \tag{187}$$

και θα την συμβολίζουμε ως $\mathbf{a}_i \sim \mathcal{D}ir(\mathbf{q}_i)$.

Στη σχέση (187) το διάνυσμα $\mathbf{q}_i = [q_{i1}, q_{i2}, \ldots, q_{iK}]$ είναι το σύνολο των υπερπαραμέτρων της $\pi(\mathbf{a}_i|\mathbf{q}_i)$. Η υπερπαράμετρος q_{ij} εκφράζει το πλήθος των μεταβάσεων που έχουν ήδη παρατηρηθεί, ανεξάρτητα από την τελική έκβαση των $\mathbf{x}_i$. Θέτοντας τις υπερπαραμέτρους ίσες με μονάδα, η κατανομή $\pi(\mathbf{a}_i|\mathbf{q}_i)$ είναι ομοιόμορφη, υπό την έννοια ότι η εκτίμηση μέγιστης πιθανοφάνειας των $\mathbf{a}_i$ ταυτίζεται με τη μέγιστη *εκ των υστέρων* εκτίμηση των $\mathbf{a}_i$, δεδομένων των $\mathbf{x}_i$.

Η συνάρτηση $\delta(\cdot, \cdot)$ είναι η συνάρτηση του Dirac και διασφαλίζει την απόδοση μηδενικής πιθανότητας σε $\mathbf{a}_i$ για τα οποία $\sum_{j=1}^{K} a_{ij} \neq 1$. Τέλος, η κανονικοποιητική σταθερά $\mathcal{B}(\mathbf{q}_i)$ είναι η συνάρτηση Βήτα (Beta) και η οποία γράφεται συναρτήσει της Γάμμα ως εξής

$$\mathcal{B}(\mathbf{q}_i) = \frac{\prod_{j=1}^{K} \Gamma(q_{ij})}{\Gamma(\sum_{j=1}^{K} q_{ij})}. \tag{188}$$

8.3.4 Υπολογισμός της εκ των προτέρων πιθανότητας των κρυφών μεταβλητών

Έχοντας ορίσει τις εκ των προτέρων κατανομές των παραμέτρων του μοντέλου προχωρούμε στον υπολογισμό της (185) που επαναλαβάνουμε για ευκολία στην ανάγνωση

$$\pi_{\mathbf{x}}(\mathbf{x}) = \int_{\mathcal{A}} p_{\mathbf{x}|\mathbf{a}}(\mathbf{x}|\mathbf{a}) \pi_{\mathbf{a}}(\mathbf{a}) d\mathbf{a} \tag{189}$$

όπου εννοούμε τη δεσμευμένη ως προς K πιθανότητα του $\mathbf{x}$. Συμβολίζουμε αρχικά με $\mathcal{G}$ τον $K \times K$ πίνακα, όπου το στοιχείο του g_{ij} εκφράζει το πλήθος των μεταβάσεων από την i-οστή στη j-οστή κατάσταση που παρατηρούνται στο $\mathbf{x}$. Αντικαθιστώντας τις (186) και (187) στην παραπάνω σχέση

υπολογίζουμε

$$\pi_{\mathbf{x}}(\mathbf{x}) = \pi_{x^1}(x^1) \int_{\mathcal{A}^K} \prod_{i=1}^{K} \prod_{j=1}^{K} a_{ij}^{g_{ij}} \mathcal{D}ir(\mathbf{a}_i|\mathbf{q}_i) d\mathbf{a} \tag{190}$$

$$= \pi_{x^1}(x^1) \prod_{i=1}^{K} \int_{\mathcal{A}} \prod_{j=1}^{K} a_{ij}^{g_{ij}} \mathcal{D}ir(\mathbf{a}_i|\mathbf{q}_i) d\mathbf{a}_i \tag{191}$$

όπου $\pi_{x^1}(x^1)$ η οριακή πιθανοφάνεια των παραμέτρων $\mathbf{a}_0$ σε σχέση με την πιθανότητα εκκίνησης από την κατάσταση x^0. Η δεύτερη ισοδυναμία προκύπτει από τη Μαρκοβιανή ιδιότητα

$$P(x^t|x^{t-1}, x^{t-2}, \ldots, x^1) = P(x^t|x^{t-1}). \tag{192}$$

Έτσι, η ποσότητα $\int_{\mathcal{A}} \prod_{j=1}^{K} a_{ij}^{g_{ij}} \mathcal{D}ir(\mathbf{a}_i|\mathbf{q}_i) d\mathbf{a}_i$ είναι ανεξάρτηση των παραμέτρων $\mathbf{a}_{i'}$ για $i' \neq i$, οδηγώντας στην παραπάνω μορφή γινομένου ολοκληρωμάτων. Αρκεί λοιπόν να υπολογίσουμε αυτήν την ποσότητα, η οποία μετά από αντικατάσταση του τύπου της κατανομής Dirichlet γράφεται ως

$$\int_{\mathcal{A}} \prod_{j=1}^{K} a_{ij}^{g_{ij}} \mathcal{D}ir(\mathbf{a}_i|\mathbf{q}_i) d\mathbf{a}_i = \frac{1}{\mathcal{B}(\mathbf{q}_i)} \int_{\mathcal{A}} \prod_{j=1}^{K} a_{ij}^{g_{ij}+q_{ij}-1} d\mathbf{a}_i \tag{193}$$

από που συμπεραίνουμε ότι το παραπάνω ολοκλήρωμα ισούται με τη σταθερά κανονικοποίησης της κατανομής Dirichlet με υπερπαραμέτρους τις $\mathbf{b}_i = [b_{i1}, \ldots, b_{iK}]$, όπου $b_{ij} = q_{ij} + g_{ij}$. Έτσι, η σχέση (190) γράφεται ως

$$\pi_{\mathbf{x}}(\mathbf{x}) = \pi_{x^1}(x^1) \prod_{i=1}^{K} \frac{\mathcal{B}(\mathbf{b}_i)}{\mathcal{B}(\mathbf{q}_i)}. \tag{194}$$

Απομένει η ανάπτυξη της οριακής πιθανοφάνειας εκκίνησης $\pi_{x^1}(x^1)$. Με τον ίδιο τρόπο, επιλέγουμε κατανομή Dirichlet για τις παραμέτρους εκκίνησης $\{a_{0i}\}_{i=1}^{K}$, με υπερπαραμέτρους $\mathbf{q}_0 = [q_1, q_2, \ldots, q_K]$ και έστω $\mathbf{g}_0 = [g_1, g_2, \ldots, g_K]$ η εκκίνηση του συστήματος, όπου τώρα $g_i = \delta(x^1, i)$. Σε απόλυτη αναλογία με τα παραπάνω έχουμε

$$\pi_{x^1}(x^1) = \int_{\mathcal{A}} \prod_{j=1}^{K} a_{0j}^{g_{0j}} \mathcal{D}ir(\mathbf{a}_0|\mathbf{q}_0) d\mathbf{a}_i = \frac{1}{\mathcal{B}(\mathbf{q}_0)} \int_{\mathcal{A}} \prod_{j=1}^{K} a_{0j}^{g_{0j}+q_{0j}-1} d\mathbf{a}_0 \tag{195}$$

δηλαδή

$$\pi_{x^1}(x^1) = \frac{\mathcal{B}(\mathbf{b}_0)}{\mathcal{B}(\mathbf{q}_0)} \tag{196}$$

όπου $\mathbf{b}_0 = [b_{01}, \ldots, b_{0K}]$ και $b_{0j} = q_{0j} + b_{ij} = q_{0j} + \delta(x^1, j)$. Επομένως, καταλήγουμε ότι η εκ των προτέρων πιθανότητα του $\mathbf{x}$ δεδομένου του K καθώς και των υπερπαραμέτρων $\mathbf{q}_i, i = 0, 1, \ldots, K$ δίδεται από τον παρακάτω τύπο

$$\pi_{\mathbf{x}}(\mathbf{x}) = \prod_{i=0}^{K} \frac{\mathcal{B}(\mathbf{b}_i)}{\mathcal{B}(\mathbf{q}_i)} = \prod_{i=0}^{K} \frac{\Gamma(\sum_{j=0}^{K} q_{ij}) \prod_{j=0}^{K} \Gamma(b_{ij})}{\Gamma(\sum_{j=0}^{K} b_{ij}) \prod_{j=0}^{K} \Gamma(q_{ij})}. \tag{197}$$

Ο λογάριθμος της παραπάνω ποσότητας, προστιθέμενος στη λογαριθμική οριακή πιθανοφάνεια των εσωτερικών παραμέτρων, όπως αυτή υπολογίζεται από το τμηματικό BIC συν έναν τελευταίο παράγοντα που εξετάζουμε στη συνέχεια, αποτελεί το προτεινόμενο κριτήριο.

Ενδιαφέρον έχει η σύνδεση της σχέσης (196) με την προσέγγιση κατά Laplace που υιοθετήσαμε για να υπολογίσουμε την οριακή πιθανοφάνεια των εξωτερικών παραμέτρων.

8.3.5 Ο ρόλος των υπερπαραμέτρων

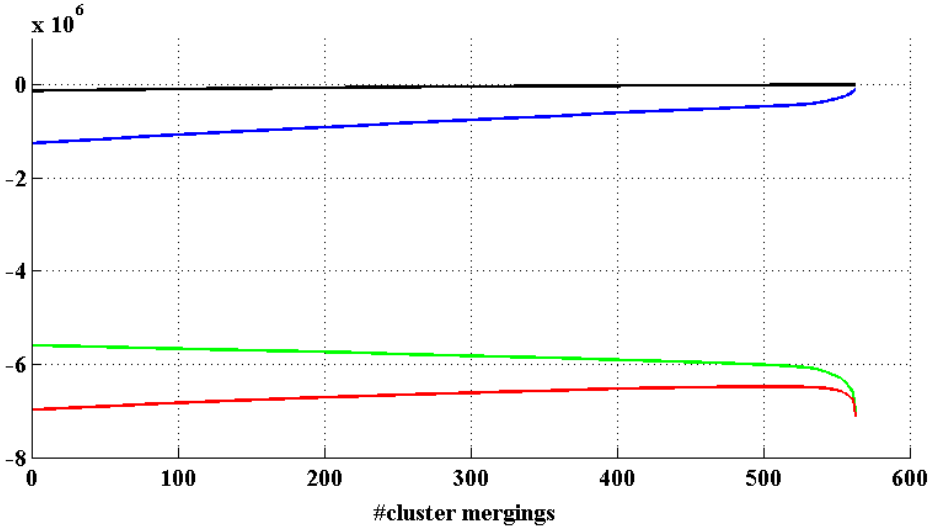
Εξετάζουμε τώρα τον ρόλο των υπερπαραμέτρων $\mathbf{q}_i$ στη παραπάνω μοντελοποίηση. Οι κατανομές Dirichlet αποτελούν έναν συστηματικό τρόπο ενσωμάτωσης στο σύστημα της εκ των προτέρων γνώσης που διαθέτουμε όσον αφορά της ιδιότητες των εξισώσεων συστήματος. Όπως δείξαμε, οι εξισώσεις αυτές διέπονται από έντονα μαρκοβιανά χαρακτηριστικά τα οποία ενδέχεται να διαφέρουν ανάλογα με το πεδίο εφαρμογής. Ένα θεμελιακό κοινό χαρακτηριστικό είναι η τάση του συστήματος για εμμονή στην παρούσα κατάσταση, ιδιότητα η οποία μπορεί να γραφεί ποιοτικά ως $a_{ii} \gg a_{ij}$ για $i \neq j$. Άλλα χαρακτηριστικά μπορεί να είναι η αραιότητα στην διασύνδεση μεταξύ των καταστάσεων, η οποία και πάλι εκφράζεται μέσα από έναν αραιό (sparse) πίνακα μετάβασης. Η ιδιότητα αυτή είναι πιο κατάλληλη για το πεδίο των εκπομπών ειδησεογραφικού χαρακτήρα, στις οποίες η πλειοψηφία των συμμετεχόντων έχει διασύνδεση μόνο με τον ή τους κεντρικούς παρουσιαστές - πιθανώς συν έναν ρεπόρτερ.

Σε αυτή τη φάση, θα μελετήσουμε μόνο την πρώτη ιδιότητα, αυτή της εμμονής στην παρούσα κατάσταση. Ο τρόπος με τον οποίο επιβάλλουμε στο σύστημα την ιδιότητα αυτή είναι η τοποθέτηση υψηλών τιμών στις αντίστοιχες υπερπαραμέτρους, δηλαδή $q_{ii} \gg 1$. Με τον τρόπο αυτόν, η *εκ των υστέρων* πιθανότητα εκτιμήσεων παραμέτρων με την ιδιότητα $a_{ii} \gg a_{ij}$ για $i \neq j$ αυξάνεται

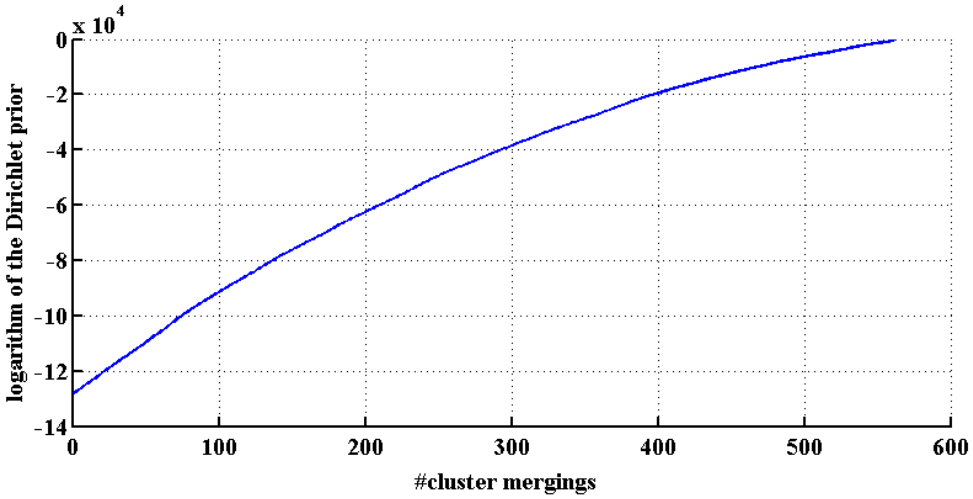
σημαντικά. Αντίθετα, $\mathbf{x}$ τα οποία περιέχουν ταχείες μεταβάσεις μεταξύ των καταστάσεων θα έχουν μικρή εκ των προτέρων πιθανότητα αφού δεν θα ικανοποιούν την παραπάνω ιδιότητα.
Ενδιαφέρον παρουσιάζει επίσης η σύγκριση της συμπεριφοράς του όρου εξισώσεων συστήματος με τον όρο ποινής του τμηματικού BIC. Υποθέτουμε την πρώτη στρατηγική, που συνοψίζεται με τη φράση όλοι οι συμμετέχοντες παίρνουν τον λόγο. Προσθέτοντας τον λογάριθμο του όρου αυτού στο τμηματικό BIC καταλήγουμε στο παρακάτω κριτήριο

$$BIC^S_{sqrt} = l(\hat{\varphi}; \mathbf{x}|\mathbf{y}) - \underbrace{\frac{\alpha}{2} P \sum_{k=1}^{K} \sqrt{n_k} \log n_k}_{\mathcal{T}_{in}} - \underbrace{\sum_{k=0}^{K} \log \frac{\mathcal{B}(\mathbf{q}_k)}{\mathcal{B}(\mathbf{b}_k)}}_{\mathcal{T}_{ex}} \tag{198}$$

Όπως ήδη αναδείξαμε, ο όρος ποινής $\mathcal{T}_{in}$ ευνοεί $\mathbf{x}$ τα οποία παρουσιάζουν μεγάλη διακύμανση στα $\{n_k\}_{k=1}^K$. Η ιδιότητα αυτή προέκυψε από την απαίτησή μας η ΔBIC εξίσωση να είναι αυτόνομη, που μας οδήγησε στη συγκεκριμένη επιλογή των εκ των προτέρων κατανομών των εσωτερικών παραμέτρων $\varphi_k = \{\mu_k, \Sigma_k\}, k = 1, \ldots, K$. Η διακύμανση αυτή έχει όμως και κόστος, αυτό της υπερεκτίμησης του αριθμού των ομάδων. Στον αντίποδα, ο όρος $\mathcal{T}_{ex}$ αποθαρρύνει την διάσπαση μιας ομάδας σε μικρά τμήματα. Για σταθερό n_k, το $\mathcal{T}_{ex}$ αποδίδει τη μικρότερη ποινή όταν τα δείγματα είναι συνεχόμενα ενώ αυξάνεται όσο τα δείγματα αυτά διασπώνται σε μικρότερα τμήματα.



Σχῆμα 30: *Παράδειγμα μεταβολής των όρων από τους οποίους το κριτήριο συντελείται με τις αναδρομές του αλγορίθμου ιεραρχικής ομαδοποίησης. Ο $x-$άξονας δείχνει τον αριθμό των ενώσεων ομάδων δηλαδή τον α/α της αναδρομής. Ο αλγόριθμος ξεκινάει με 563 τμήματα ομιλίας και εντοπίζει 65 ομιλητές (δηλαδή στο $x = 498$). Κόκκινο: Η τιμή του κριτηρίου (μέγιστο στο $x = 498$). Πράσινο: $l(\hat{\varphi}; \mathbf{x}|\mathbf{y})$, Μπλε: $\mathcal{T}_{in}$ και Μαύρο: $-\mathcal{T}_{ex}$*



Σχῆμα 31: *Παράδειγμα μεταβολής του όρου* $-T_{ex}$ *με τις αναδρομές του αλγορίθμου ιεραρχικής ομαδοποίησης. Ο* $x-$*άξονας δείχνει τον αριθμό των συνενώσεων ομάδων δηλαδή τον α/α της αναδρομής. Ο αλγόριθμος ξεκινάει με 563 τμήματα ομιλίας και εντοπίζει 65 ομιλητές. Το διάγραμμα δείχνει την αύξηση της ποσότητας καθώς τα τμήματα ενοποιούνται και δημιουργούν μεγαλύτερες ομάδες.*

9 ΠΑΡΑΡΤΗΜΑ 2: ΒΑΣΙΚΕΣ ΑΡΧΕΣ ΚΑΙ ΘΕΩΡΗΜΑΤΑ ΤΗΣ ΓΕΩΜΕΤΡΙΑΣ ΤΗΣ ΠΛΗΡΟΦΟΡΙΑΣ

Στο παρόν Παράρτημα εξετάζουμε παρουσιάζουμε ορισμένες από τις βασικές αρχές και θεωρήματα της γεωμετρίας της πληροφορίας (Information Geometry). Ο όρος αυτός εισήχθη από τον Shunichi Amari και περικλείει τα απαραίτητα εργαλεία που μας βοηθούν να κατανοήσουμε τη στατιστική μοντελοποίηση, εκτίμηση και συμπερασματολογία με γεωμετρικούς όρους. Συμπεράσματα του Παραρτήματος χρησιμοποιούνται κυρίως στον αλγόριθμο της μετατόπισης του μέσου.

9.1 Το manifold των κατανομών, οι συνδέσεις και η γενικευμένη αρχή της ορθογωνιότητας

Ας θεωρήσουμε μια παραμετρική οικογένεια κατανομών $S = \{p(\mathbf{y}|\boldsymbol{\xi})\}$, όπου $\boldsymbol{\xi} = \{\xi_i\}_{i=1}^{P}$ μία τυχαία παραμετροποίησή της. Η οικογένεια αυτή μπορεί να θεωρηθεί ως ένα $P-$διάστατο manifold, με τις παραμέτρους $\boldsymbol{\xi}$ να ορίζουν ένα σύστημα συντεταγμένων (coordinate system). Για να θεωρήσουμε ένα manifold ως Riemann, θα πρέπει να προσθέσουμε μία δομή (structure), γνωστή ως τανυστή μετρικής (metric tensor). Αποδεικνύεται ότι για ένα ευρύτατο σύνολο κατανομών, ο τανυστής μετρικής ταυτίζεται με την πληροφορία Fisher

$$\mathcal{I}_{i,j}(\boldsymbol{\xi}) = \mathcal{E}_{\boldsymbol{\xi}} \left\{ \frac{\partial \log p(\mathbf{y}|\boldsymbol{\xi})}{\partial \xi_i} \frac{\partial \log p(\mathbf{y}|\boldsymbol{\xi})}{\partial \xi_j} \right\} = \int \frac{\partial \log p(\mathbf{y}|\boldsymbol{\xi})}{\partial \xi_i} \frac{\partial \log p(\mathbf{y}|\boldsymbol{\xi})}{\partial \xi_j} p(\mathbf{y}|\boldsymbol{\xi}) d\mathbf{y} \tag{199}$$

Όταν ο πίνακας $\mathcal{I}(\boldsymbol{\xi})$ είναι θετικά ορισμένος, μπορούμε να ορίσουμε την απειροστή τετραγωνική απόσταση μεταξύ $\boldsymbol{\xi}$ και $\boldsymbol{\xi} + d\boldsymbol{\xi}$ με την ακόλουθη τετραγωνική μορφή

$$d\boldsymbol{\xi}^T \mathcal{I}(\boldsymbol{\xi}) d\boldsymbol{\xi} = \sum_{i,j} \mathcal{I}_{i,j}(\boldsymbol{\xi}) d\xi_i d\xi_j \tag{200}$$

Για μια περιοχή γύρω από το $\boldsymbol{\xi}$, η απόσταση αυτή μπορεί να θεωρηθεί ως η Mahalanobis απόσταση, με πίνακα ακρίβειας τον τανυστή μετρικής $\mathcal{I}(\boldsymbol{\xi})$. Η ιδιότητα αυτή εκφράζει και τη βασικότατη ιδιότητα της διαφορικής γεωμετρίας, σύμφωνα με την οποία για κάθε σημείο του manifold, υπάρχει

μία περιοχή η οποία μπορεί να θεωρηθεί τοπικά Ευκλείδια. Η σημαντική διαφορά με την Ευκλείδια γεωμετρία έγκειται στο ότι ο πίνακας ακρίβειας $\mathcal{I}(\boldsymbol{\xi})$ δεν είναι σταθερός, αλλά γενικά εξαρτάται από το σημείο $\boldsymbol{\xi}$. Όπως θα δείξουμε στη συνέχεια, ένα σύνολο από ευρύτατα χρησιμοποιούμενες στατιστικές αποκλίσεις έχουν την παραπάνω τετραγωνική μορφή ως κυρίαρχο όρο (τετραγωνικό όρο της ανάλυσης Taylor).

Ας δούμε πώς προκύπτει λοιπόν ο τανυστής μετρικής $\mathcal{I}(\boldsymbol{\xi})$. Έστω μία παραμετροποιημένη διαφορίσιμη καμπύλη $\boldsymbol{\xi} = \boldsymbol{\xi}(t)$ στο S, δηλαδή ένα μονοπαραμετρικό υποσύνολο της S. Ορίζουμε το εφαπτομενικό διάνυσμα (tangent vector) $\dot{\boldsymbol{\xi}}(t) = (d/dt)\boldsymbol{\xi}(t)$ της καμπύλης στο σημείο t. Σε μία περιοχή γύρω από το σημείο $\boldsymbol{\xi}_0 = \boldsymbol{\xi}(t_0)$, το διάνυσμα αυτό εκφράζεται με τη βοήθεια των εφαπτομενικών διανυσμάτων $\mathbf{e}_i$, $i = 1, \ldots, P$. Τα διανύσματα αυτά αποτελούν έναν διανυσματικό χώρο (vector space), ο οποίος ορίζεται από τα εφαπτομενικά διανύσματα του συνόλου των διαφορίσιμων καμπυλών οι οποίες διέρχονται από το σημείο $\boldsymbol{\xi}_0$. Ορίζοντας το εφαπτομενικό διάνυσμα $\dot{\boldsymbol{\xi}}(t) = \sum_i \dot{\xi}_i \mathbf{e}_i$ ως

$$\dot{\boldsymbol{\xi}}(t) = \frac{d}{dt} \log p(\mathbf{y}|\boldsymbol{\xi}(t)) = \sum_i \dot{\xi}_i \partial_i \log p(\mathbf{y}|\boldsymbol{\xi}(t)) \tag{201}$$

όπου $\partial_i = \frac{\partial}{\partial \xi_i}$, προκύπτει ότι η βάση με την οποία ορίζουμε τον χώρο αποτελείται (spanned) από τα διανύσματα

$$\mathbf{e}_i = \partial_i \log p(\mathbf{y}|\boldsymbol{\xi}(t)) \tag{202}$$

Τα διανύσματα βάσης $\{\mathbf{e}_i\}_{i=1}^P$ αποτελούν συναρτήσεις των $\mathbf{y}$ και $\boldsymbol{\xi}(t)$ ορίζουν την ευαισθησία της σ.π.π. στο $\mathbf{y}$ σε σχέση με την απειροστή μεταβολή των παραμέτρων $\boldsymbol{\xi}(t) \mapsto \boldsymbol{\xi}(t + dt)$.

Έστω δύο καμπύλες $\boldsymbol{\xi}_1(t)$ και $\boldsymbol{\xi}_2(t)$, οι οποίες διασταυρώνονται στο σημείο t_0, δηλαδή $\boldsymbol{\xi}_1(t_0) = \boldsymbol{\xi}_2(t_0) = \boldsymbol{\xi}_0$. Το εσωτερικό γινόμενο (inner product) των εφαπτομενικών διανυσμάτων τους στο t_0 υπολογίζεται ως εξής

$$< \dot{\boldsymbol{\xi}}_1(t_0), \dot{\boldsymbol{\xi}}_2(t_0) >= \mathcal{E}_{\boldsymbol{\xi}_0} \left\{ \frac{d}{dt} \log p(\mathbf{y}|\boldsymbol{\xi}_1(t_0)) \frac{d}{dt} \log p(\mathbf{y}|\boldsymbol{\xi}_2(t_0)) \right\} = \dot{\boldsymbol{\xi}}_1(t_0)^T \mathcal{I}(\boldsymbol{\xi}_0) \dot{\boldsymbol{\xi}}_2(t_0) \tag{203}$$

Ο ορισμός του εσωτερικού γινομένου μετατρέπει τον διανυσματικό χώρο $\{\mathbf{e}_i\}_{i=1}^P$ σε χώρο Hilbert. Ο ρόλος του τανυστή μετρικής γίνεται φανερός καθώς μας επιτρέπει (α) να μεταφράσουμε της απειροστές μεταβολές $\boldsymbol{\xi}(t) \mapsto \boldsymbol{\xi}(t + dt)$ σε στατιστικές αποκλίσεις και (β) να γενικεύσουμε γεωμετρικές έννοιες όπως γωνίες και αρχή ορθογωνιότητας στον χώρο των στατιστικών μοντέλων.

Έστω, λοιπόν, η απειροστή μεταβολή $\boldsymbol{\xi}(t_0) \mapsto \boldsymbol{\xi}(t_0 + dt)$. Θέτουμε $d\boldsymbol{\xi} = \boldsymbol{\xi}(t_0 + dt) - \boldsymbol{\xi}(t_0)$ και με βάση την (203) ορίζουμε την απόκλιση ως

$$< d\boldsymbol{\xi}, d\boldsymbol{\xi} >_{\boldsymbol{\xi}_0} = d\boldsymbol{\xi}^T \mathcal{I}(\boldsymbol{\xi}_0) d\boldsymbol{\xi} \tag{204}$$

Έτσι, με τη βοήθεια του τανυστή μετρικής αντιστοιχίσαμε την απειροστή μεταβολή $d\boldsymbol{\xi}$ σε απειροστή στατιστική απόκλιση, η οποία είναι γενικά συνάρτηση του $\boldsymbol{\xi}_0$.

Ας εξετάσουμε επίσης την αρχή της ορθογωνιότητας. Για να θεωρηθούν οι δύο καμπύλες ορθογώνιες στο t_0, απαιτείται $< \dot{\boldsymbol{\xi}}_1(t_0), \dot{\boldsymbol{\xi}}_2(t_0) >= 0$. Διαφορίζοντας την ταυτότητα $\int p(\mathbf{y}|\boldsymbol{\xi}) d\mathbf{y} = 1$ ως προς $\boldsymbol{\xi}$ λαμβάνουμε

$$\mathcal{E}_{\boldsymbol{\xi}_0} \left\{ \frac{d}{dt} \log p(\mathbf{y}|\boldsymbol{\xi}(t_0)) \right\} = 0 \tag{205}$$

Ταυτόχρονα, δείξαμε ότι

$$< \dot{\boldsymbol{\xi}}_1(t_0), \dot{\boldsymbol{\xi}}_2(t_0) >= \int_{\mathcal{Y}} \left[\frac{d}{dt} \log p(\mathbf{y}|\boldsymbol{\xi}_1(t_0)) \frac{d}{dt} \log p(\mathbf{y}|\boldsymbol{\xi}_2(t_0)) \right] p(\mathbf{y}|\boldsymbol{\xi}(t_0)) d\mathbf{y} \tag{206}$$

Έτσι, η αρχή της ορθογωνιότητας ισοδυναμεί με το ασυσχέτιστο των δύο εφαπτομενικών διανυσμάτων $\dot{\boldsymbol{\xi}}_1(t_0) = \sum_i \dot{\xi}_{1,i}(t_0) \mathbf{e}_i$ και $\dot{\boldsymbol{\xi}}_2(t_0) = \sum_i \dot{\xi}_{2,i}(t_0) \mathbf{e}_i$ στο σύνολο του δειγματικού χώρου $\mathcal{Y}$, με βάρος $p(\mathbf{y}|\boldsymbol{\xi}(t_0))$.

9.2 Δυικά επίπεδα manifold - οι οικογένειες των εκθετικών κατανομών

9.2.1 Συνδέσεις, γραμμικές συντεταγμένες και γεωδαιτικές

Ορίζουμε ένα manifold ως e-επίπεδο (e-flat), όταν υπάρχει μια παραμετροποίηση $\boldsymbol{\theta}$ η οποία πληροί την παρακάτω σχέση

$$\mathcal{E}_{\boldsymbol{\theta}} \left\{ \frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\mathbf{y}|\boldsymbol{\theta}) \frac{\partial}{\partial\theta_k} \log p(\mathbf{y}|\boldsymbol{\theta}) \right\} = 0 \tag{207}$$

όπου $i, j, k \in \{1, \ldots, P\}$, $\forall \boldsymbol{\theta} \in \Theta$. Αν πληρείται η παραπάνω ιδιότητα, τότε η παραμετροποίηση $\boldsymbol{\theta}$ είναι γραμμική (affine). Αυτό σημαίνει ότι η γεωδαιτική μεταξύ δύο σημείων $\boldsymbol{a}, \boldsymbol{b} \in \Theta$ θα δίνεται από την καμπύλη $\boldsymbol{\theta}(t) = \boldsymbol{a}t + \boldsymbol{b}(1 - t)$. Η ιδιότητα αυτή είναι πολύ σημαντική, καθώς απλοποιεί

ιδιαίτερα τις πράξεις μεταξύ μελών της οικογένειας. Σημειώνεται ότι αν δεν ορίζονται αφινικές συντεταγμένες σε ένα manifold, οι πράξεις αυτές θα πρέπει να λαμβάνεται υπόψη η σύνδεση (connection), η οποία προσδιορίζεται από τους συντελεστές Christophel, Γ_{ijk}. Η δομή αυτή αποτελείται από $P \times P \times P$ βαθμωτούς συντελεστές και ορίζει τόσο τη σύνδεση όσο και τη συμμεταβλητή παράγωγο (covariant derivative), τον τρόπο δηλαδή με τον οποίο διαφορίζονται οι τανυστές άνω του μηδενικού βαθμού. Μηδενικού βαθμού τανυστές είναι τα βαθμωτά πεδία (scalar fields) όπου οι συμμεταβλητή παράγωγος ταυτίζεται με τη μερική παράγωγο.

Αποδεικνύεται ότι την ιδιότητα (207) πληρούν οι εκθετικές οικογένειες, οι οικογένειες δηλαδή με την παρακάτω μορφή

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) \tag{208}$$

Όπως θα αναλυθεί στην επόμενη παράγραφο, οι παράμετροι καλούνται φυσικές ή κανονικές (natural ή cannonical). Η απόδειξη ότι η ιδιότητα (207) πληρείται προκύπτει εύκολα, καθώς

$$\frac{\partial^2}{\partial\theta_i \partial\theta_j} \log p(\mathbf{y}|\boldsymbol{\theta}) = -\frac{\partial^2}{\partial\theta_i \partial\theta_j} \psi(\boldsymbol{\theta}) \tag{209}$$

είναι δηλαδή ανεξάρτητο του $\mathbf{y}$, ενώ $\mathcal{E}_{\boldsymbol{\theta}} \left\{ \frac{\partial}{\partial\theta_i} \log p(\mathbf{y}|\boldsymbol{\theta}) \right\} = 0$. Η γεωδαιτική μεταξύ δύο μελών $\boldsymbol{a}, \boldsymbol{b} \in \Theta$ σε μορφή σ.π.π. θα έχει την ακόλουθη μορφή

$$p(\mathbf{y}|\boldsymbol{\theta}(t)) \propto p(\mathbf{y}|\boldsymbol{a})^t p(\mathbf{y}|\boldsymbol{b})^{1-t} \tag{210}$$

κατάλληλα κανονικοποιημένη ώστε $\int_{\mathcal{Y}} p(\mathbf{y}|\boldsymbol{\theta}(t)) d\mathbf{y} = 1$, είναι επομένως γραμμική ως προς τη λογαριθμική σ.π.π.. Η γεωδαιτική αυτή λέγεται εκθετική γεωδαιτική (exponential ή απλά e-geodesic). Με δυικό τρόπο, ένα manifold ως m-επίπεδο (m-flat), όταν υπάρχει μια παραμετροποίηση $\boldsymbol{\eta}$ η οποία πληροί την παρακάτω σχέση

$$\mathcal{E}_{\boldsymbol{\eta}} \left\{ \frac{1}{p(\mathbf{y}|\boldsymbol{\eta})} \frac{\partial^2}{\partial\eta_i \partial\eta_j} p(\mathbf{y}|\boldsymbol{\eta}) \frac{\partial}{\partial\eta_k} \log p(\mathbf{y}|\boldsymbol{\eta}) \right\} = 0 \tag{211}$$

όπου $i, j, k \in \{1, \ldots, P\}$, $\forall \boldsymbol{\eta} \in \mathbf{H}$. Η οικογένεια των μιγμάτων κατανομών

$$p(\mathbf{y}|\boldsymbol{\eta}) = \sum_{i>0} \eta_i q_i(\mathbf{y}) + (1 - \sum_{i>0} \eta_i) q_0(\mathbf{y}) \tag{212}$$

με ελεύθερες παραμέτρους τα βάρη $\eta_i \geq 0$, $\sum_{i\geq 0} \eta_i = 1$. Λόγω της γραμμικότητας της σ.π.π. ως προς τις παραμέτρους $\boldsymbol{\eta}$, ισχύει ότι $\frac{\partial^2}{\partial \eta_i \partial \eta_j} p(\mathbf{y}|\boldsymbol{\eta}) = 0$, σχέση που αποδεικνύει ότι η συγκεκριμένη οικογένεια πληροί την ιδιότητα (211). Η παραμετροποίηση $\boldsymbol{\eta}$ είναι λοιπόν γραμμική για την οικογένεια των μιγμάτων, και η m-γεωδαιτική μεταξύ δύο μελών της $\tilde{\boldsymbol{a}}, \tilde{\boldsymbol{b}} \in \mathbf{H}$ δίνεται από την εξίσωση

$$p(\mathbf{y}|\boldsymbol{\eta}(t)) = tp(\mathbf{y}|\tilde{\boldsymbol{a}}) + (1-t)p(\mathbf{y}|\tilde{\boldsymbol{b}}) \tag{213}$$

Από παρατήρηση της (212) προκύπτει ωστόσο ένα απλό συμπέρασμα. Ας θεωρήσουμε και πάλι μια οποιαδήποτε εκθετική οικογένεια και ας επιχειρήσουμε να βρούμε αφινικές συντεταγμένες για τη σύνδεση μίγματος. Από τη σχέση (212) προκύπτει καθαρά ότι τέτοιες συντεταγμένες υφίστανται και δεν είναι άλλες από τις αναμενόμενες (expectation) παραμέτρους, τις παραμέτρους δηλαδή που κωδικοποιούν τις επαρκείς στατιστικές $\mathbf{t}(\mathbf{y})$ ενός δείγματος $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^n$, πάνω στην οποία βασίζεται η εκτίμηση μέγιστης πιθανοφάνειας. Πιο συγκεκριμένα

$$\boldsymbol{\eta} = \int_{\mathcal{Y}} \mathbf{t}(\mathbf{y}) p(\mathbf{y}|\boldsymbol{\theta}(\boldsymbol{\eta})) d\mathbf{y} \tag{214}$$

Έστω δύο δείγματα παρατηρήσεων $\mathbf{y}_a$ και $\mathbf{y}_b$, και έστω $\mathbf{t}(\mathbf{y}_a)$ και $\mathbf{t}(\mathbf{y}_b)$ οι αναμενώμενες τιμές των επαρκών στατιστικών τους. Ισχύει ότι

$$\mathbf{t}(\mathbf{y}_a \cup \mathbf{y}_b) = t\mathbf{t}(\mathbf{y}_a) + (1-t)\mathbf{t}(\mathbf{y}_b),\ t = \frac{n_a}{n_a + n_b} \tag{215}$$

όπου $\mathbf{t}(\mathbf{y}_a \cup \mathbf{y}_b)$ το διάνυσμα των επαρκών στατιστικών της ένωσης των δύο συνόλων. Επομένως, η σύνδεση μίγματος έχει την ιδιότητα της προσθετικότητας (additivity) μεταξύ ανεξαρτήτων δειγμάτων από παρατηρήσεις. Πέρα λοιπόν από τις φυσικές παραμέτρους $\boldsymbol{\theta}$, η εκθετική οικογένεια έχει και τις αναμενόμενες παραμέτρους ως αφινικές, αν χρησιμοποιηθεί η m-γεωδαιτική. Πρόκειται λοιπόν για ένα δυικά επίπεδο manifold αφού όπως θα αναδείξουμε στη συνέχεια, οι δύο αυτές παραμετροποιήσεις έχουν αντίστροφους τανυστές μετρικής $\mathcal{I}(\boldsymbol{\theta}) = \mathcal{I}(\boldsymbol{\eta})^{-1}$.

Για να εξετάσουμε σε μεγαλύτερο βάθος τις ιδιότητες των δυικά επίπεδων manifold θα εστιάσουμε στις εκθετικές οικογένειες κατανομών. Θα δείξουμε ότι κάθε οικογένεια κατανομών που είναι e-επίπεδη (m-επίπεδη), είναι αυτομάτως και m-επίπεδη (e-επίπεδη). Έτσι, τόσο οι εκθετικές οικογένειες όσο τα μίγματα κατανομών αποτελούν δυικά επίπεδα manifold. Θα δείξουμε λοιπόν ότι η

επιλογή της σύνδεσης εξαρτάται από (και επομένως υπονοεί) μία συγκεκριμένη επιλογή απόκλισης. Τονίζουμε ότι ενώ θα δώσουμε έμφαση στην κανονική οικογένεια, τα αποτελέσματα γενικεύονται για κάθε εκθετική κατανομή.

9.3 Εκθετικές οικογένειες: Γενική μορφή και βασικές ιδιότητες

Ως εκθετικές οικογένειες ορίζονται μια πληθώρα γνωστών κατανομών, οι οποίες, μπορούν να εκφρασθούν με μια συγκεκριμένη μαθηματική μορφή και ως εκ τούτου να αντιμετωπισθούν με κοινή μεθοδολογία. Η γενική αυτή μορφή έχει σ.π.π. την ακόλουθη

$$p(\mathbf{y}|\boldsymbol{\theta}) = h(\mathbf{y}) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) \tag{216}$$

όπου

$$\psi(\boldsymbol{\theta}) = \log \int_{\mathcal{Y}} \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y})) h(\mathbf{y}) d\mathbf{y} \tag{217}$$

είναι η λογαριθμική κανονικοποιητική σταθερά (γνωστή και ως συνάρτηση *log-partition*) και $h(\mathbf{y})d\mathbf{y}$, $h : \mathcal{Y} \mapsto R^+$ το μέτρο αναφοράς. Το μέτρο αυτό, στην περίπτωση της κανονικής κατανομής με άγνωστη μέση τιμή και άγνωστη μεταβλητότητα είναι σταθερό. Έτσι, μπορεί να απορροφηθεί από την $\psi(\boldsymbol{\theta})$ και να θέσουμε $h(\mathbf{y}) = 1$ στην (216). Επιπλέον, συμβολίζουμε με $\boldsymbol{\theta} = \{\theta_i\}_{i=1}^P$ το P-διάστατο διάνυσμα των *φυσικών παραμέτρων* (*natural parameters*), και με $\mathbf{t}(\mathbf{y})$ το επίσης P-διάστατο διάνυσμα των *επαρκών στατιστικών* (*sufficient statistics*) των $\mathbf{y}$, δηλαδή μία απεικόνιση (mapping) $\mathcal{Y} \mapsto \Re^P$.

Για τη μονοδιάστατη κανονική κανατομή, οι παραπάνω συναρτήσεις έχουν την ακόλουθη μορφή

$$\boldsymbol{\theta} = \left(\frac{\mu}{\sigma^2}, -\frac{1}{2\sigma^2}\right), \mathbf{t}(\mathbf{y}) = \left(y, y^2\right) \tag{218}$$

και

$$\psi(\boldsymbol{\theta}) = \frac{\mu^2}{2\sigma^2} + \frac{1}{2}\log(2\pi\sigma^2) \tag{219}$$

όπου με (μ, σ^2) συμβολίζουμε μέση τιμή και μεταβλητότητα, αντίστοιχα.

Η συνάρτηση $\psi(\boldsymbol{\theta})$ έχει ένα σημαντικό ρόλο να παίζει καθώς με διαφόρισή της ως προς τις φυσικές

παραμέτρους λαμβάνουμε τις παραμέτρους αναμενόμενης τιμής (*expectation parameters*) $\boldsymbol{\eta}(\boldsymbol{\theta})$. Πιο αναλυτικά,

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \left(\mu, \sigma^2 + \mu^2\right) \tag{220}$$

όπου η τελευταία ισότητα αφορά στην κανονική κατανομή. Καλούνται έτσι, καθώς όπως φαίνεται από την παρακάτω σχέση

$$\boldsymbol{\eta}(\boldsymbol{\theta}) = \mathcal{E}_{p(\mathbf{y}|\boldsymbol{\theta})}\{\mathbf{t}(\mathbf{y})\} \tag{221}$$

ισούνται με την αναμενόμενη τιμή των επαρκών στατιστικών $\mathbf{t}(\mathbf{y})$, με σ.π.π. αναφοράς την $p(\mathbf{y};\boldsymbol{\theta})$. Το γεγονός αυτό είναι τεράστιας σημασίας, καθώς αποδεικνύει ότι η οικογένεια των εκθετικών κατανομών επιδέχεται σύμπτυξης της πληροφορίας που φέρει το δείγμα $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^n$ σε ένα διάνυσμα P-διαστάσεων (το $\boldsymbol{\eta}$), όπου τυπικά $n \gg P$. Αν επιπλέον η πραγματική κατανομή που γεννά τα δείγματα $\mathbf{y} = \{\mathbf{y}^{(i)}\}_{i=1}^n$ ανήκει στον χώρο των εξεταζομένων πιθανοτικών μοντέλων, η σύμπτυξη αυτή πραγματοποιείται με μηδενική απώλεια πληροφορίας (Information loss).
Επιπλέον, η παράγωγος δεύτερης τάξης οδηγεί στον πίνακα πληροφορίας του Fisher των φυσικών παραμέτρων

$$\mathcal{I}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\boldsymbol{\eta} \tag{222}$$

που ισούνται με

$$\mathcal{I}(\boldsymbol{\theta}) = \begin{pmatrix} \sigma^2 & 2\mu\sigma^2 \\ 2\mu\sigma^2 & 4\mu^2\sigma^2 + 2\sigma^4 \end{pmatrix} \tag{223}$$

για τη μονοδιάστατη κανονική κατανομή. Προσδιορίζει το κατώτερο όριο κάθε μη-πολωμένου εκτιμητή $T(\mathbf{y}) \mapsto \hat{\boldsymbol{\eta}}$ της $\boldsymbol{\eta}$ που αντίστοιχεί σε δείγμα μίας και μόνο παρατήρησης.

Απόδειξη της σχέσης $\boldsymbol{\eta}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}}\psi(\boldsymbol{\theta})$

Έχουμε ότι

$$\psi(\boldsymbol{\theta}) = \log \int_{\mathcal{Y}} \exp(\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{y}))h(\mathbf{y})d\mathbf{y} \tag{224}$$

Επομένως

$$\frac{\partial\psi(\boldsymbol{\theta})}{\partial\boldsymbol{\theta}} = \frac{\int_{\mathcal{Y}}\exp(\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{y}))\mathbf{t}(\mathbf{y})h(\mathbf{y})d\mathbf{y}}{\int_{\mathcal{Y}}\exp(\boldsymbol{\theta}\cdot\mathbf{t}(\mathbf{y}))h(\mathbf{y})d\mathbf{y}} = \mathcal{E}_{p(\mathbf{y};\boldsymbol{\theta})}\{\mathbf{t}(\mathbf{y})\} = \boldsymbol{\eta}(\boldsymbol{\theta}) \tag{225}$$

Απόδειξη της σχέσης $\mathcal{I}(\boldsymbol{\theta}) = \nabla_{\boldsymbol{\theta}} \nabla_{\boldsymbol{\theta}} \psi(\boldsymbol{\theta})$

Έχουμε ότι

$$\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \frac{\partial}{\partial \boldsymbol{\theta}} \int_{\mathcal{Y}} \mathbf{t}(\mathbf{y}) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) h(\mathbf{y}) d\mathbf{y} \tag{226}$$

Επομένως

$$\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \int_{\mathcal{Y}} \mathbf{t}(\mathbf{y})(\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta})) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) h(\mathbf{y}) d\mathbf{y} \tag{227}$$

Επιπλέον, ισχύει ότι

$$\int_{\mathcal{Y}} \boldsymbol{\eta}(\boldsymbol{\theta})(\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta})) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) h(\mathbf{y}) d\mathbf{y} = \boldsymbol{\eta}(\boldsymbol{\theta}) \int_{\mathcal{Y}} (\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta})) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) h(\mathbf{y}) d\mathbf{y} = \mathbf{0} \tag{228}$$

αφού $\mathcal{E}_{p(\mathbf{y};\boldsymbol{\theta})}\{\mathbf{t}(\mathbf{y})\} = \boldsymbol{\eta}(\boldsymbol{\theta})$. Αφαιρώντας την (228) από την (227) λαμβάνουμε

$$\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T} = \int_{\mathcal{Y}} (\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta})) \cdot (\mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta})) \exp(\boldsymbol{\theta} \cdot \mathbf{t}(\mathbf{y}) - \psi(\boldsymbol{\theta})) h(\mathbf{y}) d\mathbf{y} \tag{229}$$

Έτσι, η $\frac{\partial^2 \psi(\boldsymbol{\theta})}{\partial \boldsymbol{\theta} \partial \boldsymbol{\theta}^T}$ ισούται με τον πίνακα συμμεταβλητότητας της εκτίμησης των παραμέτρων $\boldsymbol{\eta}$, για μοναδιαίο δείγμα. Ταυτόχρονα, είναι και ο πίνακας ακρίβειας (αντίστροφος της συμμεταβλητότητας), δηλαδή ο πίνακας πληροφορίας Fisher ως προς τις παραμέτρους $\boldsymbol{\theta}$, αφού

$$\frac{\partial \log p(\mathbf{y}|\boldsymbol{\theta})}{\partial \boldsymbol{\theta}} = \mathbf{t}(\mathbf{y}) - \boldsymbol{\eta}(\boldsymbol{\theta}) \tag{230}$$

Μία επίσης χρήσιμη παρατήρηση που εξάγεται από την (222) είναι ότι ο πίνακας $\mathcal{I}(\boldsymbol{\theta})$ ισούται με την Ιακωβιανή (Jacobian) του μετασχηματισμού $\boldsymbol{\eta} \mapsto \boldsymbol{\theta}$, αφού $\{\mathcal{I}(\boldsymbol{\theta})\}_{ij} = \frac{\partial \boldsymbol{\eta}_i}{\partial \boldsymbol{\theta}_j}$. Χρησιμοποιώντας ορολογία γεωμετρίας της πληροφορίας, ο πίνακας $\mathcal{I}(\boldsymbol{\theta})$ είναι ο τανυστής μετρικής πάνω στο manifold των παραμέτρων, [98].

9.3.1 Ο δυισμός μεταξύ των συστημάτων συντεταγμένων

Μία σημαντική ιδιότητα της εξεταζόμενης οικογένειας κατανομών είναι ο δυισμός των δύο παραμετρικών χώρων. Όπως θα δείξουμε, μπορούμε να θεωρήσουμε τις παραμέτρους $\boldsymbol{\eta}$ ως το δυικό σύστημα συντεταγμένων των $\boldsymbol{\theta}$. Η ιδιότητα αυτή προκύπτει από την κυρτότητα της $\psi(\boldsymbol{\theta})$ ως προς

τις $\boldsymbol{\theta}$ και η οποία δημιουργεί μεταξύ των δύο παραμετρικών χώρων μια αμφιμονοσήμαντη αντιστοίχιση (one-to-one mapping) και εγγυάται ότι ο $\mathcal{I}(\boldsymbol{\theta})$ είναι θετικά ορισμένος. Για να αναδείξουμε τη δυικότητα αυτή, ορίζουμε τη δυική συνάρτηση δυναμικού

$$\phi(\boldsymbol{\eta}) = -\frac{1}{2}\log(2\pi e\sigma^2) \tag{231}$$

Εντελώς συμμετρικά με την $\psi(\boldsymbol{\theta})$, η $\phi(\boldsymbol{\eta})$ γεννά τις φυσικές παραμέτρους ως εξής

$$\boldsymbol{\theta}(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}}\phi(\boldsymbol{\eta}) \tag{232}$$

Η δυική αυτή συνάρτηση δυναμικού ισούται με την (αρνητική) εντροπία του Shannon της κατανομής. Όπως και την περίπτωση της (222), αποδεικνύεται ότι ο πίνακας

$$\mathcal{I}(\boldsymbol{\eta}) = \mathcal{I}(\boldsymbol{\theta})^{-1} = \nabla_{\boldsymbol{\eta}}\nabla_{\boldsymbol{\eta}}\phi(\boldsymbol{\eta}) = \nabla_{\boldsymbol{\eta}}\boldsymbol{\theta} \tag{233}$$

είναι ο πίνακας πληροφορίας του Fisher. Η σχέση αυτή είναι ιδιαίτερα σημαντική. Έστω δύο εφαπτομενικά διανύσματα $\dot{\boldsymbol{\xi}}_1(t_0) = \sum_i \dot{\xi}_{1,i}(t_0)\mathbf{e}_i$ και $\dot{\boldsymbol{\xi}}_2(t_0) = \sum_i \dot{\xi}_{2,i}(t_0)\mathbf{e}_i$ στην τυχαία παραμετροποίηση $\boldsymbol{\xi}$. Ας ορίσουμε το πρώτο στην παραμετροποίηση $\boldsymbol{\theta}$ ως $\dot{\boldsymbol{\theta}}_1(t_0) = \sum_i \dot{\theta}_{1,i}(t_0)\mathbf{e}_i^{\boldsymbol{\theta}}$ και το δεύτερο στην $\boldsymbol{\eta}$ ως $\dot{\boldsymbol{\eta}}_2(t_0) = \sum_i \dot{\eta}_{2,i}(t_0)\mathbf{e}_i^{\boldsymbol{\eta}}$, θα λάβουμε

$$< \dot{\boldsymbol{\theta}}_1(t_0), \dot{\boldsymbol{\eta}}_2(t_0) >= \sum_i \dot{\theta}_{1,i}(t_0)\dot{\eta}_{2,i}(t_0) \tag{234}$$

Ο δυισμός λοιπόν των συστημάτων συντεταγμένων $\{\mathbf{e}_i^{\boldsymbol{\theta}}\}$ και $\{\mathbf{e}_i^{\boldsymbol{\eta}}\}$ συνεπάγεται

$$< \mathbf{e}_i^{\boldsymbol{\theta}}, \mathbf{e}_j^{\boldsymbol{\eta}} >= \delta(i,j) \tag{235}$$

σχέση ή οποία μας επιτρέπει να εκφράζουμε τις στατιστικές αποκλίσεις χωρίς την ανάγκη υπολογισμού των διανυσμάτων βάσης ή του τανυστή μετρικής.

Όπως αναφέραμε, η ύπαρξη της αμφιμονοσήμαντης αντιστοιχίας είναι αποτέλεσμα της κυρτότητας της $\psi(\boldsymbol{\theta})$ ως προς τις $\boldsymbol{\theta}$ και ο μετασχηματισμός που τις συνδέει καλείται *Legendre*. Ορίζεται ως εξής

$$\phi(\boldsymbol{\eta}) = \max_{\boldsymbol{\theta}}\{\boldsymbol{\theta}\cdot\boldsymbol{\eta} - \psi(\boldsymbol{\theta})\} \tag{236}$$

και στη δυική του μορφή

$$\psi(\boldsymbol{\theta}) = \max_{\boldsymbol{\eta}}\{\boldsymbol{\theta}\cdot\boldsymbol{\eta} - \phi(\boldsymbol{\eta})\} \tag{237}$$

όπου οι συναρτήσεις δυναμικού ικανοποιούν τη σχέση

$$\psi(\boldsymbol{\theta}) + \phi(\boldsymbol{\eta}) = \boldsymbol{\theta} \cdot \boldsymbol{\eta}. \tag{238}$$

Βασισμένοι στις παραπάνω στατιστικές ποσότητες, μπορούμε να εστιάσουμε στο παράδειγμα της πολυμεταβλητής κανονικής κατανομής. Οι φυσικές και οι αναμενόμενες παράμετροι θα έχουν ως εξής

$$\boldsymbol{\theta} = \left(\Sigma^{-1}\mu,\, -\frac{1}{2}\Sigma^{-1}\right), \boldsymbol{\eta} = \left(\mu,\, \Sigma + \mu\mu^T\right) \tag{239}$$

ενώ οι συναρτήσεις δυναμικού

$$\psi(\boldsymbol{\theta}) = \frac{1}{2}\mu^T\Sigma^{-1}\mu + \frac{1}{2}\log((2\pi)^d|\Sigma|) \tag{240}$$

και

$$\phi(\boldsymbol{\eta}) = -\frac{1}{2}\log((2\pi e)^d|\Sigma|). \tag{241}$$

Επισημαίνουμε ότι τα δεύτερα μέλη των παραμέτρων είναι πίνακες, δηλαδή $\boldsymbol{\theta} = (\boldsymbol{\theta}_1, \boldsymbol{\Theta}_2)$ και $\boldsymbol{\eta} = (\boldsymbol{\eta}_1, \boldsymbol{H}_2)$, όπου $\boldsymbol{\theta}_1, \boldsymbol{\eta}_1 \in \Re^d$ και $\boldsymbol{\Theta}_2, \boldsymbol{H}_2 \in \Re^{d\times d}$, ενώ P ισούται με τη διάσταση του (μ, Σ), δηλαδή $P = d + d(d+1)/2$. Τέλος, το εσωτερικό γινόμενο μεταξύ $\boldsymbol{\theta}$ και $\boldsymbol{\eta}$ υπολογίζεται ως εξής

$$\boldsymbol{\theta} \cdot \boldsymbol{\eta} = Tr\{\boldsymbol{\theta}_1\boldsymbol{\eta}_1^T + \boldsymbol{\Theta}_2\boldsymbol{H}_2^T\} = \boldsymbol{\theta}_1^T\boldsymbol{\eta}_1 + Tr\{\boldsymbol{\Theta}_2\boldsymbol{H}_2^T\} \tag{242}$$

όπου με $|A|$, A^T και $Tr\{A\}$ εννοούμε ορίζουσα, ανάστροφο και ίχνος του A, αντίστοιχα.

9.4 Η οικογένεια των δ-αποκλίσεων

Αναφερθήκαμε παραπάνω σε δύο τύπους συνδέσεων, την εκθετική και την σύνδεση μίγματος. Θα δείξουμε τώρα ότι οι συνδέσεις αυτές αποτελούν τα δύο άκρα μίας οικογένειας συνδέσεων και είναι συνυφασμένες με την απόκλιση που επιθυμούμε να χρησιμοποιήσουμε. Τα δύο αυτά άκρα αντιστοιχούν στις αποκλίσεις Kullback-Leibler, αποκλίσεις που με τη σειρά τους αποτελούν τα δύο άκρα της συγκεκριμένης οικογένειας αποκλίσεων. Για να γίνουμε πιο ακριβείς, τα μέλη τόσο των συνδέσεων όσο και των αποκλίσεων παραμετροποιούνται με μία παράμετρο $\delta \in [0, 1]$. Η

παράμετρος αυτή ορίζει την απόκλιση και ώς εκ τούτου τη σύνδεση που θα χρησιμοποιήσουμε. Θα αναδείξουμε τέλος τη υποκείμενη γεωμετρία πάνω στην οποία βασίζεται η Μπεϋσιανή στατιστική, συνδέοντας τις εκ των προτέρων πιθανότητες με τις παραπάνω αποκλίσεις.

Ξεκινούμε την ανάλυσή μας με την οικογένεια των αποκλίσεων που θα χρησιμοποιήσουμε. Ο παρακάτω συμβολισμός είναι γενικός καθότι αναφέρεται στο σύνολο των πεπερασμένων θετικών μέτρων, δηλαδή $p, q \in \tilde{\mathcal{P}}$ υποσύνολο του οποίου αποτελούν τα πιθανοτικά μέτρα, $\mathcal{P} \subset \tilde{\mathcal{P}}$. Ορίζουμε τη δ-απόκλιση ως εξής

$$D_\delta(p||q) = \begin{cases} \frac{\int p d\mathbf{y}}{1-\delta} + \frac{\int q d\mathbf{y}}{\delta} - \frac{\int p^\delta q^{1-\delta} d\mathbf{y}}{\delta(1-\delta)}, \text{ εάν } \delta \in (0,1) \\ \int q - p + p \log(\frac{p}{q}) d\mathbf{y}, \text{ εάν } \delta = 1 \\ \int p - q + q \log(\frac{q}{p}) d\mathbf{y}, \text{ εάν } \delta = 0 \end{cases} \tag{243}$$

Σημειώνεται ότι $\frac{p^\delta}{\delta} \to \log p$, $\delta \to 0$. Για $\delta = 1$ έχουμε την Kullback-Leibler απόκλιση, ενώ για $\delta = 0$ έχουμε και πάλι την Kullback-Leibler απόκλιση, αλλά με αντίστροφα ορίσματα. Επιπλέον, η μόνη συμμετρική απόκλιση είναι η περίπτωση $\delta = 1/2$, όπου

$$D_{1/2}(p||q) = 2 \int (\sqrt{p} - \sqrt{q})^2 \, d\mathbf{y} \tag{244}$$

και είναι η τετραγωνική απόσταση Hellinger επί δύο.

Οι παραπάνω οικογένεια αποκλίσεων είναι η μόνη οικογένεια με μία σειρά ελκυστικών ιδιοτήτων

1. **Ομογένεια:** $D_\delta(cp||cq) = cD_\delta(p||q)$
2. **Θετικά ορισμένη:** $D_\delta(p||q) \geq 0$, με ισότητα αν και μόνο αν $q = p$
3. **Δυικότητα:** $D_\delta(p||q) = D_{1-\delta}(q||p)$
4. **Αμετάβλητη σε μετασχηματισμούς:** Έστω $\mathcal{T} : \mathcal{X} \mapsto \mathcal{Y}$, τότε $D_\delta(p||q) = D_\delta(p_\mathcal{T}||q_\mathcal{T})$, όπου $p_\mathcal{T} = p \circ \mathcal{T}^{-1}$
5. **Μοναδικότητα:** Η οικογένεια είναι η μοναδική που πληροί το σύνολο των παραπάνω ιδιοτήτων.

6. **Ανάπτυγμα Taylor:** Η απόκλιση $D_\delta(\boldsymbol{\theta}+\epsilon\boldsymbol{\lambda}||\boldsymbol{\theta})$ έχει ως ανάπτυγμα

$$D_\delta(\boldsymbol{\theta}+\epsilon\boldsymbol{\lambda}||\boldsymbol{\theta}) = \frac{1}{2}\sum_{i,j}\mathcal{I}(\boldsymbol{\theta})_{i,j}\lambda^i\lambda^j\epsilon^2 + \frac{1}{6}\sum_{i,j,k}[\overset{0}{\Gamma}_{ijk} + \overset{\delta}{\Gamma}_{kij} + \overset{1}{\Gamma}_{jki}]\lambda^i\lambda^j\lambda^k\epsilon^3 + o(\epsilon^3) \quad (245)$$

όπου $\overset{\delta}{\Gamma}_{ijk} = \int p[\partial_i\partial_j l + \delta\partial_i l\partial_j l]\partial_k l d\mathbf{y}$ τα σύμβολα Christoffel της δ-γεωμετρίας.

7. **Σχέσεις Eguchi:**

$$\mathcal{I}(\boldsymbol{\theta})_{i,j} = -\partial_i\partial_j D_\delta(p_\theta||p_{\theta'})|_{\theta=\theta'},\ \ \overset{\delta}{\Gamma}_{ijk} = -\partial_i\partial_j\partial_k D_\delta(p_\theta||p_{\theta'})|_{\theta=\theta'} \quad (246)$$

8. **Νόμος γενικευμένων συνημιτόνων:**

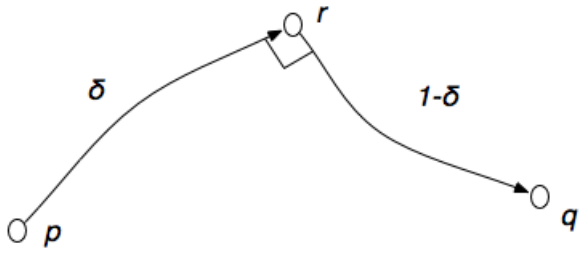
$$D_\delta(p||r) + D_\delta(r||q) = D_\delta(p||q) + \int (l_\delta(p) - l_\delta(r))(l_{1-\delta}(q) - l_{1-\delta}(r))d\mathbf{y} \quad (247)$$

από την οποία προκύπτει η παρακάτω ιδιότητα

9. **Γενικευμένο Πυθαγόριο θεώρημα:** Αν η δ-γεωδαιτική που συνδέει την p με την r είναι ορθογώνια στην r με την $(1-\delta)$-γεωδαιτική που συνδέει την r με την q τότε το ολοκλήρωμα μηδενίζεται, και ως εκ τούτου ισχύει

$$D_\delta(p||r) + D_\delta(r||q) = D_\delta(p||q) \quad (248)$$

σχέση η οποία καλείται Γενικευμένο Πυθαγόριο θεώρημα.



Σχῆμα 32: *Γενικευμένο Πυθαγώριο θεώρημα λόγω ορθογωνιότητας των γεωδαιτικών*

Από τις παραπάνω αποκλίσεις εστιάζουμε στις Kullback-Leibler ($\delta = \{0, 1\}$) καθώς και στην Hellinger ($\delta = 1/2$).

9.5 Η φυσική κλίση και ο ρόλος της στην εκμάθηση

9.5.1 Εισαγωγή και ιστορική αναδρομή

Κλείνουμε την περιγραφή των βασικών ιδιοτήτων της γεωμετρίας της πληροφορίας με τη φυσική κλίση (Natural Gradient). Η φυσική κλίση βασίζεται στην παρατήρηση ότι σε έναν μη-ευκλείδιο χώρο, η κατευθυντική παράγωγος (directional derivative) δεν αντιστοιχεί στην κατεύθυνση μέγιστης αύξησης της αντικειμενικής συνάρτησης. Αντίθετα, η κατεύθυνση αυτή δίνεται από τη φυσική κλίση, εφόσον η γεωμετρία του προβλήματος είναι γνωστή και υπακούει στις ιδιότητες του Riemann. Όπως απέδειξε ο S.-I. Amari, τα εκθετικά μοντέλα πληρούν τις ιδιότητες αυτές και συγκεκριμένα έχουν τον πίνακα πληροφορίας του Fisher ως τανυστή μετρικής (metric tensor), [99]. Ως εκ τούτου, ένα σύνολο προβλημάτων βελτιστοποίησης που απαιτεί αναδρομικούς αλγορίθμους (λόγω έλλειψης κλειστής φόρμουλας υπολογισμού) θα πρέπει να λαμβάνουν υπόψη τις συγκεκριμένες ιδιότητες του χώρου. Χαρακτηριστικό προβλήμα στο οποίο η φυσική κλίση βελτίωσε θεαματικά την ταχύτητα σύγκλισης των αλγορίθμων είναι ο τυφλός διαχωρισμός πηγών (Blind Source Separation) μέσω Ανάλυσης σε Ανεξάρτητες Συνιστώσες (Independent Component Analysis, ICA). Ο Amari απέδειξε ότι ο πίνακας αποσύζευξης υπακούει σε μη-Ευκλείδια γεωμετρία και ως εκ τούτου η φυσική κλίση θα πρεπει να αντικαταστήσει την απλή κλίση, [100]. Η φυσική κλίση χρησιμοποιείται σε πληθώρα άλλων προβλημάτων βελτιστοποίησης, όπως σε εκμάθηση τύπου Variational Bayes και νευρωνικών δικτύων τύπου Multilayer Perceptron, με θεαματική αύξηση της αποτελεσματικότητας των αλγορίθμων, [101], [102].

9.5.2 Η φυσική κλίση

Ερχόμαστε λοιπόν να εξετάσουμε πώς προκύπτει η φυσική κλίση. Έστω $S = \{\mathbf{w} \in \Re^k\}$ ένας k-διάστατος παραμετρικός χώρος επί του οποίου ορίζεται μία συνάρτηση $L(\mathbf{w})$. Όταν ο S είναι ο Ευκλείδιος χώρος με ένα ορθοκανονικό σύστημα συντεταγμένων $\mathbf{w}$, το τετραγωνικό μήκος μίας

απειροστής αύξησης $d\mathbf{w}$ δίδεται από τον τύπο

$$|d\mathbf{w}|^2 = \sum_{i=1}^{k} (dw_i)^2 = d\mathbf{w}^T d\mathbf{w} \tag{249}$$

όπου $d\mathbf{w} = \{dw_i\}_{i=1}^k$. Όταν ωστόσο το σύστημα συντεταγμένων δεν είναι ορθοκανονικό, το τετραγωνικό μήκος δίδεται από την τετραγωνική μορφή

$$|d\mathbf{w}|^2 = \sum_{i,j} \mathcal{I}(\mathbf{w})_{i,j} dw_i dw_j = d\mathbf{w}^T \mathcal{I}(\mathbf{w}) d\mathbf{w}. \tag{250}$$

Όταν λοιπόν βρισκόμαστε σε χώρο Riemann το μήκος αυτό δίδεται από τη σχέση (250). Είναι προφανές ότι ο Ευκλείδιος χώρος χαρακτηρίζεται από σταθερό (ανεξάρτητο του $\mathbf{w}$) τένσορα μετρικής, ο οποίος μπορεί να γίνει ο μοναδιαίος πίνακας με έναν απλό γραμμικό μετασχηματισμό.
Η κατεύθυνση της πίο απότομης καθόδου (steepest descent) της συνάρτησης $L(\mathbf{w})$ ορίζεται από το διάνυσμα $d\mathbf{w}$ το οποίο ελαχιστοποιεί την $L(\mathbf{w} + d\mathbf{w})$, όπου $|d\mathbf{w}|$ είναι σταθερό, δηλαδή με περιορισμό

$$|d\mathbf{w}|^2 = \epsilon^2 \tag{251}$$

για μία μικρή τιμή του ϵ.

Θεώρημα

Η κατεύθυνση της πιο απότομης καθόδου της συνάρτησης $L(\mathbf{w})$ σε έναν χώρο Riemann δίνεται από τον τύπο

$$-\tilde{\nabla} L(\mathbf{w}) = -\mathcal{I}^{-1}(\mathbf{w}) \nabla L(\mathbf{w}) \tag{252}$$

όπου $\nabla L(\mathbf{w})$ η συνήθης κλίση, δηλαδή

$$\nabla L(\mathbf{w}) = \left(\frac{\partial}{\partial w_1} L(\mathbf{w}), \ldots, \frac{\partial}{\partial w_k} L(\mathbf{w}) \right)^T. \tag{253}$$

Απόδειξη

Θέτουμε $d\mathbf{w} = \epsilon \mathbf{u}$ και αναζητούμε το $\mathbf{u} \in \Re^k$ το οποίο ελαχιστοποιεί την ποσότητα

$$L(\mathbf{w} + d\mathbf{w}) = L(\mathbf{w}) + \epsilon \nabla L(\mathbf{w})^T \mathbf{u} \tag{254}$$

υπό τον ακόλουθο περιορισμό

$$|\mathbf{u}|^2 = \mathbf{u}^T \mathcal{I}(\mathbf{w})\mathbf{u} = 1 \tag{255}$$

Με χρήση της μεθόδου Lagrange έχουμε

$$\frac{\partial}{\partial u_i}\left(\nabla L(\mathbf{w})^T \mathbf{u} - \lambda \mathbf{u}^T \mathcal{I}(\mathbf{w})\mathbf{u}\right) = 0. \tag{256}$$

Η παραπάνω σχέση δίνει

$$\mathbf{u} = \frac{1}{2\lambda}\mathcal{I}(\mathbf{w})^{-1}\nabla L(\mathbf{w}) \tag{257}$$

όπου η σταθερά λ προκύπτει από τον περιορισμό (255).

Ορίζουμε λοιπόν τη φυσική κλίση της $L(\mathbf{w})$ σε έναν χώρο Riemann ως

$$\tilde{\nabla} L(\mathbf{w}) = \mathcal{I}(\mathbf{w})^{-1}\nabla L(\mathbf{w}) \tag{258}$$

και όπως αποδείξαμε αντιπροσωπεύει την κατεύθυνση της πιο απότομης καθόδου.

Έτσι, ο αλγόριθμος που βασίζεται στην πιο απότομη κάθοδο θα πρέπει να τροποποιηθεί ως ακολούθως

$$\mathbf{w}^{t+1} = \mathbf{w}^t - n^t \tilde{\nabla} L(\mathbf{w}^t) \tag{259}$$

όπου n^t ο ρυθμός εκμάθησης (learning rate) που ορίζει το μέγεθος του βήματος.

References

[1] T. Stafylakis, V. Katsouros, and G. Carayannis, "Efficient combination of parametric spaces, models and metrics for speaker diarization," in *Proceedings of IEEE Workshop on Automatic Speech Recognition and Understanding, ASRU*, 2007.

[2] T. Stafylakis, V. Katsouros, and G. Carayannis, "Redefining the Bayesian Information Criterion for speaker diarisation," in *Proceedings of Interspeech*, September 2009.

[3] T. Stafylakis, G. Tzimiropoulos, V. Katsouros, and G. Carayannis, "A new penalty term for the BIC with respect to speaker diarization," in *Proceedings of ICASSP*, 2010, pp. 4978–4981.

[4] T. Stafylakis, V. Katsouros, and G. Carayannis, "The Segmental Bayesian Information Criterion and its applications to Speaker Diarization," *IEEE Selected topics in Signal Processing*, pp. 857 – 866, October 2010.

[5] T. Stafylakis and X. Anguera, "Improvements to the equal-parameter BIC for speaker diarization," in *Proc. Interspeech'10*, September 2010, pp. 314 – 317.

[6] T. Stafylakis, X. Anguera, V. Katsouros, and G. Carayannis, "Closed-form expressions vs. BIC: A comparison for speaker clustering," in *Proceedings of ICASSP*, May 2011.

[7] T. Stafylakis, V. Katsouros, and G. Carayannis, "Speaker clustering via the mean shift algorithm," in *Proc. Speaker Odyssey'10*, July 2010.

[8] A. P. Dempster, N. M. Laird, and D. B. Rubin, "Maximum likelihood from incomplete data via the EM algorithm," *Journal of the Royal Statistical Society Ser. B*, vol. 39, 1977.

[9] C. Fraley and A.E. Laftery, "How many clusters? Which clustering method? Answers via Model-Based Cluster Analysis," *The Computer Journal*, vol. 41, no. 8, 1998.

[10] Guillaume Lathoud and Iain A. Mccowan, "Location based speaker segmentation," in *Proceedings of the 2003 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-03), Hong Kong*, 2003, pp. 176–179.

[11] Guillaume Lathoud, Iain A. McCowan, and Jean-Marc Odobez, "Unsupervised Location-Based Segmentation of Multi-Party Speech," in *Proceedings of the 2004 ICASSP-NIST Meeting Recognition Workshop*, 2004, IDIAP-RR 04-14.

[12] J. Ajmera, G. Lathoud, and I. McCowan, "Clustering And Segmenting Speakers And Their Locations In Meetings," Tech. Rep., LIDIAP, 2003.

[13] U. Anliker, J. F. Randall, and G. Tröster, "Speaker separation and tracking system," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–14, 2006.

[14] V. Digalakis, D. Rtischev, L. Neumeyer, and Edics Sa, "Speaker adaptation using constrained estimation of Gaussian mixtures," *IEEE Transactions on Speech and Audio Processing*, vol. 3, pp. 357–366, 1995.

[15] V. Digalakis, P. Monaco, H. Murveit, and Vassilios Digalakis, "Genones: Generalized mixture tying in continuous hidden Markov model-based speech recognizers," *IEEE Transactions on Speech and Audio Processing*, vol. 4, pp. 281–289, 1996.

[16] L. Welling, R. Haeb-Umbach, X. Aubert, and N. Haberland, "A study on speaker normalization using vocal tract normalization and speaker adaptive training," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, 1998, pp. 797–800.

[17] D. Sündermann, A. Bonafonte, H. Höge, and H. Ney., "Time domain vocal tract length normalization," in *Proceesings of IEEE International Symposium on Signal Processing and Information Technology (ISSPIT), Rome, Italy*, December 2004, pp. 191–194.

[18] Giulia Garau Steve and Steve Renals, "Applying Vocal Tract Length Normalization to meeting recordings," in *Proceedings of European Conference on Speech Communication and Technology (Interspeech)*, 2005.

[19] Ke Chen, "Towards better making a decision in speaker verification," *Pattern Recognition, Elsevier*, vol. 36, no. 2, pp. 329–346, February 2003.

[20] H. Winston, H.-M. Hsu, and Shih-Fu Chang, "A statistical framework for fusing mid-level perceptual features in news story segmentation," in *ICME '03: Proceedings of the 2003 International Conference on Multimedia and Expo*, Washington, DC, USA, 2003, pp. 413–416, IEEE Computer Society.

[21] Winston Hsu, Shih-Fu Chang, Chih-Wei Huang, Lyndon Kennedy, Ching-Yung Lin, and Giridharan Iyengar, "Discovery and fusion of salient multi-modal features towards news story segmentation," in *SPIE Symposium on Electronic Imaging: Science and Technology - SPIE Storage and Retrieval of Image/Video Database*, San Jose, CA, January 2004.

[22] Vincent Della Pietra, Vincent Della Pietra, and John Lafferty, "Inducing Features of Random Fields," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 19, pp. 380–393, 1995.

[23] J. Ajmera, H. Bourlard, I. Lapidot, and I. Mccowan, "Unknown-Multiple Speaker clustering using HMM," in *Proceedings of ICSLP-2002*, 2002, pp. 573–576.

[24] Marina Meilă, "Comparing clusterings: an axiomatic view," in *ICML '05: Proceedings of the 22nd international conference on Machine learning*, New York, NY, USA, 2005, pp. 577–584, ACM.

[25] D.A. Reynolds P. Torres-Carrasquillo, "Approaches and applications of audio diarization," in *Proceedings of ICASSP*, 2005, pp. V–953–V 956.

[26] S. E. Tranter and D. A. Reynolds, "Speaker diarisation for broadcast news," in *Proceedings Odyssey Speaker and Language Recognition Workshop*, 2004, pp. 337–344.

[27] P. Deleglise, Y. Esteve, S. Meignier, and T. Merlin, "The LIUM speech transcription system: a CMU Sphinx III-based System for French Broadcast News," in *Proceedings of Interspeech, Lisbon, Portugal*, 2005.

[28] Sylvain Meignier, Daniel Moraru, Corinne Fredouille, Jean-Francois Bonastre, and Laurent Besacier, "Step-by-step and integrated approaches in broadcast news speaker diarization," *Computer Speech and Language*, vol. 20, pp. 303–330, 2006.

[29] Sylvain Meignier, Jean François Bonastre, and Stephane Igounet, "E-HMM approach for learning and adapting sound models for speaker indexing," in *Proc. Odyssey Speaker and Language Recognition Workshop*, 2001, pp. 175–180.

[30] J. Ajmera, H. Bourlard, and I. Lapidot, "Improved Unknown-Multiple Speaker clustering using HMM," Tech. Rep., IDIAP/EPFL, 2002.

[31] C. Barras, X. Zhu, S. Meignier, and J.L. Gauvain, "Improving Speaker Diarization," in *Proceedings of Fall 2004 Rich Transcription Workshop (RT-04)*, November 2004.

[32] Steven B. Davis and Paul Mermelstein, "Comparison of parametric representations for monosyllabic word recognition in continuously spoken sentences," *Readings in speech recognition*, pp. 65–74, 1990.

[33] Florian Hönig, Georg Stemmer, Christian Hacker, and Fabio Brugnara, "Revising Perceptual Linear Prediction (PLP)," in *Proceedings of the 9th European Conference on Speech Communication and Technology*, ISCA, Ed., Bonn, 2005, pp. 2997–3000.

[34] L. Ma, D.J. Smith, and Ben P. Milner, "Context awareness using environmental noise classification," in *Proceedings of Eurospeech*, 2003.

[35] Lie Lu, Hao Jiang, and Hongjiang Zhang, "A robust audio classification and segmentation method," in *ACM Multimedia*, 2001, pp. 203–211.

[36] E. Scheirer and M. Slaney, "Construction and evaluation of a robust multifeature music/speech discriminator," in *Proceeding of ICASSP*, Apr. 1997, vol. II.

[37] Ascension Gallardo-Antolin, Xavier Anguera, and Chuck Wooters, "Multi-stream speaker diarization systems for the meetings domain," in *Proceedings of Interspeech*, 2006.

[38] S. Meignier, D. Moraru, C. Fredouille, L. Besacier, and J.F. Bonastre, "Benefits of prior acoustic segmentation for automatic speaker segmentation," in *Proc. Int. Conf. on Acoustics, Speech and Signal Processing*, May 2004, pp. 397–400.

[39] Christophe Biernacki, Gilles Celeux, and Gérard Govaert, "Choosing starting values for the EM algorithm for getting the highest likelihood in multivariate Gaussian mixture models," *Comput. Stat. Data Anal.*, vol. 41, no. 3-4, pp. 561–575, 2003.

[40] Janez Zibert, Nikola Pavesic, and FranceMihelic, "Speech/non-speech segmentation based on phoneme recognition features," *EURASIP J. Appl. Signal Process.*, vol. 2006, pp. 1–13, 2006.

[41] M. Omar, U. Chaudhari, and G. Ramaswamy, "Blind Change Detection for Audio Segmentation," in *Proceedings of ICASSP*, 2005, vol. 1, pp. 501–504.

[42] Thomas Kemp, Michael Schmidt, Martin Westphal, and Alex Waibel, "Strategies for automatic segmentation of audio data," in *Proceedings of ICASSP*, 2000, pp. 1423–1426.

[43] Matthew A. Siegler, Uday Jain, Bhiksha Raj, and Richard M. Stern, "Automatic segmentation, classification and clustering of broadcast news audio," in *Proc. DARPA Speech Recognition Workshop*, 1997, pp. 97–99.

[44] M. Cettolo and M. Vescovi, "Efficient audio segmentation algorithms based on the BIC," in *Proceedings of ICASSP*, 2003, vol. 6, pp. 537–540.

[45] Lie Lu and Hong Jiang Zhang, "Unsupervised speaker segmentation and tracking in real-time audio content analysis," *Multimedia Systems*, vol. 10, no. 4, pp. 332–343, 2005.

[46] Perrine Delacourt, David Kryze, and Christian J. Wellekens, "Speaker-based segmentation for audio data indexing," in *Speech Communication*, 1999, pp. 111–126.

[47] S. Kwon and S. Narayanan, "A method for on-line speaker indexing using generic reference models," in *Proc. 8th European Conf. Speech Communication and Technology, Geneva, Switzerland*, 2003, pp. 2653–2656.

[48] Alain Tritschler and Ramesh A. Gopinath, "Improved speaker segmentation and segments clustering using the Bayesian Information Criterion," in *Proceedings of Eurospeech*, 1999, pp. 679–682.

[49] M. Vescovi, M. Cettolo, and R. Rizzi, "A DP algorithm for speaker change detection," in *Proc. 8th European Conf. Speech Communication and Technology, Geneva, Switzerland*, 2003, pp. 2997–3000.

[50] S. E. Tranter, M. J. F. Gales, R. Sinha, S. Umesh, and P. C. Woodland, "The development of the Cambridge university RT-04 Diarisation system," in *Proc. Fall 2004 Rich Transcription Workshop (RT-04)*, 2004.

[51] S.E. Tranter and D.A. Reynolds, "An overview of automatic speaker diarization systems," *IEEE Trans. Audio, Speech, and Language Processing*, vol. 14, pp. 1557–1565, 2006.

[52] W.H. Tsai and H.M. Wang, "Speech utterance clustering based on the maximization of within-cluster homogeneity of speaker voice characteristics," *J. Acoust. Sos. Am.*, vol. 120, no. 3, pp. 1631–1645, 2006.

[53] Sancho Salcedo-Sanz, Ascensión Gallardo-Antolín, José M. Leiva-Murillo, and Carlos Bousoño-Calzón, "Offline speaker segmentation using genetic algorithms and mutual information," *IEEE Trans. Evolutionary Computation*, vol. 10, no. 2, pp. 175–186, 2006.

[54] A. Hyvarinen, "Fast and robust fixed-point algorithms for Independent Component Analysis," *IEEE Transactions on Neural Networks*, vol. 10, no. 3, pp. 626–643, 1999.

[55] José M. Leiva-Murillo, Sancho Salcedo-Sanz, Ascensión Gallardo-Antolín, and Antonio Artés-Rodríguez, "A simulated annealing approach to speaker segmentation in audio databases," *Eng. Appl. Artif. Intell., Elsevier*, vol. 21, no. 4, pp. 499–508, 2008.

[56] Ozgur Cetin and Elizabeth Shriberg, "Speaker overlaps and ASR errors in meetings: Effects before, during, and after the overlap," in *Proc. of ICASSP*, 2006.

[57] Kofi Boakye, Beatriz Trueba-Hornero, Oriol Vinyals, and Gerald Friedland, "Overlapped Speech Detection for Improved Speaker Diarization in Multiparty Meetings," in *Proceedings of IEEE ICASSP*, April 2008, pp. 4353–4356.

[58] H. Jeffreys, "An invariant form for the prior probability in estimation problems," *Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences*, vol. 186, no. 1007, pp. 453–461, 1946.

[59] E.T. Jaynes, "On the rationale of maximum-entropy methods," *Proceedings of the IEEE*, vol. 70, pp. 939–952, 1982.

[60] Adam L. Berger, Stephen D. Della Pietra, and Vincent J. D. Della Pietra, "A maximum entropy approach to natural language processing," *Computational Linguistics*, vol. 22, no. 1, pp. 39–71, 1996.

[61] J. Jiwoon and R. Manmatha, "Using maximum entropy for automatic image annotation," in *Proceedings of Image and Video Retrieval: Third International Conference, CIVR 2004, Dublin, Ireland*, San Jose, CA, July 2004.

[62] Y. H. Abdel-Haleem, S. Renals, and N. D. Lawrence, "Acoustic space dimensionality selection and combination using the maximum entropy principle," in *Proc. IEEE ICASSP*, 2004.

[63] Robert Malouf, "A comparison of algorithms for maximum entropy parameter estimation," in *COLING-02: proceedings of the 6th conference on Natural language learning*, Morristown, NJ, USA, 2002, pp. 1–7, Association for Computational Linguistics.

[64] Stanley F. Chen, Ronald Rosenfeld, and Associate Member, "A survey of smoothing techniques for ME models," *IEEE Transactions on Speech and Audio Processing*, vol. 8, pp. 37–50, 2000.

[65] Joshua Goodman, "Exponential priors for maximum entropy models," in *In Proceedings of the Annual Meeting of the Association for Computational Linguistics*, 2003, pp. 305–312.

[66] Robert Tibshirani, "Regression shrinkage and selection via the lasso," *Journal of the Royal Statistical Society, Series B*, vol. 58, pp. 267–288, 1994.

[67] H. Gish, M.-H. Siu, and R. Rohlicek, "Segregation of speakers for speech recognition and speaker identification," in *Proceedings of the Acoustics, Speech, and Signal Processing, 1991. ICASSP*, Washington, DC, USA, 1991, pp. 873–876, IEEE Computer Society.

[68] G. Schwarz, "Estimating the dimension of a model," *Annals of Statistics*, vol. 6, 1978.

[69] S.S. Chen and P.S. Gopalakrishnam, "Speaker, Environment and Channel Change Detection and Clustering via the Bayesian Information Criterion," in *Proceedings of DARPA Broadcast News Transcription and Understanding Workshop*, 1998.

[70] Robert E. Kass and Larry Wasserman, "A Reference Bayesian test for nested hypotheses and its relation to the Schwarz criterion," *Journal of the American Statistical Association*, vol. 90, pp. 928–934, 1995.

[71] A. Barron J. Rissenan and B. Yu, "The Minimum Description Length principle in coding and modeling," *IEEE Trans. Information Theory*, vol. 44, pp. 2743 – 2760, 1998.

[72] Andrew Y. Ng, Michael I. Jordan, and Yair Weiss, "On spectral clustering: Analysis and an algorithm," in *Advances in Neural Information Processing Systems 14*. 2001, pp. 849–856, MIT Press.

[73] Nicolas Chopin and Florian Pelgrin, "Bayesian inference and state number determination for Hidden Markov Models: an application to the information content of the yield curve about inflation," *Journal of Econometrics*, vol. 123, no. 2, pp. 327–344, December 2004.

[74] Christophe Biernacki, Gilles Celeux, and Gérard Govaert, "Assessing a mixture model for clustering with the Integrated Completed Likelihood," *IEEE Trans. Pattern Anal. Mach. Intell.*, vol. 22, no. 7, pp. 719–725, 2000.

[75] D. J. C. Mackay, "Information theory, inference, and learning algorithms," *Cambridge University Press New York*, 2003.

[76] Robert E. Kass and Adrian E. Raftery, "Bayes Factors," *Journal of the American Statistical Association*, vol. 90, no. 430, pp. 773–795, 1995.

[77] M. J. Beal, *Variational Algorithms for Approximate Bayesian Inference*, Ph.D. thesis, Gatsby Computational Neuroscience Unit, University College London, 2003.

[78] C. C. Rodriguez, "Entropic priors for discrete probabilistic networks and for mixtures of gaussians models," in *Bayesian Inference and Maximum Entropy Methods*. 2001, pp. 410–432, Inst. Physics.

[79] X. Zhu, C. Barras, S. Meignier, and J.L Gauvain, "Combining Speaker Identification and BIC for Speaker Diarization," in *Proceedings of Interspeech*, September 2005, pp. 2441 – 2444.

[80] Emily B. Fox, Erik B. Sudderth, Michael I. Jordan, and Alan S. Willsky, "The Sticky HDP-HMM: Bayesian nonparametric Hidden Markov Models with Persistent States," 2009.

[81] C.-Y. Sin and H. White, "Information criteria for selecting possibly misspecified parametric models," *Journal of Econometrics*, vol. 71, no. 1-2, pp. 207–225, 1996.

[82] Han Hong and Bruce Preston, "Bayesian averaging, prediction and nonnested model selection," NBER Working Papers 14284, National Bureau of Economic Research, Inc, Aug 2008.

[83] S. Galliano, E. Geoffrois, G. Gravier, J.-F. Bonastre, D. Mostefa, and K. Choukri, "Corpus description of the ESTER evaluation campaign for the Rich Transcription of French Broadcast News," in *Proc. Language Evaluation and Resources Conference*, 2006.

[84] S. Galliano, E. Geoffrois, D. Mostefa, K. Choukri, J.-F. Bonastre, and G. Gravier, "The ESTER phase II evaluation campaign for the rich transcription of french broadcast news," in *Proceedings of European Conference on Speech Communication and Technology (Interspeech)*, September 2005, pp. 1149 – 1152.

[85] K. Fukunaga and L. Hostetler, "The estimation of the gradient of a density function, with applications in pattern recognition," *IEEE Trans. on Information Theory*, vol. 21, no. 1, pp. 32–40, January 1975.

[86] Y. Cheng, "Mean Shift, Mode Seeking, and Clustering," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 17, no. 8, pp. 790 – 799, August 1995.

[87] D. Comaniciu, V. Ramesh, and P. Meer, "The variable bandwidth Mean Shift and data-driven scale selection," in *Proc. 8th Intl. Conf. on Computer Vision*, 2001, pp. 438–445.

[88] D. Comaniciu and P. Meer, "Mean shift: A robust approach towards feature space analysis," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 24, no. 5, pp. 603 – 619, May 2002.

[89] Vincent Wan and Steve Renals, "Evaluation of kernel methods for speaker verification and identification," 2002.

[90] Vincent Wan and Steve Renals, "Speaker verification using sequence discriminant support vector machines," 2005.

[91] Michael Collins, Robert E. Schapire, and Yoram Singer, "Logistic Regression, Adaboost and Bregman Distances," 2000.

[92] C. C. Rodriguez, "Entropic priors for discrete probabilistic networks and for mixtures of gaussians models," *AIP Conference Proceedings*, vol. 617, no. 1, pp. 410–432, 2002.

[93] Hichem Snoussi, "The geometry of prior selection," *Neurocomputing*, vol. 67, pp. 214–244, 2005.

[94] Douglas A. Reynolds, Thomas F. Quatieri, and Robert B. Dunn, "Speaker verification using adapted Gaussian mixture models," in *Digital Signal Processing*, 2000, vol. 10, pp. 19–41.

[95] Fabio Valente, *Variational Bayesian methods for audio indexing*, Ph.D. thesis, September 2005.

[96] Jenq neng Hwang, Shyh rong Lay, and Alan Lippman, "Nonparametric multivariate density estimation: A comparative study," *IEEE Trans. Signal Processing*, vol. 42, pp. 2795–2810, 1994.

[97] M.H. Hansen and B. Yu, "Model selection and the principle of Minimum Description Length," *Journal of the American Statistical Association*, vol. 96, pp. 746–774, June 2001.

[98] Shun ichi Amari, "Information geometry of the EM and em algorithms for neural networks," *Neural Networks*, vol. 8, pp. 1379–1408, 1995.

[99] Shun ichi Amari, "Natural gradient works efficiently in learning," *Neural Computation*, vol. 10, no. 2, pp. 251–276, 1998.

[100] Shun ichi Amari, "Natural gradient learning for over- and under-complete bases in ica," *Neural Computation*, vol. 11, no. 8, pp. 1875–1883, 1999.

[101] Antti Honkela, Matti Tornio, Tapani Raiko, and Juha Karhunen, "Neural information processing," chapter Natural Conjugate Gradient in Variational Inference, pp. 305–314. Springer-Verlag, Berlin, Heidelberg, 2008.

[102] Shun ichi Amari, Hyeyoung Park, and Kenji Fukumizu, "Adaptive method of realizing natural gradient learning for multilayer perceptrons," *Neural Computation*, vol. 12, no. 6, pp. 1399–1409, 2000.