ΜΙΚΤΑ ΜΟΝΤΕΛΑ Π-ΣΙΓΜΟΕΙΔΩΝ ΚΑΤΑΝΟΜΩΝ

Η
ΜΕΤΑΠΤΥΧΙΑΚΗ ΕΡΓΑΣΙΑ ΕΞΕΙΔΙΚΕΥΣΗΣ

Υποβάλλεται στην

ορισθείσα από την Γενική Συνέλευση Ειδικής Σύνθεσης
του Τμήματος Πληροφορικής
Εξεταστική Επιτροπή

από τον

Αναστάσιο Αλιβάνογλου

ως μέρος των Υποχρεώσεων

για τη λήψη

του

ΜΕΤΑΠΤΥΧΙΑΚΟΥ ΔΙΠΛΩΜΑΤΟΣ ΣΤΗΝ ΠΛΗΡΟΦΟΡΙΚΗ

ΜΕ ΕΞΕΙΔΙΚΕΥΣΗ ΣΤΙΣ ΤΕΧΝΟΛΟΓΙΕΣ-ΕΦΑΡΜΟΓΕΣ

Ιούνιος 2007

ΑΦΙΕΡΩΣΗ

Στα αγαπημένα πρόσωπα της οικογένειάς μου,
Λάζαρο, Ευαγγελία και Χρυσούλα….

ΕΥΧΑΡΙΣΤΙΕΣ

Θα ήθελα να ευχαριστήσω θερμά τον επιβλέποντα καθηγητή κ. Αριστείδη Λύκα για την ουσιαστική και ειλικρινή βοήθεια του, καθ' όλη την διάρκεια της εκπόνησης της μεταπτυχιακής μου εργασίας. Οι συμβουλές και η υπομονή που επέδειξε ήταν στοιχεία καθοριστικά για την ολοκλήρωση των υποχρεώσεών μου. Επίσης, θα ήθελα να ευχαριστήσω την οικογένειά μου, για την αμέριστη ηθική και οικονομική συμπαράσταση που μου παρείχε σε όλη την διάρκεια φοίτησης μου στο Πανεπιστήμιο Ιωαννίνων.

ΠΕΡΙΕΧΟΜΕΝΑ

Σελ.

ΕΥΡΕΤΗΡΙΟ ΠΙΝΑΚΩΝ

Πίνακας Σελ.

ΕΥΡΕΤΗΡΙΟ ΣΧΗΜΑΤΩΝ

Σχήμα Σελ

ΠΕΡΙΛΗΨΗ

Αναστάσιος Αλιβάνογλου του Λαζάρου και της Ευαγγελίας. MSc, Τμήμα Πληροφορικής, Πανεπιστήμιο Ιωαννίνων, Ιούνιος, 2007. Μικτά μοντέλα Π-σιγμοειδών κατανομών. Επιβλέποντας: Αριστείδης Λύκας.

Ο τομέας της αναγνώρισης προτύπων ασχολείται με ένα πλήθος προβλημάτων επεξεργασίας πληροφορίας όπως για παράδειγμα την αναγνώριση χειρόγραφων χαρακτήρων, την εξαγωγή κανόνων για δημιουργία συστημάτων λήψης απόφασης κ.α. Προκειμένου να είμαστε αποδοτικοί στην επίλυση αυτών των προβλημάτων είναι αναγκαίο να δημιουργηθεί ένα ευέλικτο και αποδοτικό μοντέλο περιγραφής δεδομένων. Το GMM (Gaussian Mixture Model) αποτελεί ένα από τα πιο διαδεδομένα μοντέλα περιγραφής, διότι τόσο οι καλές αναλυτικές του ιδιότητες όσο και η δυνατότητα περιγραφής πολλών τύπων δεδομένων το καθιστούν ξεχωριστό. Εντούτοις, και σε αυτό το μοντέλο υπάρχουν κάποιες βασικές αδυναμίες που συνοψίζονται κυρίως σε δύο βασικά σημεία. Πρώτον, τα αποτελέσματα της ομαδοποίησης που δίνει (τα κέντρα και οι πίνακες συμμεταβλητότητας) δεν μπορούν να παράγουν ερμηνεύσιμους, από τον άνθρωπο, κανόνες. Δεύτερον, αδυνατεί να περιγράψει με ικανοποιητικό τρόπο τα δεδομένα που ακολουθούν ομοιόμορφη κατανομή. Μια προσέγγιση στη λύση αυτών των προβλημάτων δίνει η συνάρτηση πυκνότητας πιθανότητας Π-sigmoid που προτείνεται στην εργασία αυτή. Η τελευταία μπορεί αφενός να περιγράψει ομοιόμορφα δεδομένα και αφετέρου να δημιουργήσει ερμηνεύσιμους κανόνες, αφού λύση που παρέχει είναι μια λίστα από υποδιαστήματα ανά διάσταση. Η τομή των υποδιαστημάτων αυτών ορίζει στις D διαστάσεις ένα υπερ-ορθογώνιο. Στη συνέχεια ορίζεται το μικτό μοντέλο Π-sigmoid κατανομών (ΠsMM) και για την εκπαίδευση προτείνεται η χρήση του αλγορίθμου GEM (Generalized EM). Για την βελτιστοποίηση στο M-βήμα του GEM γίνεται χρήση της μεθόδου βελτιστοποίησης BFGS. Επίσης, προτείνουμε μια τροποποίηση της Π-

sigmoid κατανομής για να μπορεί περιγραφεί και περιστραμμένες ομάδες δεδομένων. Το γεγονός αυτό οδήγησε στην ανάπτυξη ειδικής τεχνικής για τον καθορισμό των πινάκων περιστροφής, τόσο για την απλή κατανομή όσο και για ένα ΠsΜΜ. Η απόδοση του ΠsΜΜ εξετάζεται μέσα από πειράματα σε τεχνητά και πραγματικά δεδομένα, καθώς επίσης σε προβλήματα κατηγοριοποίησης δεδομένων και κατάτμησης εικόνων.

EXTENDED ABSTRACT IN ENGLISH

Alivanoglou, Anastasios, Initials. MSc, Computer Science Department, University of Ioannina, Greece. June, 2007. Π-sigmoid Mixture Models. Thesis Supervisor: Aristidis Likas.

The term pattern recognition encompasses a wide range of information processing problems of great practical significance, from speech recognition an the classification of handwritten characters, to fault detection in machinery and medical diagnosis. Often these are problems which many humans solve in a seemingly effortless fashion. However, their solution using computers has, in many cases, proved to be immensely difficult. In order to have the best opportunity of developing effective solutions, it is important to adopt a principled approach based on sound theoritical concepts.

The most general, and the most natural, framework in which to formulate solutions to pattern recognition problems is a statistical one, which recognizes the probabilistic nature both of the information we seek to process, and of the form in which we should express the results. Statistical pattern recognition is a well established field with a long history.

In statistical pattern recognition we assume that the data has been generated as a result of a statistical process and we seek the statistical model that best fits the data, where the statistical model is described in terms of a distribution and a set of parameters for that distribution. At a high level, this process involves deciding on a statistical model for the data and estimating the parameters of that model from the data. This thesis deals with a particular type of statistical model, mixture models, which model the data by using a convex combination of statistcal distributions. Each distribution corresponds to a cluster and the parameters of each distribution provide a description of the corresponding cluster.

In this thesis we propose a new probability density function, the Π-sigmoid. It took its name from its ability to form the shape of the greek letter “Π”. More specifically, we demonstrate the properties and the different shapes that can take for particular values of its parameters. We then describe Π-sigmoid mixture models (ΠsMM) and we consider how parameters can be estimated for this statistical model. We first show how a procedure known as maximum likelihood estimation (MLE) can be used to estimate parameters for the simple case of a single Π-sigmoid and then present how can we extend this approach for estimating the parameters of a ΠsMM. This can be achieved via the Generalized Expectation Maximization (GEM) algorithm, which makes an initial guess for the parameters, and then iteratively improves these estimates. Moreover, we provide techniques to adjust orientation of the distribution as well as the way to deal with noisy datasets.

We assess the performance of the method using both real and synthetic datasets. We also test its effectiveness in classification problems and image segmentation.

ΚΕΦΑΛΑΙΟ 1. ΕΙΣΑΓΩΓΗ

1.1. Στατιστική αναγνώριση προτύπων

Ο όρος αναγνώριση προτύπων αναφέρεται σε ένα πλήθος προβλημάτων επεξεργασίας πληροφορίας όπως είναι η αναγνώριση φωνής, η αναγνώριση χειρόγραφων χαρακτήρων κτλ. Τέτοιου είδους προβλήματα είναι απλά για την ανθρώπινη νοημοσύνη, π.χ. ένας άνθρωπος έχει την ικανότητα να αναγνωρίζει χειρόγραφους χαρακτήρες ακόμη και στις περιπτώσεις που αυτοί είναι γραμμένοι με ένα ιδιόμορφο τρόπο. Ωστόσο η επίλυση τέτοιων προβλημάτων χρησιμοποιώντας υπολογιστικές μηχανές έχει αποδειχτεί ιδιαίτερα δύσκολη και έχει αποτελέσει το επίκεντρο σημαντικής ερευνητικής προσπάθειας. Προκειμένου να κατασκευαστούν αποδοτικά συστήματα για την επίλυση προβλημάτων αναγνώρισης προτύπων πρέπει να υιοθετηθεί μια γενική προσέγγιση η οποία θα προσφέρει ένα πλαίσιο αρχών και εννοιών πάνω στο οποίο θα βασιστεί στη συνέχεια η ερευνητική προσπάθεια.

Η στατιστική προσέγγιση προσπαθεί να αναδείξει την πιθανοτική φύση του προβλήματος. Ο τομέας της στατιστικής αναγνώρισης προτύπων είναι ο παλαιότερος και καλύτερα θεμελιωμένος και βασίζεται σε έννοιες της θεωρίας στατιστικής και πιθανοτήτων.

Στο παρών κεφάλαιο, θα μελετήσουμε έννοιες και μεθόδους της στατιστικής αναγνώρισης προτύπων τα οποία θα βοηθήσουν στην κατανόηση των όσων θα αναφερθούν σε επόμενα κεφάλαια.

1.2. Εκτίμηση συνάρτησης πυκνότητας πιθανότητας

Ίσως, η πιο άμεση και αποτελεσματική προσέγγιση για την εκτίμηση μιας άγνωστης κατανομής, στηρίζεται στην παραδοχή ότι η τελευταία αποτελεί μια συνάρτηση εξαρτώμενη από ένα διάνυσμα ρυθμιζόμενων παραμέτρων. Η διαδικασία της εκτίμησης, συνεπώς, ανάγεται στη βελτιστοποίησης των παραμέτρων αυτών, η οποία θα οδηγήσει στο καλύτερο δυνατό ταίριασμα της συνάρτησης με την κατανομή που ακολουθούν τα δεδομένα μας. Θα συμβολίζουμε την συνάρτηση πυκνότητας πιθανότητας p(x) η οποία εξαρτάται από το διάνυσμα παραμέτρων Θ ως p(x; Θ).

1.2.1. Η κανονική κατανομή ως παραμετρικό μοντέλο

Μια από τις πιο διαδεδομένες παραμετρικές συναρτήσεις της μορφής p(x;Θ) είναι η κανονική (Gaussian) κατανομή, η οποία χαρακτηρίζεται από καλές στατιστικές και αναλυτικές ιδιότητες. Στην μονοδιάστατη εκδοχή της η κανονική κατανομή μπορεί να γραφτεί στην μορφή:

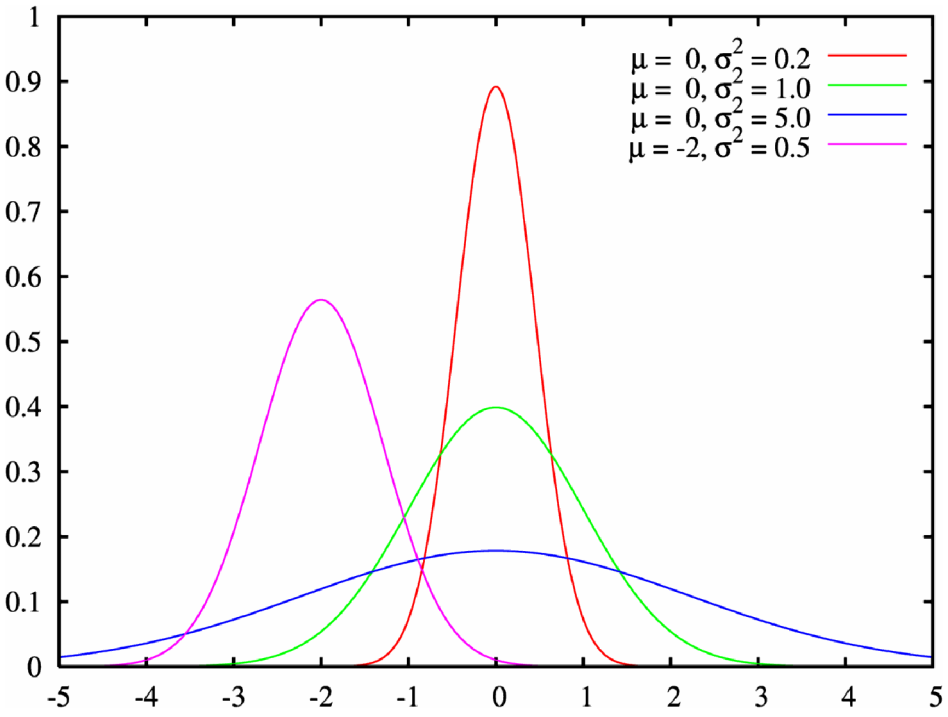
$$N(x;\mu,\sigma^2)=\frac{1}{(2\pi\sigma^2)^{1/2}}\exp\left\{-\frac{(x-\mu)^2}{2\sigma^2}\right\} \tag{1.1}$$

όπου μ και σ^2 είναι το κέντρο και η διασπορά αντίστοιχα, ενώ το σ (το οποίο είναι η τετραγωνική ρίζα της διασποράς) ονομάζεται τυπική απόκλιση. Ο συντελεστής μπροστά από την εκθετική συνάρτηση στην σχέση (1.1) είναι η σταθερά κανονικοποίησης η οποία διασφαλίζει ότι $\int_{-\infty}^{\infty} N(x;\mu,\sigma^2)dx=1$. Αναφορικά με το κέντρο και την διασπορά της μονοδιάστατης κανονικής κατανομής, είναι εύκολα να δειχτεί ότι ικανοποιούν τις παρακάτω σχέσεις:

$$\mu=\mathrm{E}[x]=\int_{-\infty}^{\infty} xp(x)dx \tag{1.2}$$

$$\sigma^2=\mathrm{E}[(x-\mu)^2]=\int_{-\infty}^{\infty}(x-\mu)^2\,p(x)dx \tag{1.3}$$

όπου με E[.] συμβολίζουμε την αναμενόμενη τιμή.



Σχήμα 1.1 Η κανονική συνάρτηση πυκνότητας πιθανότητας για διάφορες τιμές των παραμέτρων της.

Γενικότερα, η πολυδιάστατη κανονική κατανομή στις D διαστάσεις έχει την ακόλουθη μορφή:

$$N(x\,|\,\mu,\Sigma)=\frac{1}{(2\pi)^{D/2}\,|\Sigma|^{1/2}}\exp\left\{-\frac{1}{2}(x-\mu)^{T}\Sigma^{-1}(x-\mu)\right\} \tag{1.4}$$

Στην νέα αυτή πολυδιάστατη εκδοχή, το κέντρο μ είναι πλέον ένα D-διάστατο διάνυσμα ενώ ο Σ είναι ένας $D\times D$ πίνακας συμμεταβλητότητας και |Σ| είναι η ορίζουσά του. Ο παράγοντας μπροστά από το εκθετικό μέρος της συνάρτησης, όπως και στην μονοδιάστατη εκδοχή, μας εγγυάται ότι θα ισχύει $\int_{-\infty}^{\infty} N(x;\mu,\Sigma)dx=1$. Αναφορικά με τον συμβολισμό, θα γράφουμε X~N(x; μ, Σ), όταν το X είναι μια τυχαία μεταβλητή που ακολουθεί την κανονική κατανομή με κέντρο μ και πίνακα συμμεταβλητότητας Σ. Είναι φανερό ότι η συνάρτηση πυκνότητας N(x; μ, Σ) χαρακτηρίζεται από τις παραμέτρους μ και Σ οι οποίες ικανοποιούν τις παρακάτω σχέσεις:

$$\mu = \mathrm{E}[x] \tag{1.5}$$

$$\Sigma = \mathrm{E}[(x-\mu)(x-\mu)^{\mathrm{T}}] \tag{1.6}$$

Από την σχέση (1.6) συνάγεται ότι ο πίνακας Σ είναι συμμετρικός και θετικά ημιορισμένος. Η ιδιότητα αυτή της συμμετρίας μας οδηγεί στο συμπέρασμα ότι ο Σ αποτελείται από D(D+1)/2 ανεξάρτητες παραμέτρους το οποίο φαίνεται και από τον παρακάτω ορισμό:

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \mathsf{L} & \sigma_{1D} \\ \sigma_{12} & \mathsf{O} & & \sigma_{2D} \\ \mathsf{M} & & \mathsf{O} & \mathsf{M} \\ \sigma_{1D} & \sigma_{2D} & \mathsf{L} & \sigma_D^2 \end{bmatrix}$$

Όπου σ_d^2, d=1,…,D είναι η διασπορά στην d-οστή διάσταση και σ_{ij}, $i \neq j$ είναι η συμμεταβλητότητα της i-οστής και της j-οστής συνιστώσας των διανυσμάτων x. Συνυπολογίζοντας τώρα και τα D στοιχεία του κέντρου μ, έχουμε συνολικά D(D+3)/2 ανεξάρτητες παραμέτρους οι οποίες χαρακτηρίζουν την κανονική κατανομή.

Σε αυτό το σημείο θα μας απασχολήσει η γεωμετρική μορφή της κανονικής κατανομής. Η συναρτησιακή εξάρτηση της κατανομής αυτής από το x, διατυπώνεται διαμέσου της παρακάτω τετραγωνικής μορφής:

$$\Delta^2 = (x-\mu)^{\mathrm{T}}\Sigma^{-1}(x-\mu) \tag{1.7}$$

η οποία εμφανίζεται στο εκθετικό μέρος της (1.4). Η ποσότητα Δ ονομάζεται απόσταση mahalanobis μεταξύ του x και του μ.

Σταθερή απόσταση Δ^2

Σε αυτή την περίπτωση όλα τα x ανήκουν σε μια υπερελλειψοειδή επιφάνεια με κέντρο μ και σχήμα που καθορίζεται από τον πίνακα Σ. Είναι προφανές ότι η τιμή της κανονικής συνάρτησης (1.4) για όλα αυτά τα σημεία είναι σταθερή και αυτό μας οδηγεί στο συμπέρασμα ότι τα δεδομένα x που παράγονται από αυτή, σχηματίζουν στις D διαστάσεις υπερελλειψοειδείς πυρήνες.

Διαγώνιος πίνακας συμμεταβλητότητας Σ

Πολλές φορές είναι αρκετά βολικό να θεωρήσουμε ότι οι συνιστώσες των διανυσμάτων x_i ενός συνόλου δεδομένων $X=\{x_1,\ldots,x_N\}$ είναι ανεξάρτητες, δηλαδή ισχύει $\sigma_{ij}=\mathrm{cov}(x_i,x_j) = 0, \forall i,j , i \neq j$. Αυτό πρακτικά σημαίνει ότι ο πίνακας Σ εκφυλίζεται σε ένα διαγώνιο πίνακα της μορφής $\Sigma=\mathrm{diag}(\sigma_1^2, \ldots, \sigma_D^2)$. Με βάση αυτή την υπόθεση η κανονική κατανομή παίρνει τη μορφή:

$$N(x;\mu,\Sigma) = \frac{1}{(2\pi)^{D/2}\sigma_1 \text{ L } \sigma_D}\exp\left\{-\frac{(x_1-\mu_1)^2}{2\sigma_1^2} - \text{L} - \frac{(x_D-\mu_D)^2}{2\sigma_D^2}\right\} \tag{1.8}$$

ή

$$N(x;\mu,\Sigma) = \prod_{d=1}^{D}\frac{1}{(2\pi)^{1/2}\sigma_d}\exp\left\{-\frac{(x_d-\mu_d)^2}{2\sigma_d^2}\right\} \tag{1.9}$$

Η παραπάνω διατύπωση χαρακτηρίζεται από 2D ανεξάρτητες παραμέτρους, και ουσιαστικά τα διανύσματα για τα οποία η N(x; μ, Σ) δίνει ίδιες τιμές, ορίζουν στις D διαστάσεις μια υπερέλλειψη που έχει τους άξονες της ευθυγραμμισμένους με τους κύριους άξονες.

Πίνακας Σ της μορφής $\sigma^2 I_D$

Σε αυτή την πιο απλή περίπτωση, ο πίνακας Σ είναι διαγώνιος για τον οποίο υποθέτουμε ότι η διασπορά για κάθε μια συνιστώσα είναι σταθερή, δηλαδή ισχύει $\Sigma=\mathrm{diag}(\sigma^2,\ldots, \sigma^2)= \sigma^2 I_D$, όπου I_D είναι ο $D\times D$ μοναδιαίος πίνακας. Τα διανύσματα που προκαλούν την ίδια τιμή στην κανονική κατανομή, σχηματίζουν πλέον στις D διαστάσεις μια υπερσφαίρα με κέντρο μ και ακτίνα σ^2. Αναφορικά τώρα με τις παραμέτρους, αυτές μειώνονται στις D+1, ενώ η μορφή της κατανομής παίρνει την παρακάτω μορφή:

$$N(x;\mu,\sigma^2) = \frac{1}{(2\pi\sigma)^{D/2}}\exp\left\{-\frac{\|x-\mu\|^2}{2\sigma^2}\right\} \tag{1.10}$$

η ποσότητα ||x-μ|| συμβολίζει την ευκλείδεια απόσταση των διανυσμάτων x και μ. Το πλεονέκτημα της τελευταίας αυτής προσέγγισης είναι ο μικρός αριθμός των παραμέτρων που χρειάζονται για την περιγραφή της κατανομής, το οποίο είναι

ταυτόχρονα και μειονέκτημα, αφού με αυτό τον τρόπο υποβαθμίζεται η γενικευτική ικανότητα του μοντέλου.

1.2.2. Μικτές κατανομές

Μια μικτή κατανομή ορίζεται ως μια ειδική περίπτωση γραμμικού συνδυασμού ενός πεπερασμένου αριθμού συναρτήσεων πυκνότητας πιθανότητας [4]. Πιο συγκεκριμένα η πιθανότητα μιας τυχαίας μεταβλητής x η οποία ακολουθεί μια μικτή κατανομή, υπολογίζεται από το άθροισμα M συναρτήσεων πυκνότητας πιθανότητας σταθμισμένων με βάρη. Η διατύπωση ενός τέτοιου μοντέλου αποτυπώνεται στην παρακάτω σχέση:

$$p(x;\Theta) = \sum_{j=1}^{M} \pi_j f(x;\theta_j) \tag{1.11}$$

όπου $f(x;\theta_j)$ αναπαριστά την j-οστή συνιστώσα κατανομή του μικτού μοντέλου και θ_j είναι το αντίστοιχο διάνυσμα παραμέτρων της. Τα π_j αποκαλούνται παράμετροι μίξης οι οποίες θα πρέπει να ικανοποιούν τους κάτωθι περιορισμούς:

$$0 \le \pi_j \le 1, \forall j \tag{1.12}$$

$$\sum_{j=1}^{M} \pi_j = 1 \tag{1.13}$$

Οι παραπάνω περιορισμοί είναι ταυτόσημοι με αυτούς που πρέπει να πληρούνται έτσι ώστε μια ποσότητα να θεωρείται πιθανότητα. Άρα τα βάρη π_j μπορούν να ερμηνευτούν ως η εκ των προτέρων πιθανότητα σύμφωνα με την οποία ένα διάνυσμα μπορεί να παραχθεί από την j-οστή συνιστώσα κατανομή του ολικού μοντέλου. Θα πρέπει να αναφερθεί ότι οι παράμετροι του μικτού μοντέλου ορίζονται ως $\Theta=\{(\pi_j,\theta_j),\ j=1,\ldots,M\}$. Η συνάρτηση $f(x;\theta_j)$ εκφράζει την δεσμευμένη κατανομή σύμφωνα με την οποία η j-οστή συνιστώσα παράγει το διάνυσμα x. Η διαδικασία παραγωγής ενός προτύπου από ένα μικτό μοντέλο της μορφής (1.11) είναι η επιλογή του j–οστού πυρήνα της μίξης με πιθανότητα π_j και εν συνεχεία η παραγωγή του προτύπου από την δεσμευμένη κατανομή $f(x;\theta_j)$. Μια πολύ σημαντική ιδιότητα των μικτών κατανομών είναι ότι επιλέγοντας κατάλληλα τις συνιστώσες κατανομές

και ρυθμίζοντας κατάλληλα τις παραμέτρους Θ του μοντέλου, μπορούμε να προσεγγίσουμε οποιαδήποτε συνεχή συνάρτηση πυκνότητας, με οσοδήποτε καλή ακρίβεια. Το τελευταίο εξαρτάται άμεσα και από τον αριθμό M των συνιστωσών κατανομών, αλλά και από την δυνατότητα που έχουμε για την σωστή επιλογή τιμών για τις παραμέτρους του μοντέλου.

Είναι ενδιαφέρον να δούμε της πληροφορίες ομαδοποίησης που μπορεί μας εξασφαλίσει η εκτίμηση συνάρτησης πυκνότητας πιθανότητας με μικτές κατανομές. Ας υποθέσουμε ότι έχουμε εκτελέσει μια διαδικασία μάθησης με την βοήθεια της οποίας προσδιορίστηκαν οι τιμές των παραμέτρων της μικτής κατανομής. Καταρχάς, η εκ των προτέρων πιθανότητα μιας συνιστώσας κατανομής, εκφράζει την αναλογία των δεδομένων που παράγονται από την κατανομή αυτή σε σχέση με το σύνολο των δεδομένων. Επιπλέον, μέσω των συστατικών κατανομών παίρνουμε πληροφορίες σχετικά με τα χαρακτηριστικά της κάθε ομάδας (π.χ. κέντρο, διακύμανση). Και τέλος για ένα οποιοδήποτε δεδομένο x μπορούμε να υπολογίσουμε την εκ των υστέρων πιθανότητα να έχει παραχθεί από την j-οστή κατανομή, όπου με την χρήση του θεωρήματος Bayes μπορεί να εκφραστεί ως:

$$p(j \mid x, \Theta) = \frac{\pi_j f(x; \theta_j)}{\sum_{k=1}^{M} \pi_k f(x; \theta_k)} \tag{1.14}$$

Οι εκ των υστέρων πιθανότητες θα πρέπει να ικανοποιούν τον περιορισμό:

$$\sum_{j=1}^{M} p(j \mid x, \Theta) = 1 \tag{1.15}$$

1.3. Μέθοδοι εκτίμησης παραμέτρων

1.3.1. Μέγιστη Πιθανοφάνεια

Σε αυτή την ενότητα θα παρουσιάσουμε μια μέθοδο εκτίμησης των παραμέτρων για τα παραμετρικά μοντέλα δίνοντας έμφαση στην εφαρμογή της στην κανονική κατανομή.

Έχοντας αποφασίσει για την μορφή της παραμετρικής συνάρτησης που θα χρησιμοποιήσουμε αυτό που μας απομένει, είναι η εύρεση των κατάλληλων τιμών για τις παραμέτρους της συνάρτησης για την βέλτιστη δυνατή περιγραφή των δεδομένων από αυτή. Μια αποτελεσματική και ευρέως χρησιμοποιούμενη μέθοδος είναι και η "Μέγιστη Πιθανοφάνεια" (Maximum Likelihood). Σε αυτή τη μέθοδο επιστρατεύουμε τεχνικές από τον τομέα της Βελτιστοποίησης, για να μεγιστοποιήσουμε τη συνάρτηση της πιθανοφάνειας που θα ορίσουμε στην συνέχεια. Υποθέστε ότι έχουμε μια παραμετρική συνάρτηση πυκνότητας της μορφής $p(x; \Theta)$ που προφανώς χαρακτηρίζεται από ένα σύνολο άγνωστων παραμέτρων Θ, τις οποίες θέλουμε να εκτιμήσουμε. Για παράδειγμα, σε μια πολυδιάστατη κανονική κατανομή θα θέλαμε να εκτιμήσουμε τις παραμέτρους $\Theta=\{\mu, \Sigma\}$. Θεωρώντας τώρα, ένα σύνολο δεδομένων $X=\{x_1,\ldots,x_N\}$ από D-διάστατα τυχαία δείγματα, τα οποία παράχθηκαν ανεξάρτητα από την $p(x; \Theta)$, η από κοινού συνάρτηση πυκνότητας πιθανότητας $p(X; \Theta)$ δίνεται από την σχέση:

$$p(X;\Theta) = \prod_{i=1}^{N} p(x_i;\Theta) = L(\Theta \mid X) \tag{1.16}$$

όπου η ποσότητα $L(\Theta|X)$ είναι γνωστή ως συνάρτηση "πιθανοφάνειας" του Θ δοθέντος του συνόλου δεδομένων X. Ο εκτιμητής μέγιστης πιθανοφάνειας των παραμέτρων είναι εξ' ορισμού η τιμή $\hat{\Theta}$ η οποία μεγιστοποιεί την ποσότητα $L(\Theta| X)$ δηλαδή:

$$\hat{\Theta} = \arg\max_{\Theta} L(\Theta \mid X) \tag{1.17}$$

Διαισθητικά αντιλαμβανόμαστε ότι η εκτίμηση $\hat{\Theta}$ θα οδηγήσει στην καλύτερη δυνατή περιγραφή των δεδομένων από την συνάρτηση $p(x;\Theta)$, η οποία θα "ταιριάξει", με κάποια σχετική ακρίβεια, στο δοθέν σύνολο δεδομένων. Για αναλυτικούς λόγους, στη διαδικασία βελτιστοποίησης της πιθανοφάνειας, είναι βολικότερο να δουλεύουμε με το λογάριθμο της σχέσης (1.16). Λόγω του ότι ο λογάριθμος είναι γνησίως αύξουσα συνάρτηση, η τιμή $\hat{\Theta}$ η οποία μεγιστοποιεί την

λογαριθμική πιθανοφάνεια, μεγιστοποιεί επίσης και την συνάρτηση της πιθανοφάνειας. Σε αυτή την περίπτωση ορίζουμε την λογαριθμική πιθανοφάνεια ως:

$$LL(\Theta \mid X) = \log \prod_{i=1}^{N} p(x_i;\Theta) = \sum_{i=1}^{N} \log\, p(x_i;\Theta) \tag{1.18}$$

Η LL(Θ| X) είναι συνήθως παραγωγίσιμη συνάρτηση ως προς το διάνυσμα παραμέτρων Θ και άρα το μέγιστο της $\hat{\Theta}$ μπορεί να υπολογιστεί γενικά, με την βοήθεια γνωστών μεθόδων του διαφορικού λογισμού. Πιο συγκεκριμένα η απαραίτητη συνθήκη που θα πρέπει να ικανοποιείται για την εκτίμηση της μέγιστης πιθανοφάνειας είναι η εξής:

$$\nabla_{\Theta} LL(\Theta \mid \mathrm{X}) = \frac{\partial LL(\Theta \mid \mathrm{X})}{\partial \Theta} = \sum_{i=1}^{N} \frac{\partial \log p(x_i;\Theta)}{\partial \Theta} = 0 \tag{1.19}$$

όπου ∇_{Θ} είναι ο τελεστής παραγώγισης ως προς το διάνυσμα παραμέτρων Θ. Από την σχέση (1.19) προκύπτει ότι εάν το Θ είναι ένα διάνυσμα με k συνιστώσες, θα πρέπει να επιλύσουμε ένα σύστημα με k εξισώσεις για να βρούμε την απαιτούμενη λύση. Στην περίπτωση όπου η επίλυση αυτού του συστήματος δεν μπορεί να πραγματοποιηθεί αναλυτικά και σε κλειστή μορφή, πρέπει να καταφύγουμε σε πιο εξεζητημένες τεχνικές όπως είναι οι επαναληπτικές μέθοδοι βελτιστοποίησης.

Εφαρμογή στην κανονική κατανομή

Αν στη σχέση (1.18) βάλουμε στην θέση της δεσμευμένης κατανομής p(x; Θ), την πολυδιάστατη κανονική κατανομή προκύπτει:

$$LL(\Theta \mid X) = \sum_{i=1}^{N} \left[-\frac{D}{2}\log(2\pi) - \frac{1}{2}\log|\Sigma| - \frac{1}{2}(x_i - \mu)^{\mathrm{T}} \Sigma^{-1} (x_i - \mu) \right] \tag{1.20}$$

Παίρνοντας τώρα τις μερικές παραγώγους της σχέσης (1.20) ως προς τις παραμέτρους μ και Σ και θέτοντας αυτές ίσες με το μηδέν, προκύπτουν οι παρακάτω εξισώσεις:

$$\frac{\partial LL(\Theta \mid X)}{\partial \mu} = \sum_{i=1}^{N} \Sigma^{-1}(x_i - \mu) = 0 \tag{1.21}$$

$$\frac{\partial LL(\Theta \mid X)}{\partial \Sigma} = \sum_{i=1}^{N} \left[-\frac{1}{2}\Sigma^{-1} + \frac{1}{2}\Sigma^{-2}(x_i - \mu)(x_i - \mu)^{\mathrm{T}} \right] = 0 \tag{1.22}$$

Αναδιατάσσοντας κάθε μια από τις παραπάνω εξισώσεις παίρνουμε τις παρακάτω εκτιμήσεις για τις παραμέτρους μ και Σ:

$$\hat{\mu} = \frac{1}{N}\sum_{i=1}^{N} x_i \tag{1.23}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{i=1}^{N} (x_i - \mu)(x_i - \mu)^{\mathrm{T}} \tag{1.24}$$

Όπως παρατηρούμε από τις παραπάνω σχέσεις ο εκτιμητής μέγιστης πιθανοφάνειας $\hat{\mu}$ για τη μέση τιμή, είναι ο δειγματικός μέσος όρος όλων των προτύπων, ενώ αντίστοιχα για τον πίνακα συμμεταβλητότητας είναι ο δειγματικός πίνακας συμμεταβλητότητας.

Εφαρμογή σε μικτές κατανομές

Σε πολλά προβλήματα μηχανικής μάθησης τα οποία υιοθετούν ένα μικτό μοντέλο της μορφής (1.11) με K συνιστώσες, ενδιαφερόμαστε να προσδιορίσουμε την πιθανοτική συμμετοχή ενός σημείου, δοθέντων των παραμέτρων της j-οστής κατανομής. Χρησιμοποιώντας ένα τέτοιο μοντέλο και υποθέτοντας ότι οι παράμετροι του $\Theta=\{(\pi_j, \theta_j), j=1,\ldots,M\}$ είναι γνωστοί, τότε για κάθε j=1,…,M μπορούμε να υπολογίσουμε την συνεισφορά ενός σημείου στην j-οστή κατανομή, απευθείας από την σχέση (1.14). Στην πλειονότητα των περιπτώσεων όμως, οι παράμετροι Θ του μικτού μοντέλου είναι άγνωστοι και θα πρέπει να εκτιμηθούν . Μια άμεση σκέψη είναι να χρησιμοποιήσουμε και σε αυτήν την περίπτωση την μέθοδο της μέγιστης πιθανοφάνειας, στόχος της οποίας θα είναι η μεγιστοποίηση της ποσότητας

$$LL(\Theta \mid \mathrm{X}) = \sum_{i=1}^{N} \log p(x_i;\Theta) = \sum_{i=1}^{N} \log \sum_{j=1}^{M} \pi_j p(x_i;\theta_j) \tag{1.25}$$

ως προς τις παραμέτρους $\Theta=\{(\pi_j, \theta_j), j=1,\ldots,M\}$. Σε αντίθεση με την εκτίμηση μέγιστης πιθανοφάνειας της απλής κατανομής, εδώ παρουσιάζονται κάποια προβλήματα. Η βασική δυσκολία συνίσταται στο ότι η συνάρτηση $LL(\Theta| X)$, εξαιτίας του αθροίσματος που υπάρχει στον λογάριθμο, παρουσιάζει υψηλή μη γραμμικότητα, πράγμα που συνεπάγεται την ύπαρξη πολλών τοπικών μεγίστων. Επίσης, για την λογαριθμική πιθανοφάνεια μικτών κατανομών υπάρχουν διάφορα θεωρητικά ζητήματα σχετικά με την μοναδικότητα του εκτιμητή μέγιστη πιθανοφάνειας. Ειδικότερα λόγω των πολλών τοπικών μεγίστων, το ολικό μέγιστο μπορεί να προκύπτει για πολλά διαφορετικά διανύσματα παραμέτρων (που ορίζουν διαφορετικά μοντέλα), με αποτέλεσμα το βέλτιστο διάνυσμα να μην ορίζεται μοναδικά.

Θα πρέπει τέλος να σημειωθεί ότι, η ετικέτα του κάθε σημείου $\{x_i, i=1,\ldots,N\}$ που μας πληροφορεί για το ποια συνιστώσα κατανομή παρήγαγε αυτό το δείγμα, είναι άγνωστη. Αυτό δυσχεραίνει τη διαπραγμάτευση του προβλήματος μας, κάνοντας την αναλυτική διαχείριση του αδύνατη. Αν οι ετικέτες $z_i=\{ j,$ *αν* x_i *παράγεται από την κατανομή* $j\}$ των δεδομένων ήταν γνωστές, θα συλλέγαμε τα δεδομένα που παράγονται από την κάθε κατανομή και θα εκτελούσαμε Μ ξεχωριστές διαδικασίες εκτίμησης μέγιστης πιθανοφάνειας. Αυτή η άγνωστη ή κρυμμένη πληροφορία $Z=\{z_i, i=1,\ldots,N\}$ ανάγει το παρόν πρόβλημα σε ένα τυπικό πρόβλημα με ελλιπή δεδομένα, για την επίλυση των οποίων έχει σχεδιαστεί ο αλγόριθμος ΕΜ (expectation Maximization) τον οποίο αναλύουμε στην συνέχεια.

1.4. Βελτιστοποίηση μέσω του αλγορίθμου ΕΜ

1.4.1. Ορισμός του αλγορίθμου ΕΜ

Ο αλγόριθμος ΕΜ (Expectation Maximization) είναι μια γενική μέθοδος εύρεσης εκτιμητών μέγιστης πιθανοφάνειας των παραμέτρων μιας δοθείσας κατανομής, σε προβλήματα όπου κάποιες μεταβλητές δεν έχουν παρατηρηθεί (μη παρατηρήσιμες ή κρυμμένες μεταβλητές). Συνεπώς, η εφαρμογή του ΕΜ συνίσταται για την επίλυση δύο βασικών προβλημάτων. Το πρώτο υφίσταται όταν έχουμε δεδομένα από τα οποία

ορισμένες τιμές λείπουν, εξαιτίας κάποιων λαθών ή περιορισμών που υπάρχουν κατά την διάρκεια της διαδικασίας παρατήρησης. Το δεύτερο, αφορά κυρίως εφαρμογές μικτών μοντέλων, στα οποία η μεγιστοποίηση της πιθανοφάνειας είναι αναλυτικά αδύνατη. Γι' αυτό υποθέτουμε την ύπαρξη κάποιων επιπρόσθετων, αλλά κρυμμένων μεταβλητών, που προσδιορίζουν την ομάδα στην οποία ανήκει το κάθε πρότυπο.

Η γενική φιλοσοφία του ΕΜ διατυπώνεται ακολούθως. Ξεκινάμε με μια αρχική εκτίμηση $\Theta^{(0)}$ των παραμέτρων του μικτού μοντέλου, που πρέπει να εκτιμηθούν. Κάθε επανάληψη αποτελείται από δύο βήματα. Το πρώτο είναι το Ε-βήμα (expectation step) στο οποίο προσπαθούμε να υπολογίσουμε ένα τοπικό κάτω φράγμα της λογαριθμικής πιθανοφάνειας και να το μεγιστοποιήσουμε ως προς την κατανομή των κρυμμένων μεταβλητών. Παρακάτω, θα δείξουμε ότι αυτό είναι ισοδύναμο με το να υπολογίσουμε τη εκ των υστέρων κατανομή των κρυμμένων μεταβλητών, δοθέντος των παρατηρήσιμων μεταβλητών και των τρεχουσών εκτιμήσεων των παραμέτρων. Το δεύτερο βήμα είναι το Μ-βήμα (Maximization step) στο οποίο μεγιστοποιείται το κάτω φράγμα ως προς τις παραμέτρους $\Theta=\{(\pi_j, \theta_j), j=1,\ldots,M\}$ της μικτής κατανομής, υποθέτοντας πάντα ότι η κατανομή των κρυμμένων μεταβλητών που βρέθηκε στο Ε-βήμα, είναι σωστή. Αυτά τα δύο βήματα επαναλαμβάνονται μέχρι να υπάρξει σύγκλιση στην ακολουθία των παραμέτρων, δηλαδή όταν φτάσουμε σε κάποιο τοπικό μέγιστο.

Έστω τώρα ότι $X=\{x_1,\ldots,x_N\}, x_i \in \mathbb{R}^D$ είναι ένα σύνολο παρατηρήσεων το οποίο στο εξής θα το αποκαλούμε ως ελλιπές σύνολο δεδομένων. Ορίζουμε ως πλήρες σύνολο δεδομένων το $Y=(X, Z)$ όπου $Z=\{z_1,\ldots,z_N\}, z_i \in \{1,\ldots,M\}$ αναπαριστά τις Ν μη παρατηρήσιμες τιμές που υποδεικνύουν τον πυρήνα από τον οποίο παράχθηκε το κάθε πρότυπο του συνόλου Χ (π.χ. αν $z_i=j$, το x_i ανήκει στην j-οστή κατανομή). Από το τελευταίο συνάγεται ότι τα στοιχεία του Z βρίσκονται σε ένα προς ένα αντιστοιχία με τα διανύσματα του Χ, δηλαδή το x_i σχετίζεται με το z_i. Μπορούμε τώρα να ορίσουμε την από κοινού κατανομή $p(X,Z|\Theta)$ μεταξύ των παρατηρήσιμων και μη παρατηρήσιμων μεταβλητών. Με την βοήθεια αυτής της κατανομής, η συνάρτηση λογαριθμικής πιθανοφάνειας για τα πλήρη δεδομένα ορίζεται ως:

$$LLc(\Theta| Y) = LLc(\Theta| X, Z) =\log[p(X,Z; \Theta)] \qquad (1.26)$$

Επειδή όμως το σύνολο Y (συγκεκριμένα το Z) είμαι μη παρατηρήσιμο και επομένως η λογαριθμική πιθανοφάνεια LLc είναι ακαθόριστη, ο ΕΜ την λαμβάνει ως τυχαία μεταβλητή και υπολογίζει την αναμενόμενη τιμή της, ως προς την κατανομή q(Z) των μεταβλητών z_i, i=1,…,M. Η τελευταία, όπως θα δείξουμε στην συνέχεια, ισούται με $p(Z|X,\Theta^{(t)})$, όπου $\Theta^{(t)}$ είναι η τρέχουσα τιμή την παραμέτρων. Τα παραπάνω συνοψίζονται στη παρακάτω σχέση

$$G(\Theta,q(Z);\Theta^{(t)})=E_{Z|X,\Theta}\left[LLc(\Theta\mid X,Z)\right]=\sum_{Z} p(Z\mid X,\Theta)LLc(\Theta\mid X,Z) \qquad (1.27)$$

$$\text{όπου} \qquad p(Z\mid X,\Theta)=\frac{p(X,Z;\Theta)}{p(X)}$$

Στην συνέχεια θα εξετάσουμε τον τρόπο με τον οποίο προκύπτουν οι παραπάνω σχέσεις.

Με δεδομένη τώρα την από κοινού κατανομή p(X,Z;Θ) μπορούμε να επαναδιατυπώσουμε την λογαριθμική πιθανοφάνεια των ελλιπών δεδομένων με περιθωριοποίηση ως προς τις τιμές της μη παρατηρήσιμης μεταβλητής Z:

$$LL(\Theta\mid X)=\log p(X;\Theta)=\log\sum_{Z} p(X,Z;\Theta) \qquad (1.28)$$

Όπως αναφέρθηκε και προηγουμένως, το πρόβλημα με την μεγιστοποίηση της σχέσης (1.25) είναι ότι περιέχει το λογάριθμο ενός αθροίσματος και επιπλέον ότι η κρυμμένη πληροφορία Z είναι άγνωστη.

Η βασική ιδέα του τρόπου με τον οποίο γίνεται η βελτιστοποίηση μέσω του αλγορίθμου ΕΜ, είναι η κατασκευή ενός ευδιαχείριστου κάτω φράγματος G(Θ,q(Z)) της συνάρτηση LL(Θ|X), για το οποίο προφανώς θα ισχύει $G(\Theta,q(Z))\le LL(\Theta\mid X)$. Το κάτω αυτό φράγμα θα έχει ως παραμέτρους τα Θ και την κατανομή που ακολουθεί η κρυμμένη πληροφορία Z. Ιδανικά επιθυμούμε το G, σε αντίθεση με την πιθανοφάνεια LL(Θ|X), να περιέχει άθροισμα λογαρίθμων, αντί για λογάριθμο αθροίσματος. Για να κατασκευάσουμε μια τέτοια ποσότητα θα πρέπει πρώτα να ξαναγράψουμε την λογαριθμική πιθανοφάνεια με τον κάτωθι τρόπο:

$$LL(\Theta \mid X) = \log p(X;\Theta) \overset{(1.28)}{=} \log \sum_Z p(X,Z;\Theta) = \log \sum_Z q(Z) \frac{p(X,Z;\Theta)}{q(Z)} \qquad (1.29)$$

όπου q(Z) είναι προς το παρόν μια αυθαίρετη κατανομή για την μεταβλητή Z, την οποία θα προσδιορίσουμε στην συνέχεια. Εξαιτίας της κυρτότητας της συνάντησης του λογάριθμου μπορούμε να χρησιμοποιήσουμε την ανισότητα του Jensen για υπολογίσουμε το κάτω φράγμα G. Άρα με την βοήθεια της σχέσης (1.29) έχουμε:

$$\begin{aligned} LL(\Theta \mid X) &= \log \sum_Z q(Z) \frac{p(X,Z;\Theta)}{q(Z)} \\ &\geq \sum_Z q(Z) \log \frac{p(X,Z;\Theta)}{q(Z)} \qquad (\textit{Jensen's inequality}) \\ &= \sum_Z \left[q(Z) \log p(X,Z;\Theta) - q(Z) \log q(Z) \right] \\ &= G(\Theta, q(Z)) \end{aligned} \qquad (1.30)$$

Η ανισότητα (1.30) είναι αληθής για κάθε έγκυρη κατανομή q, παρόλα αυτά, εκείνο που ιδανικά επιθυμούμε, δεν είναι ένα οποιοδήποτε κάτω φράγμα, αλλά ένα βέλτιστο (ή σφιχτό) κάτω φράγμα της λογαριθμικής πιθανοφάνειας. Πιο συγκεκριμένα, επιδιώκουμε να προσδιορίσουμε εκείνη την κατανομή q(Z), ή οποία θα αναγκάσει το κάτω φράγμα G(Θ,q(Z)) να εφάπτεται στην συνάρτηση LL(Θ |X) στην τρέχουσα εκτίμηση των παραμέτρων $\Theta^{(t)}$ (Σχήμα 1.2), δηλαδή να ισχύει:

$$G(\Theta^{(t)}, q(Z)) = LL(\Theta^{(t)} \mid X) \qquad (1.31)$$

Αυτό θα μας εγγυηθεί, ότι μετά από κάθε βελτιστοποίηση της συνάρτησης G(Θ,q(Z)) ως

προς Θ, θα βελτιστοποιείται ταυτόχρονα και η συνάρτηση LL(Θ |X). Για να βρούμε τώρα το βέλτιστο κάτω φράγμα πρέπει να μεγιστοποιήσουμε την ποσότητα G(Θ,q(Z)) ως προς την άγνωστη κατανομή q(Z). Αυτό πραγματοποιείται ξαναγράφοντας το G(Θ,q(Z)) με το εξής τρόπο:

$$
\begin{aligned}
G(\Theta, q(Z)) &= \sum_{Z} q(Z) \log \frac{p(X, Z; \Theta)}{q(Z)} \\
&= E_Z \left[\log \frac{p(X, Z; \Theta)}{q(Z)} \right] \\
&= E_Z \left[\log \frac{p(Z \mid X, \Theta)\, p(X; \Theta)}{q(Z)} \right] \\
&= E_Z \left[\log \frac{p(Z \mid X, \Theta)}{q(Z)} + \log p(X; \Theta) \right] \\
&= E_Z \left[\log \frac{p(Z \mid X, \Theta)}{q(Z)} \right] + E_Z \left[\log p(X; \Theta) \right] \\
&= -E_Z \left[\log \frac{q(Z)}{p(Z \mid X, \Theta)} \right] + \log p(X; \Theta) \\
&= -D\left[q(Z) \,\|\, p(Z \mid X, \Theta) \right] + LL(\Theta \mid X)
\end{aligned}
\tag{1.32}
$$

Υποθέτοντας ότι η q(Z) είναι μια έγκυρη κατανομή, η ποσότητα $D\left[q(Z) \,\|\, p(Z \mid X, \Theta) \right]$ αναπαριστά την απόσταση Kullback-Leibler η οποία είναι ένα μέτρο απόστασης μεταξύ των κατανομών q(Z) και p(Z| X, Θ). Το μέτρο αυτό είναι εξ' ορισμού πάντοτε μη αρνητικό και ισούται με το μηδέν όταν οι δύο συγκρινόμενες κατανομές είναι ίδιες. Από αυτό συνεπάγεται ότι η ποσότητα G(Θ,q(Z)) μεγιστοποιείται ως προς q(Z) όταν $D\left[q(Z) \,\|\, p(Z \mid X, \Theta) \right]$ = 0, δηλαδή όταν ισχύει q(Z) = p(Z| X, Θ). Από την σχέση (1.32) είναι εύκολο να αποδείξουμε ότι όταν ισχύει η τελευταία ισότητα, το κάτω φράγμα G(Θ,q(Z)) είναι σφιχτό και ισούται (ή εφάπτεται) με την λογαριθμική πιθανοφάνεια στην τρέχουσα τιμή των παραμέτρων Θ.

Η παραπάνω διαδικασία προσδιορισμού της κατανομής q από την οποία προκύπτει το βέλτιστο κάτω φράγμα με δεδομένη την τρέχουσα τιμή των παραμέτρων Θ, αποτελεί το Ε-βήμα. Για να πάρουμε την επόμενη καλύτερη εκτίμηση των παραμέτρων, το οποίο αποτελεί το Μ-βήμα, πρέπει να μεγιστοποιήσουμε αυτό το κάτω φράγμα ως προς Θ, δοθείσας της κατανομής q που βρήκαμε στο Ε-βήμα. Το τελευταίο εξαρτάται από το εκάστοτε πρόβλημα και επιλύεται αρκετές φορές αναλυτικά. Από την σχέση (1.30) παρατηρούμε ότι ο όρος που πρέπει να μεγιστοποιήσουμε ως Θ για να μεγιστοποιηθεί αντίστοιχα και η ποσότητα G(Θ,q(Z)) είναι:

$$\begin{aligned}&\sum_{Z} q(Z)\log p(X,Z;\Theta)\\&=\sum_{Z} p(Z\mid X,\Theta)\log p(X,Z;\Theta)\\&=\mathrm{E}_{Z\mid X,\Theta}\left[\log p(X,Z;\Theta)\right]\end{aligned} \tag{1.33}$$

Ας υποθέσουμε τώρα ότι $\Theta^{(t)}$, $\Theta^{(t+1)}$ και $q^{(t)}$, $q^{(t+1)}$ συμβολίζουν τις τρέχουσες καλύτερες εκτιμήσεις για τις παραμέτρους Θ και q αντίστοιχα, τότε ο αλγόριθμος ΕΜ μπορεί να διατυπωθεί ως εξής:

Ε-Βήμα:

$$q^{(t+1)}(Z)=\arg\max_{q} G(\Theta^{(t)},q(Z))=p(Z\mid X,\Theta^{(t)}) \tag{1.34}$$

Μ-Βήμα:

$$\Theta^{(t+1)}=\arg\max_{\Theta} G(\Theta,q^{(t+1)}(Z))=\arg\max E_{Z\mid X,\Theta^{(t)}}[\log p(X,Z;\Theta)] \tag{1.35}$$

Μια γραφική αναπαράσταση ενός μοναδικού βήματος του αλγορίθμου ΕΜ φαίνεται και στο Σχήμα 1.2.

1.4.2. Σύγκλιση του αλγορίθμου ΕΜ

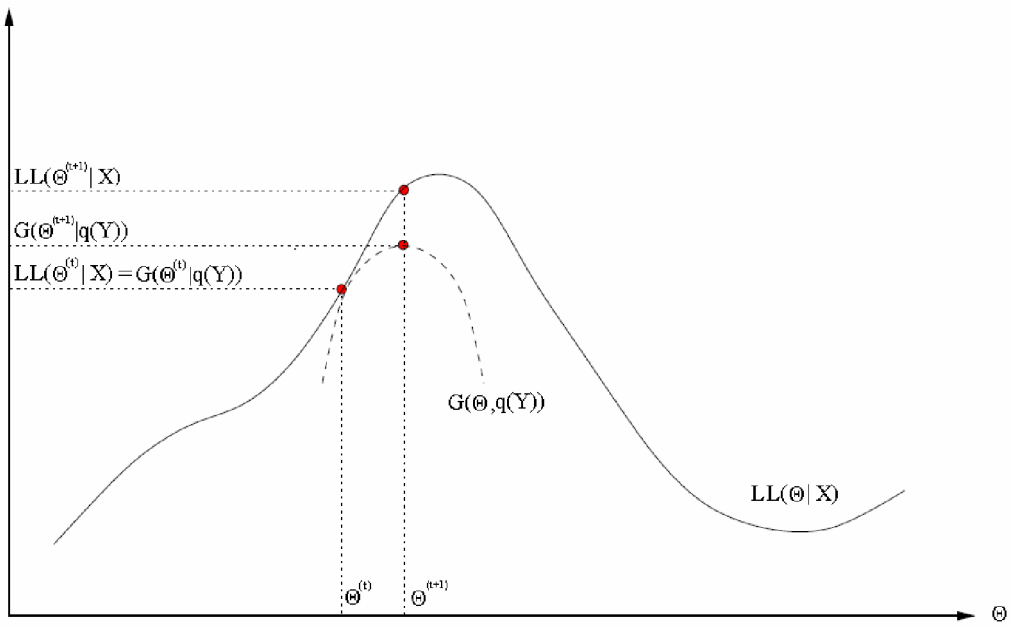
Σε αυτήν την παράγραφο θα μας απασχολήσει η σύγκλιση του αλγορίθμου ΕΜ σε μια γενική προσπάθεια απόδειξης της. Δεν θα σταθούμε στις ιδιότητες της σύγκλισης, για τις οποίες μπορεί κάποιος να αντλήσει πληροφορίες από το βιβλίο των McLachlan και Krishnan [5]. Υποθέτουμε, ότι $\Theta^{(t+1)}$ και $\Theta^{(t)}$ είναι οι εκτιμήσεις των παραμέτρων που παράχθηκαν από δύο διαδοχικά βήματα του ΕΜ. Από την στιγμή που το $\Theta^{(t+1)}$ επιλέχθηκε με τέτοιο τρόπο έτσι ώστε να μεγιστοποιεί το φράγμα G, αντιλαμβανόμαστε ότι ισχύει $G(\Theta^{(t+1)},q(Z))\geq G(\Theta^{(t)},q(Z))$ και επειδή το G είναι ένα σφιχτό κάτω φράγμα του LL έχουμε ότι $LL(\Theta^{(t+1)}\mid X)\geq G(\Theta^{(t+1)},q(Z))$ και επίσης $G(\Theta^{(t)},q(Z))=LL(\Theta^{(t)}\mid X)$. Άρα συνολικά έχουμε:

$$LL(\Theta^{(t+1)}\mid X)\geq G(\Theta^{(t+1)},q(Z))\geq G(\Theta^{(t)},q(Z))=LL(\Theta^{(t)}\mid X) \tag{1.35}$$

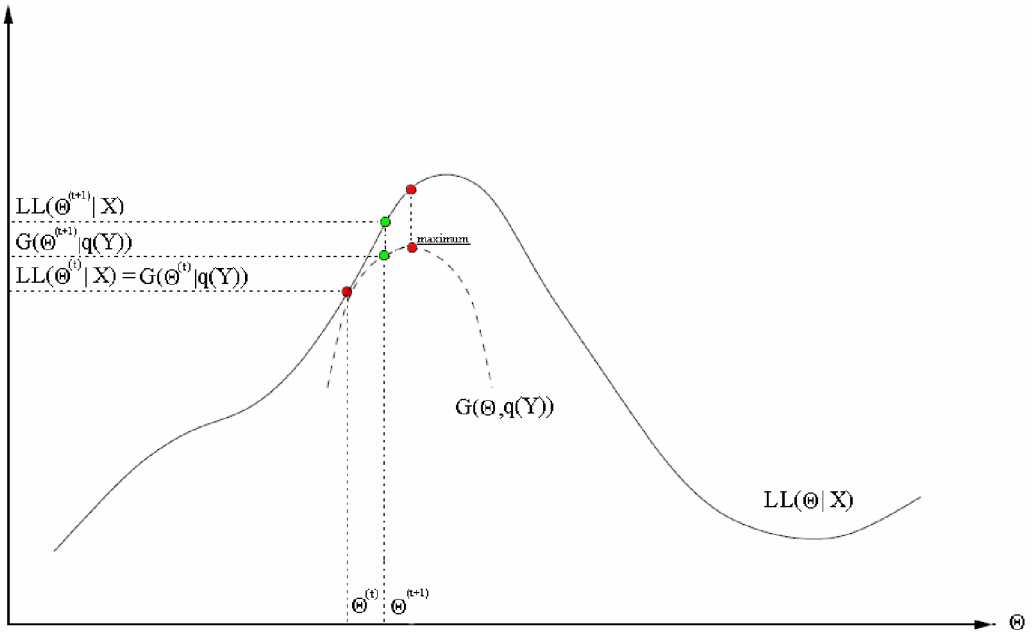
Από την παραπάνω σχέση επαγωγικά συνάγεται το συμπέρασμα ότι $LL(\Theta^{(t+1)} \mid X) \geq LL(\Theta^{(t)} \mid X)$ το οποίο δείχνει ότι η πιθανοφάνεια αυξάνει μονότονα. Η όλη διαδικασία θα συνεχιστεί μέχρι να ανιχνευτεί ένα τοπικό μέγιστο, όπου σε αυτή την περίπτωση η ακολουθία των τιμών που παράγονται από την πιθανοφάνεια σε κάθε βήμα, θα συγκλίνει και ο αλγόριθμος θα τερματίσει.

1.4.3. Ο αλγόριθμος GEM (Generalized EM)

Όπως είδαμε στην παρουσίαση του αλγορίθμου ΕΜ, στο Μ-βήμα του αλγορίθμου η νέα εκτίμηση των παραμέτρων $\Theta^{(t+1)}$ επιλέγεται έτσι ώστε το κάτω φράγμα $G(\Theta|q^{(t+1)}(Z))$ να μεγιστοποιείται. Εφόσον η αύξηση του φράγματος G προκαλεί ταυτόχρονη αύξηση και στην συνάρτηση της πιθανοφάνειας, θα μπορούσαμε να χαλαρώσουμε το προηγούμενο κριτήριο και αντί να απαιτούμε να βρεθεί το μέγιστο της συνάρτησης G ως προς Θ, να επιδιώκουμε απλώς να την αυξήσουμε. Με αυτό τον τρόπο επιτυγχάνουμε και πάλι αύξηση και της συνάρτησης της πιθανοφάνειας, το οποίο είναι και το ζητούμενο. Αυτή η προσέγγιση είναι γνωστή ως GEM (Generalized Expectation Maximization). Ο GEM είναι πολύ χρήσιμος σε περιπτώσεις όπου η μεγιστοποίηση του κάτω φράγματος G είναι δύσκολη ή δεν υπάρχει λύση διατυπωμένη σε κλειστή μορφή. Σε αυτή την περίπτωση χρησιμοποιούνται μέθοδοι αριθμητικής βελτιστοποίησης (gradient ascent) η χρήση των οποίων δεν οδηγεί πάντα στην εύρεση ολικού μεγίστου, αλλά σε μια καλύτερη εκτίμηση από το αρχικό σημείο εκκίνησης. Η απόδειξη της σύγκλισης του αλγορίθμου αυτού είναι παρόμοια με αυτή του αλγορίθμου ΕΜ και γι’ αυτό παραλείπεται. Ο τρόπος με τον οποίο λειτουργεί ο αλγόριθμος για μια επανάληψη, φαίνεται και στο Σχήμα 1.3.



Σχήμα 1.2 Γραφική απεικόνιση ενός βήματος του EM. Η συνάρτηση G(Θ,q(Z)) είναι το κάτω φράγμα της πιθανοφάνειας LL(Θ/X). Οι δύο συναρτήσεις είναι ίσες στο σημείο Θ(t). Στο Μ-βήμα το Θ(t+1) ορίζεται ως η τιμή του Θ που μεγιστοποιεί το G.



Σχήμα 1.3 Ο τρόπος λειτουργίας του αλγόριθμου GEM για μια επανάληψη. Παρατηρούμε ότι παρόλο που δεν ανιχνεύει το μέγιστο του κάτω φράγματος G, η πιθανοφάνεια αυξάνει.

1.4.4. Εφαρμογή του αλγορίθμου ΕΜ σε μικτές κανονικές κατανομές

Μέχρι τώρα παρουσιάσαμε τον αλγόριθμο ΕΜ σε ένα πιο γενικό θεωρητικό επίπεδο, παραβλέποντας λεπτομέρειες που προκύπτουν κατά την εφαρμογή του σε συγκεκριμένες εφαρμογές. Σε αυτή την παράγραφο θα εξετάσουμε τον τρόπο με τον οποίο μπορεί αυτός να εφαρμοστεί για την εκπαίδευση των παραμέτρων ενός μικτού μοντέλου μικτών κανονικών κατανομών.

Ανακαλούμε τον ορισμό ενός μικτού μοντέλου το οποίο διατυπώθηκε στην σχέση (1.11).

$$p(x;\Theta)=\sum_{j=1}^{M}\pi_j f(x;\theta_j)$$

Όπου στην περίπτωση των μικτών κανονικών κατανομών η ποσότητα $f(x;\theta_j)$ αποτελεί την συνάρτηση πυκνότητας πιθανότητας της j-οστής κανονικής κατανομής και π_j η αντίστοιχη prior πιθανότητα. Τόσο οι παράμετροι π_j όσο και οι παράμετροι θ_j για κάθε μια συνιστώσα κατανομή j, πρέπει να εκτιμηθούν.

Ε-Βήμα

Σύμφωνα με την σχέση (1.34) στο Ε-βήμα πρέπει να υπολογίσουμε την ποσότητα $p(Z|X,\Theta^{(t)})$, το οποίο είναι η κατανομή των μη παρατηρήσιμων μεταβλητών, δοθέντων των παρατηρήσιμων μεταβλητών και των τρεχουσών εκτιμήσεων των παραμέτρων Θ. Αυτό είναι ισοδύναμο με το να υπολογίσουμε τις πιθανότητες για κάθε ένα πρότυπο x_i να έχει παραχθεί από κάθε μια από τις Μ κατανομές, δηλαδή πρέπει να υπολογιστούν οι παρακάτω ποσότητες:

$$p(z_i=j\mid x_i)=\frac{\pi_j N(x_i\mid z_i=j)}{p(x_i)}=\frac{\pi_j N(x_i;\mu_j,\Sigma_j)}{\sum_{k=1}^{M}\pi_k N(x_i;\mu_k,\Sigma_k)} \qquad (1.36)$$

για κάθε δυνατό συνδυασμό του πλήθους Ν των προτύπων x_i και του πλήθους των ομάδων Μ.

Μ-βήμα

Παίρνοντας ως βάση μας τη σχέση (1.35), στο Μ-βήμα θα πρέπει να μεγιστοποιήσουμε την ποσότητα $E_{Z|X,\Theta^{(t)}}[\log p(X,Z|\Theta)]$ ως προς τις παραμέτρους $\Theta=\{\pi_j, (\mu_j, \Sigma_j)\}$, j=1,…,M. Στην περίπτωση των μικτών μοντέλων κανονικών κατανομών αυτό μπορεί να πραγματοποιηθεί αναλυτικά σε κλειστή μορφή:

$$\begin{aligned} E_{p(Y|X,Y)}[\log p(X,Y;\Theta)] &= \sum_{i=1}^{N}\sum_{j=1}^{M} p(z_i=j|x_i)\log[\pi_j p(x_i;\theta_j)] \\ &= \sum_{i=1}^{N} p(z_i=j|x_i)\log p(x_i;\theta_j) \\ &+ \sum_{i=1}^{N} p(z_i=j|x_i)\log \pi_j \end{aligned} \tag{1.37}$$

Η ποσότητα $p(z_i=j|x_i)$ έχει ήδη υπολογιστεί στο Ε-βήμα, και άρα η αντικειμενική συνάρτηση (1.37) είναι ένα άθροισμα δύο ασυσχέτιστων μεταξύ τους όρων. Ο πρώτος εμπλέκει την συνάρτηση πυκνότητας πιθανότητας της j-οστής κανονικής κατανομής και ο δεύτερος την αντίστοιχη prior πιθανότητα. Άρα για να υπολογίσουμε τις φόρμουλες ανανέωσης για την εύρεση των παραμέτρων της j-οστής συνιστώσας κατανομής θα πρέπει να μεγιστοποιήσουμε τον πρώτο όρο, ενώ για την prior πιθανότητα αρκεί η μεγιστοποίηση του δευτέρου. Εφαρμόζοντας τα παραπάνω για μια μίξη κανονικών κατανομών προκύπτουν οι εξής τύποι ανανέωσης.

$$\pi_j^{new} = \frac{1}{N}\sum_{i=1}^{N} p(z_i=j|x_i) \tag{1.38}$$

$$\mu_j^{new} = \frac{\sum_{i=1}^{N} x_i\, p(z_i=j|x_i)}{\sum_{i=1}^{N} p(z_i=j|x_i)} \tag{1.39}$$

$$\Sigma_j^{new} = \frac{\sum_{i=1}^{N} p(z_i=j|x_i)(x_i-\mu_j^{new})(x_i-\mu_j^{new})^{\mathrm{T}}}{\sum_{i=1}^{N} p(z_i=j|x_i)} \tag{1.40}$$

Εκτελώντας αυτές τις ενημερώσεις για j=1,…,M σε κάθε βήμα του ΕΜ μεγιστοποιείται η συνάρτηση της πιθανοφάνειας. Παρατηρείστε ότι οι εξισώσεις (1.39)-(1.40) είναι παρεμφερείς με αυτές που χρησιμοποιούμε για να εκτιμήσουμε τις

παραμέτρους μιας μοναδικής κανονικής κατανομής. Η μόνη διαφορά είναι ότι εδώ το κάθε πρότυπο x_i, είναι σταθμισμένο με την posterior πιθανότητα $p(z_i=j|x_i)$ βάσει της οποίας παράγεται από την εκάστοτε κατανομή.

Το βασικότερο μειονέκτημα του ΕΜ είναι ότι εξαρτάται από τις αρχικές τιμές των παραμέτρων. Στην περίπτωση που αυτές δεν είναι αρκετά κοντά στην τελική επιθυμητή λύση, ο αλγόριθμος θα μας δώσει κακής ποιότητας αποτελέσματα. Για την καλή αρχικοποίηση, συνήθως χρησιμοποιούμε ένα άλλο αλγόριθμο ο οποίος θα μας δώσει μια αρχικά αποδεκτή λύση από την οποία θα συνεχίσει ο ΕΜ για να μας δώσει την τελική εκτίμηση των παραμέτρων.

Τα πλεονεκτήματα του ΕΜ είναι η απλότητα του καθώς και σίγουρη σύγκλιση του σε κάποιο στάσιμο σημείο. Επίσης έχουμε άμεση ικανοποίηση των παρακάτω περιορισμών: $\pi_j \geq 0$ και $\sum_{j=1}^{K} \pi_j = 1$ καθώς επίσης και ότι ο πίνακας συμμεταβλητότητας που παράγεται από την σχέση (1.40) είναι συμμετρικός και θετικά ημιορισμένος.

Αναφορικά με την ταχύτητα σύγκλισης θα λέγαμε ότι σε γενικές γραμμές αυξάνει με γρήγορους ρυθμούς την πιθανοφάνεια στα πρώτα του βήματα, ενώ η σύγκλιση γίνεται αισθητά αργή καθώς πλησιάζει σε ένα τοπικό μέγιστο. Γενικότερα αποδίδει καλύτερα σε δεδομένα μικρής διάστασης, ενώ δεν υπάρχει κάποια αποδεδειγμένη θεωρία στην βιβλιογραφία που να πιστοποιεί ότι η απόδοσή του είναι καλύτερη από άλλες μεθόδους αριθμητικής βελτιστοποίησης (π.χ. Quasi-Newton).

1.4.5. Ο αλγόριθμος Greedy EM

Όπως αναφέραμε και στην προηγούμενη ενότητα ο αλγόριθμος ΕΜ εξαρτάται σημαντικά από τις αρχικές τιμές των παραμέτρων. Προκειμένου να ξεπεραστεί αυτό το σημαντικό πρόβλημα, το οποίο επηρεάζει κατά πολύ την απόδοση του αλγορίθμου, δημιουργήθηκε μια παραλλαγή του, ο αλγόριθμος Greedy EM [2]. Η φιλοσοφία του νέου αυτού αλγορίθμου στηρίζεται στο ότι μπορούμε να

εκπαιδεύσουμε τις παραμέτρους ενός μικτού μοντέλου, προσθέτοντας αυξητικά συνιστώσες κατανομές στο μοντέλο μας, έως ότου φτάσουμε σε ένα επιθυμητό αριθμό κατανομών K. Ειδικότερα, υποθέτουμε ότι μια νέα κατανομή $\phi(x;\theta)$ προστίθεται σε ένα μικτό μοντέλο $f_k(x)$ απαρτιζόμενο από K συνιστώσες, παράγοντας την παρακάτω μίξη:

$$f_{k+1}(x) = (1-a)\, f_k(x) + a\phi(x;\theta) \tag{1.41}$$

Όπου το $a \in (0,1)$. Συνεπώς η επιδίωξη μας πλέον είναι, με δεδομένη την κατανομή $f_k(x)$, να προσδιορίσουμε εκείνες τις τιμές του βάρους α και των παραμέτρων θ της $\phi(x;\theta)$, για τις οποίες η συνάρτηση της πιθανοφάνειας

$$L_{k+1} = \sum_{i=1}^{N} \log f_{k+1}(x_i) = \sum_{i=1}^{N} \log[(1-a)\, f_k(x) + a\phi(x;\theta)] \tag{1.42}$$

να μεγιστοποιείται.

Δεν θα επεκταθούμε σε περισσότερο στην θεωρητική μελέτη του αλγορίθμου, αφού ο αναγνώστης μπορεί να ανατρέξει στην αντίστοιχη αναφορά [2] για σχετικές πληροφορίες. Παρακάτω θα παρουσιάσουμε την περιγραφή και τον τρόπο υλοποίησης του με ένα συνοπτικό τρόπο. Η σχετική παρουσίαση θα γίνει για την εκπαίδευση ενός μικτού μοντέλου κανονικών κατανομών.

Χρησιμοποιώντας αρχικά ένα μοντέλο με μια μόνο κατανομή, αρχικοποιούμε της παραμέτρους του με τον εξής τρόπο: το κέντρο της κατανομής ισούται με την μέση τιμή των δεδομένων δηλ. μ = E[X] και ο πίνακας συμμεταβλητότητας, Σ=cov(X). Η εκ των προτέρων πιθανότητα του νέου αυτού πυρήνα είναι μονάδα. Η τιμή της παραμέτρου σ καθορίζεται από την παρακάτω εμπειρική σχέση

$$\sigma = \beta\left[\frac{4}{(D+2)\mathrm{N}}\right]^{1/(D+4)} \tag{1.43}$$

Όπου η τιμή της παραμέτρου β τίθεται ίση με το μισό της μέγιστης ιδιοτιμής του πίνακα Σ. Στην συνέχεια υπολογίζουμε τον πίνακα ομοιότητας K=[k_{ij}]:

$$k_{ij} = (2\pi\sigma^2)^{-\frac{D}{2}} \exp(-\frac{\left\|x_i - x_j\right\|^2}{2\sigma^2}) \tag{1.44}$$

όπου $x_i, x_j \in X$

Εκτελούμε τον κλασικό αλγόριθμο ΕΜ έως ότου επέλθει σύγκλιση δηλαδή μέχρι να ικανοποιηθεί το παρακάτω κριτήριο:

$$\left| \frac{L_k^{(t)}}{L_k^{(t-1)}} - 1 \right| < 10^{-6} \tag{1.45}$$

Κάνουμε καθολική αναζήτηση σε όλα τα πρότυπα $x_i \in X$ για να εντοπίσουμε το υποψήφιο κέντρο μ του νέου πυρήνα και επιλέγουμε εκείνο το πρότυπο από το σύνολο δεδομένων που μεγιστοποιεί την ποσότητα

$$\hat{L}_{k+1} = \sum_{i=1}^{N} \log \frac{f_k(x_i) + \phi(x_i; \mu, \Sigma)}{2} + \frac{1}{2} \frac{\left[\sum_{i=1}^{N} \delta(x_i; \mu, \Sigma)\right]^2}{\sum_{i=1}^{N} \delta^2(x_i; \mu, \Sigma)} \tag{1.46}$$

Όπου

$$\delta(x; \theta) = \frac{f_k(x) - \phi(x; \theta)}{f_k(x) + \phi(x; \theta)} \tag{1.47}$$

για το οποίο κάνουμε την παραδοχή $\Sigma=\sigma^2 I$ και το σ^2 δίνεται από την (1.43). Το τελευταίο γίνεται σκόπιμα και αποσκοπεί στην απλούστευση της διαδικασίας βελτιστοποίησης, αφού πλέον θα πρέπει να βελτιστοποιήσουμε μόνο ως προς το κέντρο μ του νεοεισελθέντος πυρήνα. Για να πραγματοποιήσουμε την διαδικασία βελτιστοποίησης, τοποθετούμε στην θέση της κατανομής $\phi(x_i; \mu = x_j, \Sigma)$ την ποσότητα k_{ij} που υπολογίσαμε στο βήμα 1. Επιλέγουμε ως κέντρο $\hat{\mu}$ του νέου πυρήνα, το πρότυπο x_i που μεγιστοποιεί τη ποσότητα: $\hat{L}_{k+1}(x_i)$, δηλαδή

$$\hat{\mu} = \arg\max_{x_i \in X} \hat{L}_{k+1}(x_i) \tag{1.48}$$

Αρχικοποιούμε τον αλγόριθμο partial EM με την εκτιμούμενη τιμή του μ που βρέθηκε στο βήμα 3, και τον πίνακα συμμεταβλητότητας $\Sigma=\sigma^2 I$. Στην συνέχεια υπολογίζουμε την παράμετρο $\hat{a}$ από την παρακάτω εξίσωση:

$$\hat{a} = \frac{1}{2} - \frac{1}{2}\frac{\sum_{i=1}^{N}\delta(x_i;\mu,\Sigma)}{\sum_{i=1}^{N}\delta^2(x_i;\mu,\Sigma)} \tag{1.49}$$

Σε περίπτωση που αυτή η ποσότητα είναι εκτός διαστήματος (0,1), αρχικοποιούμε τον αλγόριθμο partial EM του επόμενου βήματος θέτοντας: $\hat{a}$=0.5 για k=1, και $\hat{a}$=2/(k+1) για k>1.

Εφαρμόζουμε το αλγόριθμο partial EM με παραμέτρους α, μ και Σ.με αρχικές τιμές $\hat{a}$, $\hat{\mu}$ και $\hat{\Sigma}=\sigma^2 I$ για τον οποίο εκτελείται η ακόλουθη διαδικασία:

E-step

$$p(k+1 \mid x_i) = \frac{a\phi(x_i;\mu,\Sigma)}{(1-a)\, f_k(x_i) + a\phi(x_i;\mu,\Sigma)} \tag{1.50}$$

M-step

$$a = \frac{1}{N}\sum_{i=1}^{N} p(k+1 \mid x_i) \tag{1.51}$$

$$\mu = \frac{\sum_{i=1}^{N} p(k+1 \mid x_i) x_i}{\sum_{i=1}^{N} p(k+1 \mid x_i)} \tag{1.52}$$

$$\Sigma = \frac{\sum_{i=1}^{N} p(k+1 \mid x_i)(x_i-\mu)(x_i-\mu)^{\mathrm{T}}}{\sum_{i=1}^{N} p(k+1 \mid x_i)} \tag{1.53}$$

Εφαρμόζουμε επαναληπτικά τα βήματα αυτά έως ότου επέλθει σύγκλιση (αντίστοιχα όπως και στο βήμα 2).

Αν $L_{k+1} < L_k$, (όπου L_{k+1} είναι η λογαριθμική πιθανοφάνεια μετά την εισαγωγή του νέου πυρήνα στο μοντέλο, L_k είναι η λογαριθμική πιθανοφάνεια στο παλιό μοντέλο) ο αλγόριθμος τερματίζει. Ένα άλλο, πιθανό κριτήριο τερματισμού, αποτελεί και το γεγονός της υπέρβασης του πλήθους των συνολικών πυρήνων από ένα μέγιστο καθορισμένο αριθμό k_{max}. Αν καμία από τις παραπάνω συνθήκες δεν ικανοποιείται, τότε εισάγουμε τον νέο πυρήνα στο μικτό μοντέλο και μεταβαίνουμε πάλι στο βήμα 2.

Θα πρέπει εν κατακλείδι να σημειωθεί, ότι εφόσον ο αλγόριθμος EM δεν μπορεί να οδηγήσει σε μείωση της πιθανοφάνειας και η λύση του partial EM γίνεται αποδεκτή μόνο όταν ισχύει $L_{k+1} > L_k$, ο αλγόριθμος greedy EM μπορεί να μας εγγυηθεί για την μονότονη αύξηση της λογαριθμικής πιθανοφάνειας του συνόλου εκπαίδευσης.

ΚΕΦΑΛΑΙΟ 2. Η ΚΑΤΑΝΟΜΗ Π-SIGMOID

2.1. Γενικά

Σε αυτό το κεφάλαιο θα παρουσιαστεί αναλυτικά η νέα συνάρτηση πυκνότητας πιθανότητας Π-sigmoid. Γίνεται περιγραφή των ιδιοτήτων της, καθώς επίσης και της συμπεριφοράς που αυτή επιδεικνύει για τις διάφορες τιμές των παραμέτρων της. Θα πρέπει να σημειωθεί ότι η σχετική ανάλυση περιλαμβάνει αναφορά τόσο στη μονοδιάστατη όσο και στη πολυδιάστατη εκδοχή, ενώ στη δεύτερη μελετάται και ο τρόπος χειρισμού "περιστραμένων" δεδομένων, γεγονός που ολοκληρώνει την σχετική παρουσίαση.

Κατ' αρχάς, είναι σκόπιμο να αναφερθεί ο λόγος και οι αφορμές που δόθηκαν για την δημιουργία αυτής της νέας κατανομής. Ίσως, δεν θα ήταν σωστό να περιορίσουμε την παρουσίαση και χρήση της σε συγκεκριμένα πεδία και εφαρμογές. Παρόλα αυτά όμως, θα είχε εξαιρετικό ενδιαφέρον να την εξετάσουμε υπό το πρίσμα ζητημάτων που διαπραγματεύονται σ' ένα από τα σύγχρονα πεδία της τεχνητής νοημοσύνης, το data-mining ([8] [6]). Πιο συγκεκριμένα, η συνεχής αύξηση του όγκου δεδομένων στις βάσεις καθώς επίσης και η ραγδαία διεύρυνση του διαδικτύου, οδήγησαν στην συσσώρευση πολλών και ανεπιθύμητων στοιχείων, τα οποία κατέστησαν την αναζήτηση χρήσιμων πληροφοριών δυσχερή και χρονοβόρα. Στην προσπάθεια επίλυσης αυτού του προβλήματος, επιστρατεύονται τεχνικές και μέθοδοι ομαδοποίησης για την εξαγωγή και συγκέντρωση των ωφέλιμων δεδομένων όπως

επίσης και την ανίχνευση “τάσεων” σε θέματα που αφορούν το marketing, την οικονομία, το χρηματιστήριο κ.α.. Το πιο γνωστό μοντέλο που χρησιμοποιείται για την επίτευξη των παραπάνω είναι η μικτή κανονική κατανομή (Gaussian Mixture model, GMM), το οποίο έχει περιγραφεί στο προηγούμενο κεφάλαιο. Είναι ένα πολύ ισχυρό μοντέλο, γιατί τόσο η ικανότητα του να χειρίζεται αρκετούς διαφορετικούς τύπους δεδομένων, όσο και οι καλές αναλυτικές ιδιότητες τους, το καταστούν ένα σημαντικό εργαλείο για την διαπραγμάτευση τέτοιων ζητημάτων.
Ωστόσο, και σε αυτή την προσέγγιση υπάρχουν και κάποια σημαντικά μειονεκτήματα που αφορούν την ίδια την φύση της κανονικής κατανομής. Τα αποτελέσματα της ομαδοποίησης που δίνει, δηλαδή ο πίνακας συμμεταβλητότητας και το κέντρο της κάθε ομάδας, δεν είναι ερμηνεύσιμα και κατανοητά από τον άνθρωπο. Οι άνθρωποι προτιμούν την περιγραφή δεδομένων σε μορφή κανόνων, δηλαδή τον χωρισμό του χώρου δεδομένων σε ορθογώνιες περιοχές. Επίσης, παρά την ευελιξία που παρέχει η μικτή κανονική κατανομή, αναφορικά με τους διάφορους τύπους δεδομένων που μπορεί να περιγράψει, αδυνατεί να ανταποκριθεί ικανοποιητικά σε δεδομένα που είναι ομοιόμορφα κατανεμημένα.

Η προτεινόμενη κατανομή Π-sigmoid επιλύει τα προαναφερθέντα προβλήματα με αρκετά ικανοποιητικό τρόπο. Είναι ευέλικτη, και αυτό της επιτρέπει να περιγράψει επαρκώς, δεδομένα πολλών τύπων, όπως για παράδειγμα ομοιόμορφα, γκαουσιανά και ενδεχομένως αρκετά άλλα. Η ιδιότητα αυτή, της προσδίδει καλή ικανότητα μοντελοποίησης και άρα την καθιστά κατάλληλη στην περιγραφή δεδομένων με άγνωστες στατιστικές ιδιότητες. Επιπλέον, δίνει την δυνατότητα παραγωγής κατανοητών και ερμηνεύσιμων κανόνων, αφού, όπως θα δειχτεί παρακάτω, τα σχετικά όρια που παρέχει η κατανομή, είναι υπερεπίπεδα.

2.1.1. Η σιγμοειδής συνάρτηση

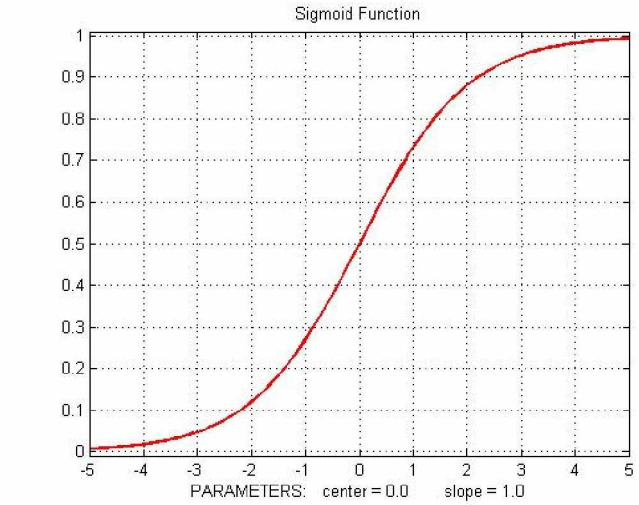
Σε αυτό το σημείο, πριν προχωρήσουμε στην περιγραφή και τον ορισμό της Π-sigmoid, κρίνεται απαραίτητο να αναφέρουμε κάποια στοιχεία για την σιγμοειδή συνάρτηση, η οποία είναι γνωστή και από στον τομέα των νευρωνικών δικτύων και ως λογιστική. Η σύντομη αυτή αναφορά σε αυτή τη συνάρτηση γίνεται επειδή αποτελεί βασικό δομικό στοιχείο στην διατύπωση και τον ορισμό της Π-sigmoid

κατανομής και συνεπώς η κατανόηση των ιδιοτήτων της θα συμβάλει σημαντικά στην καλύτερη και πληρέστερη παρουσίαση της εν λόγω κατανομής.

Η σιγμοειδής λογιστική συνάρτηση πήρε το όνομα της, από τη καμπύλη που παράγει η οποία έχει το σχήμα του γράμματος “S”. Ο τύπος της δίνεται από την παρακάτω σχέση:

$$f(x) = \frac{1}{1+e^{-x}} \tag{2.1}$$

Το κέντρο της σιγμοειδούς με τον τρόπο που διατυπώθηκε στην σχέση (2.1) είναι το μηδέν. Παρατηρούμε ότι καθώς το x τείνει στο μείον άπειρο η τιμή της συνάρτησης τείνει στο μηδέν, ενώ όταν το x τείνει στο άπειρο η συνάρτηση τείνει στη μονάδα. Με λίγα λόγια η συνάρτηση έχει άνω φράγμα το μηδέν και κάτω φράγμα τη μονάδα. Η γραφική της παράσταση φαίνεται παρακάτω:

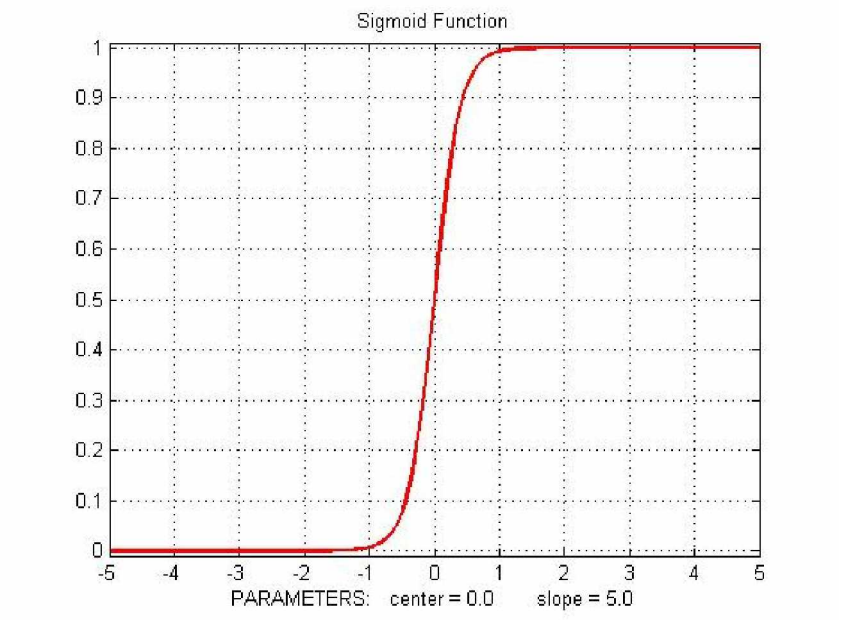


Σχήμα 2.1 Η σιγμοειδής συνάρτηση με κέντρο το μηδέν και κλίση μονάδα.

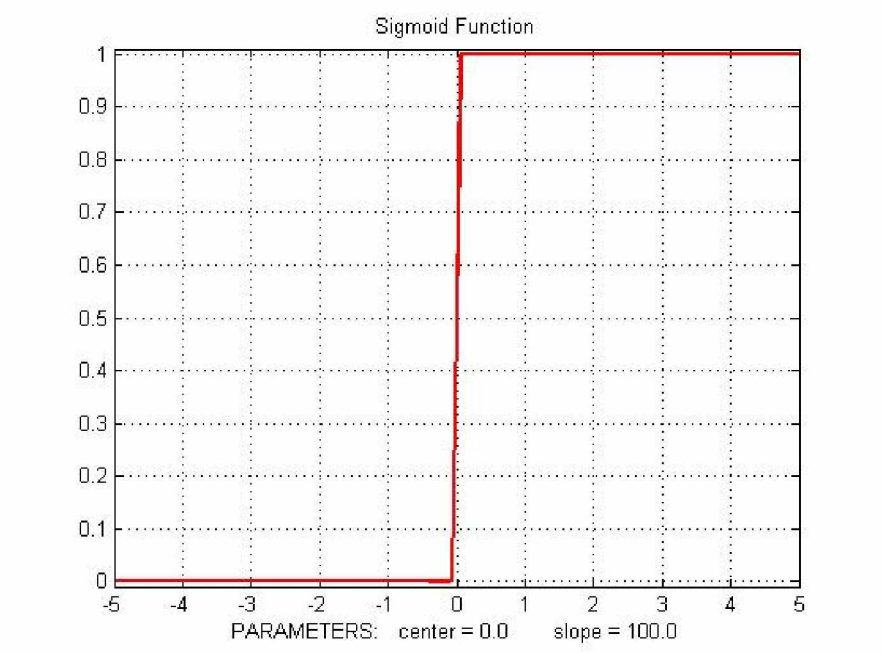
Ο τύπος (2.1) αποτελεί τη γενική μορφή μια σιγμοειδούς συνάρτησης. Θα ήταν πολύ χρήσιμο να κάνουμε κάποια τροποποίηση έτσι ώστε να υπάρχει δυνατότητα μετατόπισης του κέντρου της από το μηδέν σε οποιοδήποτε άλλο σημείο. Επίσης, η αλλαγή της κλίσης στο κέντρο της, θα προσέδιδε σημαντικές δυνατότητες στη συνάρτηση. Οι παραπάνω ιδιότητες συνοψίζονται στην παρακάτω φόρμουλα:

$$f(x) = \frac{1}{1+e^{-\lambda(x-a)}} \tag{2.2}$$

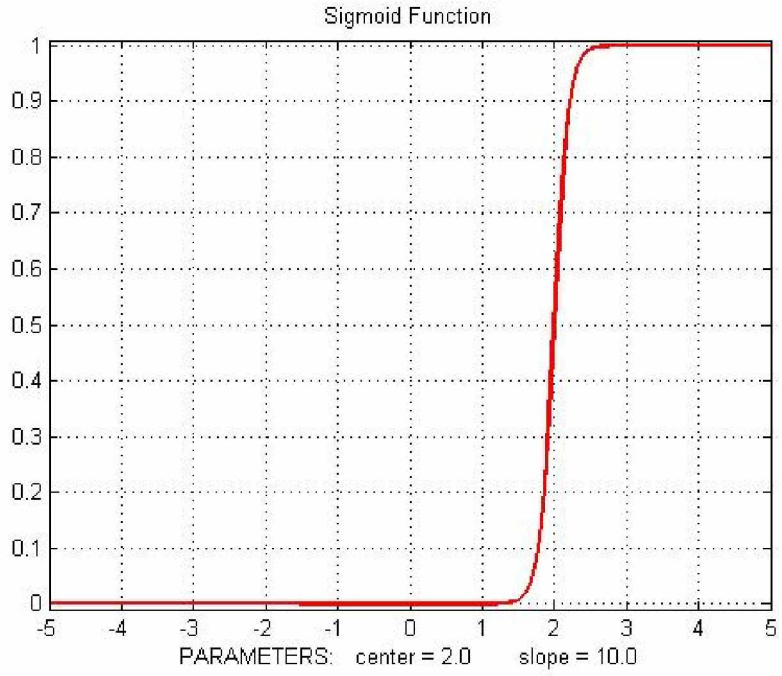
όπου α είναι το νέο κέντρο της συνάρτησης και λ η κλίση της συνάρτησης στο σημείο α. Να σημειωθεί ότι για λ=1 και α=0 παίρνουμε την σχέση (2.1). Από την νέα αυτή προσέγγιση, παρατηρούμε ότι για μεγάλες τιμές του λ θα έχουμε μια απότομη μετάβαση της συνάρτησης από το μηδέν στη μονάδα, ενώ η συμπεριφορά της για τιμές του x έξω από την περιοχή του κέντρου (δηλ. x << a και x >> a) τείνει να είναι σταθερή. Τέλος, για λ=0 η συνάρτηση γίνεται σταθερή με f(x) = 1/2. Μερικά δείγματα της γραφικής παράστασης της συνάρτησης για κάποιες επιλεγμένες τιμές των παραμέτρων φαίνονται και στα παρακάτω σχήματα.



Σχήμα 2.2 Η γραφική παράσταση της σιγμοειδούς συνάρτησης με κέντρο α=0 και κλίση λ=5.



Σχήμα 2.3 Η γραφική παράσταση της σιγμοειδούς συνάρτησης με κέντρο a=0 και κλίση λ=100. Φαίνεται καθαρά η απότομη μεταβολή της συνάρτησης από το μηδέν στην μονάδα καθώς και η γραμμική συμπεριφορά που επιδεικνύει.



Σχήμα 2.4 Η γραφική παράσταση της σιγμοειδούς συνάρτησης με κέντρο a=2 και κλίση λ=10.

Ύστερα από την παρουσίαση της βασικής συμπεριφοράς της σιγμοειδούς συνάρτησης, παρακάτω παρουσιάζονται κάποιες βασικές ιδιότητες που ικανοποιούνται:

Ιδιότητα συμμετρίας

$$\begin{aligned}\sigma(-x) &= \frac{1}{1+e^{x}} = \frac{e^{-x}}{1+e^{-x}} = \frac{1-1+e^{-x}}{1+e^{-x}} = \frac{1+e^{-x}}{1+e^{-x}} - \frac{1}{1+e^{-x}} \\ &= 1-\sigma(x)\end{aligned} \tag{2.3}$$

Αντίστροφη συνάρτηση

$$\sigma^{-1}(x) = \ln\left[\frac{x}{1-x}\right] \tag{2.4}$$

Ιδιότητα παραγώγου

$$\begin{aligned}\frac{d\sigma}{dx} &= \frac{e^{-x}}{(1+e^{-x})^{2}} = \frac{1}{1+e^{-x}}\frac{e^{-x}}{1+e^{-x}} \\ &= \sigma(x)\frac{e^{-x}}{1+e^{-x}} = \sigma(x)\frac{1}{1+e^{x}} = \sigma(x)\sigma(-x) \\ &\overset{(2.3)}{=} \sigma(x)(1-\sigma(x))\end{aligned} \tag{2.5}$$

Η τελευταία ιδιότητα (2.5) χαρακτηρίζεται πολύ σημαντική, μιας και μας επιτρέψει να εκφράσουμε την παράγωγό της σιγμοειδούς συνάρτησης, συναρτήσει του ίδιου της του εαυτού. Χρησιμοποιώντας αυτή την έκφραση της παραγώγου, πολύ εύκολα μπορεί να υπολογιστεί και η παράγωγος της τροποποιημένης σιγμοειδούς στην μορφή που παρουσιάζεται στην σχέση (2.2). Έστω z = λ(x-a), τότε από τον κανόνα της αλυσίδας έχουμε:

$$\frac{d\sigma}{dx} = \frac{d\sigma}{dz}\frac{dz}{dx} \overset{(2.5)}{=} \sigma(x)(1-\sigma(x))\lambda \tag{2.6}$$

2.2. Ορισμός της συνάρτησης πυκνότητας πιθανότητας Π-sigmoid

Ύστερα από την αναφορά που έγινε στην σιγμοειδή συνάρτηση, μπορούμε πλέον να προχωρήσουμε στον ορισμό της κατανομής Π-sigmoid.

Ορισμός: Η συνάρτηση πυκνότητας πιθανότητας Π-sigmoid ορίζεται ως η διαφορά δύο σιγμοειδών συναρτήσεων με κέντρα a , b όπου b > a. Ο τύπός της δίνεται από την σχέση:

$$f(x) = \left(\frac{1}{b-a}\right)\left[\frac{1}{1+e^{-\lambda(x-a)}} - \frac{1}{1+e^{-\lambda(x-b)}}\right], \quad \lambda > 0 \qquad (2.7)$$

το λ είναι η κλίση των δύο σιγμοειδών συναρτήσεων στα κέντρα τους a, b αντίστοιχα, η οποία θα πρέπει να είναι πάντα μεγαλύτερη του μηδενός. Για να δηλώσουμε ότι μια τυχαία μεταβλητή X ακολουθεί την Π-sigmoid κατανομή με κλίση λ και κέντρα a και b, θα γράφουμε: $X \sim \Pi s\ (a, b, \lambda)$

Ο παράγοντας 1/(b-a) είναι η σταθερά κανονικοποίησης, η οποία εγγυάται ότι το ολοκλήρωμα της συνάρτησης είναι μονάδα, δηλαδή:

$$\int_{-\infty}^{\infty} f(x)\,dx = 1 \qquad (2.8)$$

Θα αποδείξουμε τώρα το πώς προκύπτει η σταθερά κανονικοποίησης K_σ. Καταρχάς για να «αναγκάσουμε» το ολοκλήρωμα μιας συνάρτησης να γίνει μονάδα, θα πρέπει να τη διαιρέσουμε με το ολοκλήρωμα της, το οποίο υπολογίζεται από το μείον άπειρο έως το άπειρο. Άρα πρέπει να υπολογίσουμε την παρακάτω ποσότητα:

$$K_\sigma^{-1} = \int_{-\infty}^{\infty} \frac{1}{1+e^{(-\lambda(x-a))}} - \frac{1}{1+e^{(-\lambda(x-b))}}\,dx = \int_{-\infty}^{\infty} \frac{1}{1+e^{(-\lambda(x-a))}}\,dx - \int_{-\infty}^{\infty} \frac{1}{1+e^{(-\lambda(x-b))}}\,dx \qquad (2.9)$$

Υπολογίζω το πρώτο ολοκλήρωμα της σχέσης (2.9) και ακολούθως το δεύτερο με παρόμοιο τρόπο:

$$\int_{-\infty}^{\infty} \frac{1}{1+e^{(-\lambda(x-a))}}dx = \int_{-\infty}^{\infty} \frac{1}{1+\frac{1}{e^{\lambda(x-a)}}}dx = \int_{-\infty}^{\infty} \frac{1}{\frac{1+e^{\lambda(x-a)}}{e^{\lambda(x-a)}}}dx$$

$$= \int_{-\infty}^{\infty} \frac{e^{\lambda(x-a)}}{1+e^{\lambda(x-a)}}dx = \frac{1}{\lambda}\int_{-\infty}^{\infty} \frac{\lambda e^{\lambda(x-a)}}{1+e^{\lambda(x-a)}}dx \qquad (2.10)$$

Στην τελευταία σχέση παρατηρούμε ότι ο αριθμητής ισούται με την παράγωγο του παρονομαστή. Άρα θέτω $t = 1+e^{\lambda(x-a)}$ και $dt = \lambda e^{\lambda(x-a)}$ και με αντικατάσταση έχω:

$$\frac{1}{\lambda}\int_{-\infty}^{\infty} \frac{1}{t}dt = \frac{1}{\lambda}\Big|_{-\infty}^{\infty} \ln(t) = \frac{1}{\lambda}\Big|_{-\infty}^{\infty} \ln(1+e^{\lambda(x-\alpha)})$$

$$= \frac{1}{\lambda}\Big|_{-\infty}^{\infty} \ln(1+\frac{1}{e^{-\lambda(x-\alpha)}}) = \frac{1}{\lambda}\Big|_{-\infty}^{\infty} \ln(\frac{1+e^{-\lambda(x-\alpha)}}{e^{-\lambda(x-\alpha)}})$$

$$= \frac{1}{\lambda}\Big|_{-\infty}^{\infty} \ln(1+e^{-\lambda(x-\alpha)}) - \ln(e^{-\lambda(x-\alpha)})$$

$$= \Big|_{-\infty}^{\infty} \frac{1}{\lambda}\ln(1+e^{-\lambda(x-\alpha)}) + x - \alpha \qquad (2.11)$$

Από τις σχέσεις (2.9) και (2.11) προκύπτει ότι:

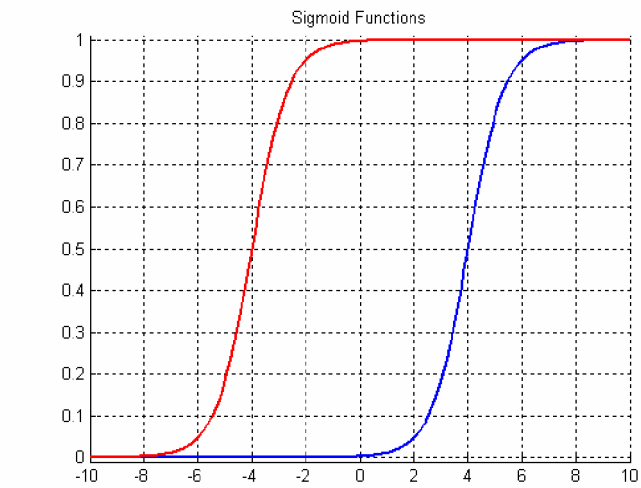
$$K_{\sigma}^{-1} = \Big|_{-\infty}^{\infty}\left[\frac{1}{\lambda}\ln(1+e^{-\lambda(x-\alpha)}) + x - \alpha\right] - \Big|_{-\infty}^{\infty}\left[\frac{1}{\lambda}\ln(1+e^{-\lambda(x-\alpha)}) + x - b\right]$$

$$= \lim_{x\to-\infty}\frac{1}{\lambda}\ln(1+e^{-\lambda(x-b)}) - \lim_{x\to-\infty}\frac{1}{\lambda}\ln(1+e^{-\lambda(x-\alpha)})$$

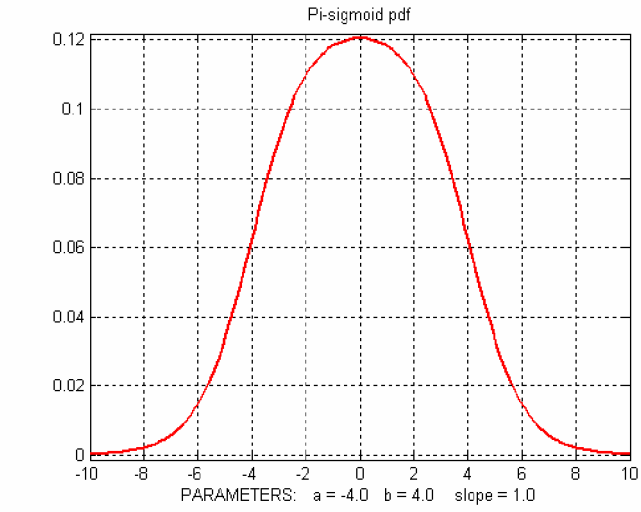
$$= b - a \qquad (2.12)$$

Αποδείχτηκε ότι η σταθερά κανονικοποίησης είναι η $K_{\sigma} = 1/(b-a)$. Μια πολύ σημαντική ιδιότητα που προκύπτει από τον παραπάνω υπολογισμό είναι ότι η σταθερά κανονικοποίησης K_{σ} βρέθηκε να είναι ανεξάρτητη από την παράμετρο λ που είναι η κλίση των δύο σιγμοειδών στα αντίστοιχα κέντρα τους. Αυτό απλουστεύει σημαντικά τόσο την διατύπωση της ίδιας της κατανομής όσο και τον υπολογισμό των εκφράσεων των παραγώγων της, όπως θα δούμε σε επόμενο κεφάλαιο. ■

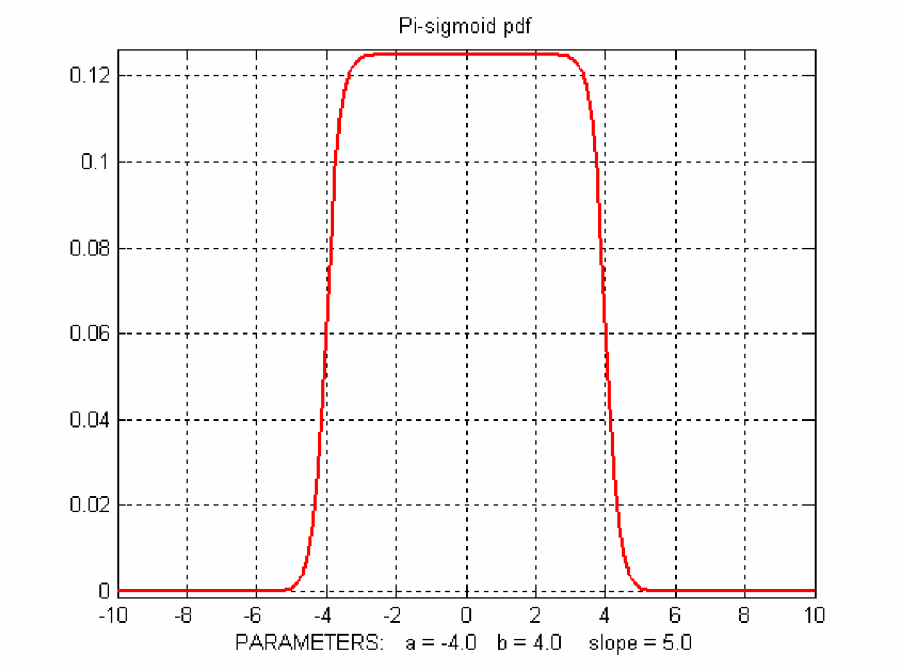
Ακολουθούν γραφικές παραστάσεις της Π-sigmoid για διάφορες τιμές των παραμέτρων της.



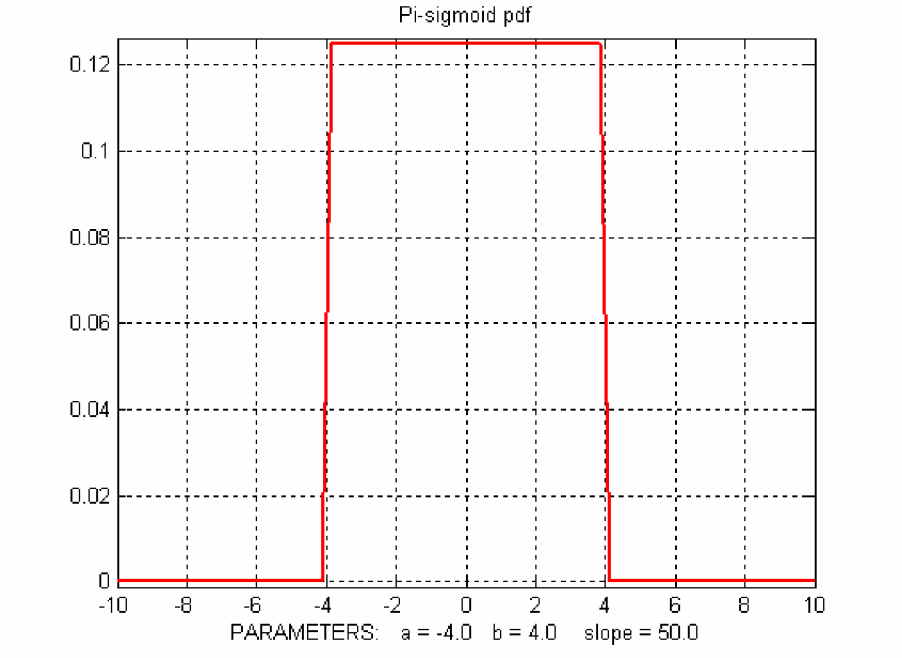
Σχήμα 2.5 Παράθεση δύο διαφορετικών σιγμοειδών συναρτήσεων με κέντρα -4 και 4 και κλίση 1.5 . Είναι σχεδόν προφανές ότι το αποτέλεσμα της διαφοράς τους θα είναι μια καμπανοειδής συνάρτηση ενώ για λ>>1 θα τείνει σε σχήμα Π.



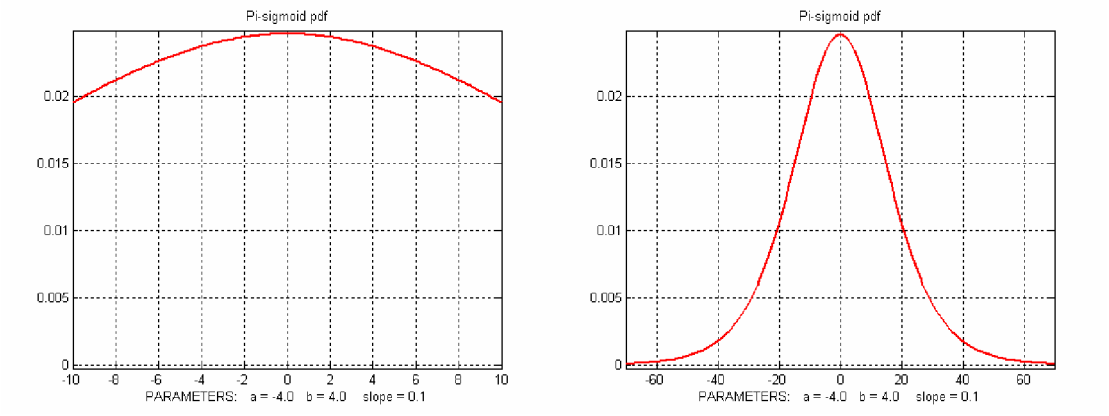
Σχήμα 2.6 Η κατανομή Π-sigmoid με παραμέτρους a=-4, b=4 και κλίση λ=1. Να σημειωθεί ότι το σχήμα αυτής της κατανομής ανταποκρίνεται στη διαφορά των σιγμοειδών συναρτήσεων του σχήματος 5.



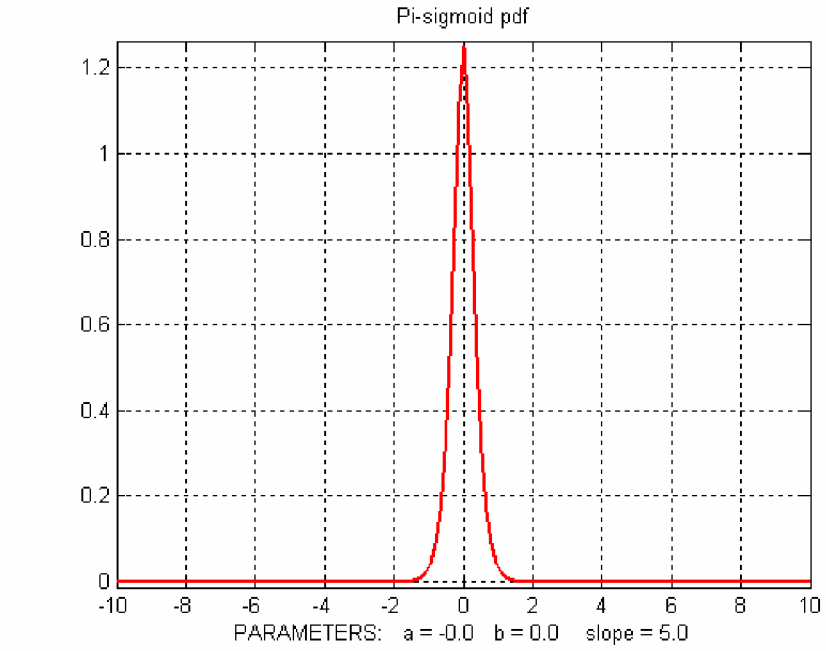
Σχήμα 2.7 Η κατανομή Π-sigmoid με παραμέτρους a=-4, b=4 και κλίση λ=5. Μεγαλώνοντας την τιμή του λ , η κατανομή αρχίζει να παίρνει εμφανώς το σχήμα Π.



Σχήμα 2.8 Η κατανομή Π-sigmoid με παραμέτρους a=-4, b=4 και κλίση λ=50. Μεγαλώνοντας ακόμα πιο πολύ την τιμή του λ, η κατανομή προσεγγίζει με πολύ ικανοποιητικό τρόπο την ομοιόμορφη.



Σχήμα 2.9 Η κατανομή Π-sigmoid με παραμέτρους a=-4, b=4 και κλίση λ=0.1. Βλέπουμε στην αριστερή γραφική παράσταση η συνάρτηση πλατειάζει σημαντικά και για αυτό ανοίγουμε το διάστημα των τιμών από το [-10 10] στο [-70 70] για να είναι ορατό το σχήμα της.



Σχήμα 2.10 Η κατανομή Π-sigmoid με παραμέτρους a=-0.001, b=0.001 και κλίση λ=5. Παρατηρείστε ότι η συνάρτηση γίνεται sharp, όταν η σχετική απόσταση του a με το b γίνει μικρή.

2.2.1. Ιδιότητες της Π-sigmoid

Ύστερα από τον μαθηματικό ορισμό αλλά και την γραφική περιγραφή της Π-sigmoid κατανομής, είναι καιρός να αναφέρουμε ορισμένα συμπεράσματα και ιδιότητες που προκύπτουν από την μελέτη της συμπεριφοράς της.

Η σημασία της παραμέτρου λ

Όπως έγινε κατανοητό και από τις γραφικές παραστάσεις, ο ρόλος της παραμέτρου λ είναι πολύ σημαντικός στην γενικότερη συμπεριφορά της κατανομής. Παρατηρούμε ότι για μεγάλες τιμές του, η συνάρτηση τείνει να γίνει ομοιόμορφη, ιδιότητα που την καθιστά σχεδόν μοναδική, αφού επιτυγχάνει το τελευταίο διατηρώντας την ιδιότητα της συνεχούς συνάρτησης. Από την άλλη, για μικρές σχετικά τιμές της παραμέτρου λ (δηλ. λ=1+ε, όπου ε μικρός θετικός αριθμός), η συνάρτηση παίρνει μια καμπανοειδή μορφή πλησιάζοντας αρκετά το σχήμα της κανονικής κατανομής, ενώ για λ<1 η πλατειάζει σημαντικά, δίνοντας το δικαίωμα της περιγραφής δεδομένων με μεγάλη διασπορά. Τέλος, θα πρέπει να αναφερθεί η σημαντική ιδιότητα της συμμετρίας που έχει η κατανομή που οφείλεται στην χρήση κοινής κλίσης λ για τις δύο σιγμοειδείς που την συνιστούν.

Η σημασία των παραμέτρων a και b

Οι δύο αυτές παράμετροι της κατανομής, παίζουν ένα παρεμφερή ρόλο με την παράμετρο της διασποράς που υπάρχει στην κανονική κατανομή. Παρατηρούμε ότι όσο πιο μεγάλη είναι η απόλυτη τιμή της διαφοράς τους |b-a|, τόσο πιο μεγάλη είναι και η διασπορά των δεδομένων. Αυτό γίνεται αρκετά εμφανές, όταν η παράμετρος λ είναι πολύ μεγάλη (λ>>1). Σε αυτή την περίπτωση, όπως φαίνεται και στο Σχήμα 2.8, η κατανομή ανέρχεται και κατέρχεται στα σημεία a και b αντίστοιχα, με σχεδόν κατακόρυφο τρόπο, δείχνοντας σχεδόν απόλυτα ότι η διασπορά των δεδομένων είναι όμοια με αυτήν της ομοιόμορφης κατανομής, δηλαδή, $(b-a)^2/12$. Θα πρέπει να σημειωθεί ότι στην έννοια της διασποράς εμπλέκεται και η παράμετρος λ, η οποία συνεισφέρει με έναν αντιστρόφως ανάλογο τρόπο σε αυτή.

Ακολουθεί η περιγραφή της αθροιστικής συνάρτησης, καθώς επίσης και κάποιων βασικών στατιστικών μεγεθών.

Αθροιστική συνάρτηση (CDF)

$$F(u;a,b,\lambda) = \frac{1}{b-a}\int_{-\infty}^{u}\left[\frac{1}{1+e^{-\lambda(x-a)}} - \frac{1}{1+e^{-\lambda(x-b)}}\right]dx$$
$$= \frac{\ln(e^{-\lambda(u-a)}+1) - \ln(e^{-\lambda(u-b)}+1) + \lambda(b-a)}{\lambda(b-a)}$$
$$= \frac{\ln(e^{-\lambda(u-a)}+1) - \ln(e^{-\lambda(u-b)}+1)}{\lambda(b-a)} + 1 \qquad (2.13)$$

Η αναμενόμενη τιμή ή κέντρο, X ~ Πs(a,b,λ)

$$E[X] = \frac{1}{b-a}\int_{-\infty}^{\infty}\left(\frac{1}{1+e^{-\lambda(x-a)}} - \frac{1}{1+e^{-\lambda(x-b)}}\right)xdx$$
$$= \frac{b+a}{2} \qquad (2.14)$$

Median

Επειδή η Π-sigmoid είναι συμμετρική συνάρτηση, το median ταυτίζεται με την αναμενόμενη τιμή , άρα ισχύει:

$$\text{Median(X)} = \frac{b+a}{2} \qquad (2.15)$$

2.3. Η πολυδιάστατη Π-sigmoid κατανομή

Σε αυτή την ενότητα θα παρουσιάσουμε την Π-sigmoid τροποποιημένη, έτσι ώστε να μπορεί να περιγράψει πολυδιάστατα δεδομένα. Καταρχάς, πριν προχωρήσουμε στην διατύπωση της, θα πρέπει να κάνουμε μια βασική παραδοχή αναφορικά με κάποια βασική ιδιότητα που θα πρέπει να ικανοποιούν τα δεδομένα που πρόκειται να μας απασχολήσουν. Θεωρούμε ότι τα χαρακτηριστικά των πολυδιάστατων δεδομένων είναι ανεξάρτητα, δηλαδή δεν συμμεταβάλλονται. Πιο συγκεκριμένα, εάν έχουμε ένα

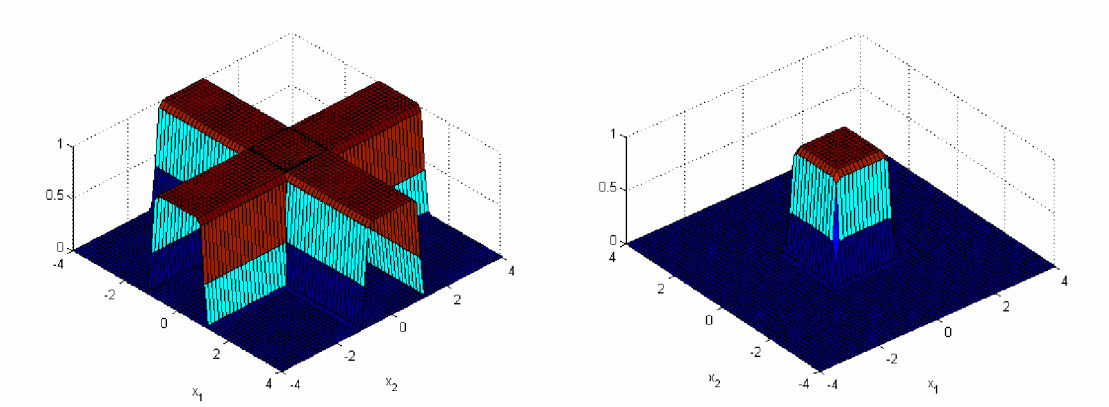
διάνυσμα x={x_1, x_2,…,x_D} τότε σύμφωνα με την ιδιότητα της υπό συνθήκη ανεξαρτησίας των χαρακτηριστικών, παίρνουμε την παρακάτω έκφραση:

$$p(x;\Theta)=\prod_{i=1}^{D} p(x_i;\Theta) \tag{2.16}$$

όπου Θ είναι οι παράμετροι της κατανομής. Αντικαταστώντας στην σχέση (2.16) τον μονοδιάστατο τύπο της Π-sigmoid (2.7) προκύπτει:

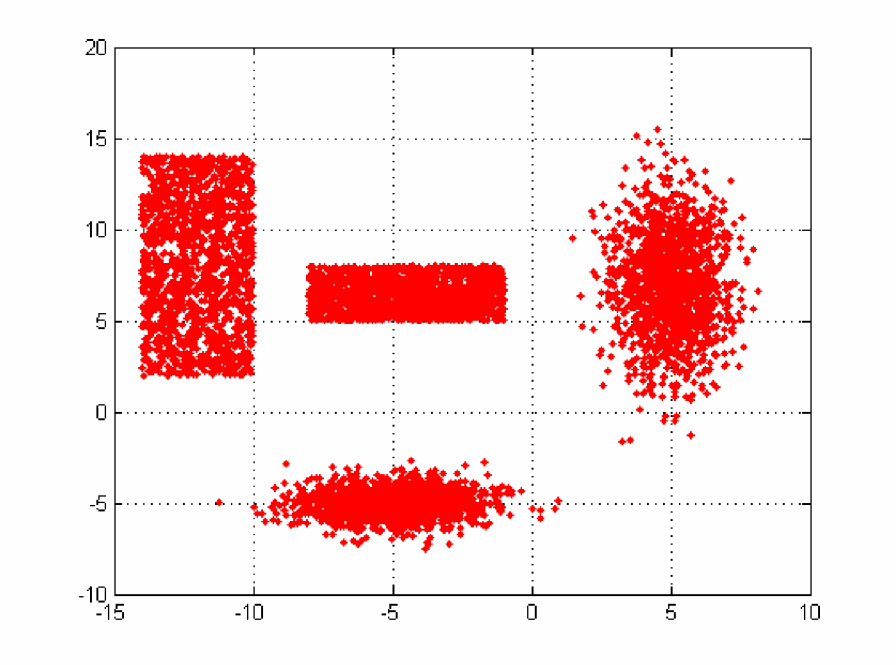
$$p(x\,|\,\Theta)=\prod_{d=1}^{D}\frac{\dfrac{1}{1+e^{-\lambda_d(x_d-a_d)}}-\dfrac{1}{1+e^{-\lambda_d(x_d-b_d)}}}{b_d-a_d} \tag{2.17}$$

Η τελευταία σχέση αποτελεί την πολυδιάστατη έκφραση της κατανομής Π-sigmoid, η οποία, όπως παρατηρούμε, διατυπώνεται πρακτικά ως η “τομή” των επιμέρους μονοδιάστατων εκδοχών της, όπως αυτές διατυπώνονται και στο γινόμενο της σχέσης (2.17). Ένα χαρακτηριστικό δισδιάστατο παράδειγμα, που αποτυπώνει την φιλοσοφία της παραπάνω σκέψης, φαίνεται και στο Σχήμα 2.11.

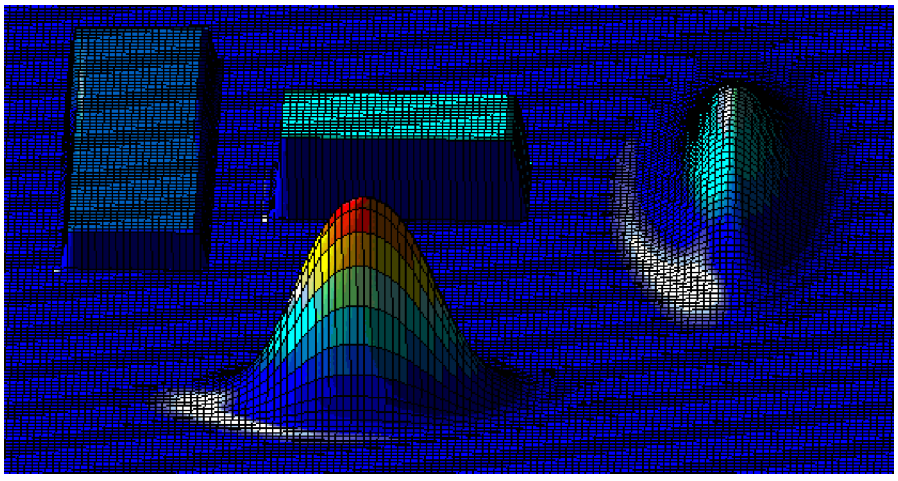


Σχήμα 2.11 Στο αριστερό σχήμα βλέπουμε την “διασταύρωση” των δύο μονοδιάστατων εκδοχών της Π-sigmoid, που εκπροσωπούν την κάθε μια από τις δύο διαστάσεις, και δεξιά βλέπουμε το αποτέλεσμα του γινομένου τους.

Παρακάτω ακολουθούν κάποιες γραφικές παραστάσεις για να γίνει κατανοητή η συμπεριφορά της πολυδιάστατης Π-sigmoid, έτσι όπως διατυπώθηκε στην σχέση (2.17).



Σχήμα 2.12 Ένα δείγμα από 4 ομάδων, τα δεδομένα των οποίων είναι ανεξάρτητα. Παρατηρούμε ότι οι άξονες συμμετρίας των ομάδων είναι παράλληλοι με τους κύριους άξονες. Τα δεδομένα των ορθογωνίων ομάδων είναι ομοιόμορφα, ενώ των άλλων δύο, γκαουσιανά.



Σχήμα 2.13 Η μορφή των γραφικών παραστάσεων των πολυδιάστατων Π-sigmoid κατανομών αντιστοιχούν σε κάθε ομάδα του Σχήματος 2.12.

Από τον τύπο (2.17) παρατηρούμε ότι για κάθε μια διάσταση έχουμε ξεχωριστές παραμέτρους a_d, b_d, λ_d, d=1,…,D. Πιο συγκεκριμένα, αν υποθέσουμε ότι τα δεδομένα είναι διάστασης D, χρειαζόμαστε συνολικά 3*D παραμέτρους προς εκτίμηση.

2.3.1. Περιστροφή στην πολυδιάστατη Π-sigmoid

Στην προηγούμενη ενότητα παρουσιάσαμε την Π-sigmoid ως πολυδιάστατη κατανομή. Το βασικό μειονέκτημα της εν λόγω εκδοχής είναι ότι για να δουλέψει ικανοποιητικά προϋποθέτει ανεξαρτησία χαρακτηριστικών. Τι γίνεται όμως όταν τα δεδομένα δεν πληρούν την παραπάνω υπόθεση; Απάντηση σε αυτό το ερώτημα έρχεται να δώσει η εισαγωγή της έννοιας της περιστροφής. Παρακάτω θα μελετήσουμε τον τρόπο με τον οποίο η Π-sigmoid, με κάποιες τροποποιήσεις, θα μπορέσει να περιγράψει δεδομένα με εξαρτημένα χαρακτηριστικά.

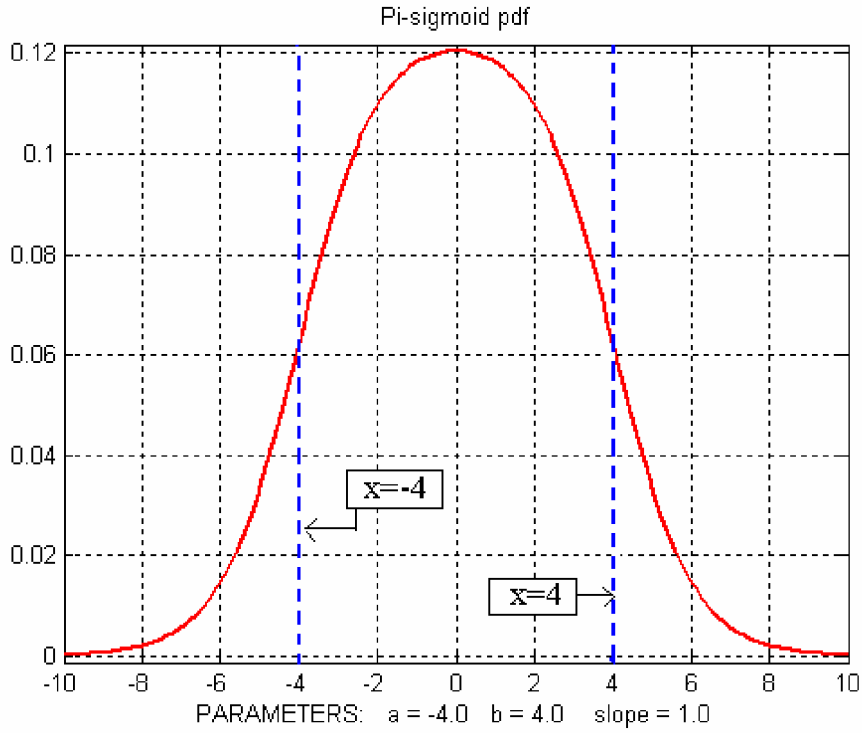
Σε αυτό το σημείο πρέπει να ανακαλέσουμε τον ορισμό της μονοδιάστατης Π-sigmoid κατανομής (2.7):

$$f(x) = \left(\frac{1}{b-a}\right)\left[\frac{1}{1+e^{-\lambda(x-a)}} - \frac{1}{1+e^{-\lambda(x-b)}}\right] \tag{2.18}$$

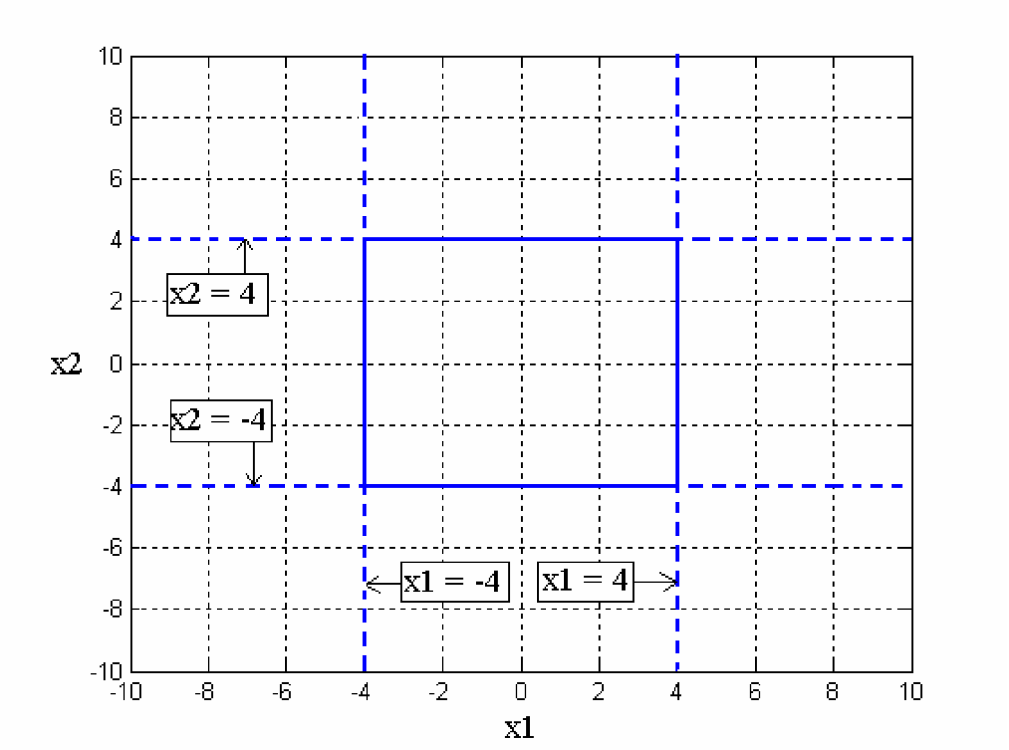
Παρατηρούμε ότι στην εκθετική συνάρτηση που βρίσκεται στον παρανομαστή των δύο σιγμοειδών, και πιο συγκεκριμένα στο εκθέτη της, υπάρχει μια συνάρτηση της μορφής (x-μ) η οποία επηρεάζει τόσο το πρόσημο όσο και την απόλυτη τιμή του εκθέτη. Για τιμές του x < μ το πρόσημο του εκθέτη [-λ(χ-μ), λ>0] γίνεται θετικό και η τιμή της σιγμοειδούς τείνει στο μηδέν όταν |x-μ|>>0, ενώ όταν x > μ, ο εκθέτης γίνεται αρνητικός και συνεπώς η τιμή της σιγμοειδούς τείνει στη μονάδα όταν |x-μ|>>0. Όπως φαίνεται και στο Σχήμα 14, οι ποσότητες (x-a) και (x-b) που υπάρχουν στον ορισμό (2.7) ορίζουν δύο κατακόρυφες ευθείες στα σημεία x=a και x=b αντίστοιχα. Οι παραπάνω παρατήρηση θα μπορούσε να αναχθεί και στις D διαστάσεις, όπου στον άξονα κάθε διάστασης ορίζεται ένα ζεύγος ευθειών, κάθετων σε αυτόν στα σημεία $x_d=a_d$ και $x_d=b_d$, d=1,…,D αντίστοιχα. Από το Σχήμα 2.15 γίνεται αντιληπτό ότι η τομή που ορίζεται από την διασταύρωση των “ζωνών” που

δημιουργούν τα ζεύγη αυτών των ευθειών ανά διάσταση, είναι γενικά στις D διαστάσεις ένα υπέρ-ορθογώνιο που έχει τις πλευρές του παράλληλες προς τους κύριους άξονες.

Το ερώτημα που τίθεται σε αυτό το σημείο είναι το αν θα μπορούσαμε να τροποποιήσουμε την Π-sigmoid έτσι ώστε να μπορεί να περιγράφει περιστραμένα υπέρ-ορθογώνια. Τη λύση στο πρόβλημα αυτό την δίνει η χρήση κεκλιμένων υπερ-επιπέδων.



Σχήμα 2.14 Γραφική απεικόνιση της σχέσης που έχουν οι ποσότητες x-a και x-b στον τρόπο με τον οποίο συμπεριφέρεται η Π-sigmoid.



Σχήμα 2.15 Δισδιάστατο παράδειγμα που απεικονίζει την ιδιότητα που προσδίδουν οι ποσότητες x_d-a_d και x_d-b_d, d=1,2 στην κατανομή Π-sigmoid.

Αντί λοιπόν, να έχουμε ζεύγη παράλληλων υπέρ-επιπέδων που να είναι κάθετα στον άξονα της εκάστοτε διάστασης, τοποθετούμε ζεύγη παράλληλων αλλά κεκλιμένων υπέρ-επιπέδων της μορφής:

$$\begin{aligned} w^T x - a &= w_1 x_1 + w_2 x_2 + \ldots + w_D x_D - a \\ w^T x - b &= w_1 x_1 + w_2 x_2 + \ldots + w_D x_D - b \end{aligned} \qquad (2.19)$$

όπου $w=[w_1\ w_2\ \ldots\ w_D]^T$, $w_d \in$ ¡ , d=1,..,D είναι ένα διάνυσμα βαρών, τα a, b ∈ ¡ είναι οι σταθεροί όροι και $x=[x_1\ x_2\ \ldots\ x_D]^T$ $x_d \in$ ¡ , d=1,..,D είναι το διάνυσμα εισόδου. Παρατηρούμε ότι το διάνυσμα βαρών w είναι ίδιο και στις δύο εξισώσεις (2.19) και αυτό μας εξασφαλίζει την παραλληλία των δύο υπέρ-επιπέδων. Αν τώρα αντικαταστήσουμε στην έκφραση της πολυδιάστατης Π-sigmoid τις ποσότητες (x_d - α_d) και (x_d - b_d) με υπέρ-επίπεδα της μορφής (2.19), τότε προκύπτει:

$$P(x;\Theta)=\prod_{d=1}^{D}\frac{\frac{1}{1+e^{-\lambda_d\left(\left(\sum_{i=1}^{D}W_{di}x_i\right)-a_d\right)}}-\frac{1}{1+e^{-\lambda_d\left(\left(\sum_{i=1}^{D}W_{di}x_i\right)-b_d\right)}}}{\frac{|b_d-a_d|}{\|W_d\|}} \tag{2.20}$$

όπου $W_d \in$ ¡ ^{D} ,$W_d=\{W_{d1},\ldots, W_{dD}\}$. Όπως παρατηρούμε από την προηγούμενη σχέση, οι σταθεροί όροι των δύο υπέρ-επιπέδων είναι οι α_d και b_d αντίστοιχα. Τα διανύσματα W_d για κάθε διάσταση d=1,…,D θα μπορούσαμε συγκεντρωτικά να πούμε ότι απαρτίζουν τις στήλες ενός τετράγωνο πίνακα W διάστασης DxD ο οποίος ορίζεται ως:

$$W = [W_1\ W_2\ \ldots\ W_D] \tag{2.21}$$

Στη σχέση (2.20) έχει αλλάξει επίσης και η σταθερά κανονικοποίησης K_σ, η οποία για κάθε μια διάσταση d=1,…,D, είναι το αντίστροφο της απόστασης των δύο παράλληλων υπέρ-επιπέδων της μορφής (2.19). Δηλαδή, συναθροιστικά για όλες τις διαστάσεις ισχύει:

$$K_\sigma=\frac{1}{\prod_{d=1}^{D}\frac{(b_d-a_d)}{\|W_d\|}} \tag{2.22}$$

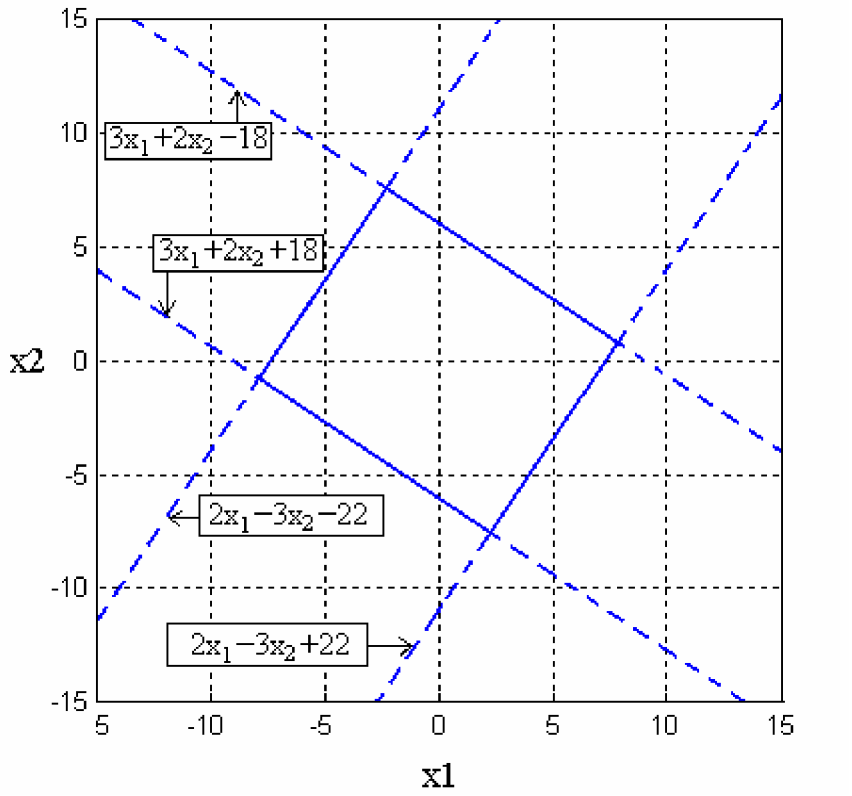
Θα πρέπει να επισημάνουμε ότι για να λειτουργεί σωστά η σταθερά κανονικοποίησης έτσι όπως αναδιατυπώνεται στην σχέση (2.22), θα πρέπει να απαιτήσουμε, ο πίνακας βαρών W να είναι ορθογώνιος. Δηλαδή θα πρέπει να ισχύει:

$$WW^T = I \tag{2.23}$$

Αυτό θα μας εξασφαλίσει ότι όλα τα ζεύγη των παράλληλων υπέρ-επιπέδων $W_d^T x-\alpha_d$ και $W_d^T x-b_d$, $\forall$ d, θα είναι ανά δύο κάθετα μεταξύ τους, σχηματίζοντας με την τομή τους, στις D διαστάσεις, περιστραμμένα υπέρ-ορθογώνια. Ένα χαρακτηριστικό παράδειγμα φαίνεται στο Σχήμα 2.16.

Σχολιάζοντας αυτό το γεγονός της περιστροφής, θα λέγαμε με αντίστοιχο τρόπο, όπως και στην μη περιστραμμένη εκδοχή, ότι για τα δεδομένα που εμπίπτουν εντός των ορίων του περιστραμμένου υπερ-ορθογωνίου, η Π-sigmoid κατανομή θα έχει την τάση να δίνει μεγάλες τιμές, ενώ για τα υπόλοιπα η τιμή της θα μειώνεται αισθητά, προσεγγίζοντας το μηδέν. Για να πραγματοποιηθεί το τελευταίο θα πρέπει ή να απομακρυνθούμε πολύ τα όρια του υπερ-ορθογωνίου ή να ισχύει $\lambda >> 1$.

Συνοψίζοντας, θα λέγαμε ότι η έννοια της περιστροφής είναι ένα πολύ σημαντικό στοιχείο το οποίο προάγει την ικανότητα του μοντελοποίησης της πολυδιάστατης κατανομής. Δίνει μια πιο ευέλικτη μορφή και βοηθά στην καλύτερη περιγραφή των δεδομένων, τα οποία είναι τυχαία τοποθετημένα στο χώρο των D διαστάσεων.



Σχήμα 2.16 Δισδιάστατο παράδειγμα, στο οποίο φαίνεται η κάθετη τομή από δύο ζεύγη παράλληλων και κεκλιμένων ευθειών(=2D υπερ-επιπέδων). Το αποτέλεσμα είναι ένα περιστραμμένο ορθογώνιο.

ΚΕΦΑΛΑΙΟ 3. ΕΚΠΑΙΔΕΥΣΗ ΜΙΚΤΩΝ ΜΟΝΤΕΛΩΝ Π-SIGMOID ΚΑΤΑΝΟΜΩΝ

3.1. Γενικά

Στο προηγούμενο κεφάλαιο μελετήσαμε διεξοδικά την κατανομή Π-sigmoid παρουσιάζοντας αναλυτικά της ιδιότητες και την συμπεριφορά της. Σε αυτό το κεφάλαιο θα ανακαλέσουμε τον ορισμό ενός μικτού μοντέλου κατανομών, έτσι όπως ορίστηκε στην σχέση (1.11), και θα το ανάγουμε σε μικτό μοντέλο Π-sigmoid κατανομών. Συγκεκριμένα, θα περιγράψουμε τον τρόπο με τον οποίο μπορεί να εκπαιδευτεί ένα τέτοιο μοντέλο μέσω του αλγορίθμου βελτιστοποίησης GEM. Η χρήση του τελευταίου συνίσταται στο γεγονός ότι ο τρόπος διαπραγμάτευσης του M-βήματος (Maximization step), δηλαδή η βελτιστοποίηση της πλήρους πιθανοφάνειας, γίνεται όπως θα δείξουμε, με χρήση αριθμητικής τεχνικής, σε αντίθεση με ένα GMM (Gaussian Mixture Model) όπου πραγματοποιείται αναλυτικά. Παράλληλα, θα γίνει αναφορά στον τρόπο αρχικοποίησης του αλγορίθμου, αλλά και στον τρόπο αντιμετώπισης του θορύβου που υπάρχει στα σύνολα δεδομένων. Θα πρέπει να σημειωθεί ότι ειδικά για την περίπτωση του μικτού μοντέλου “περιστραμμένων” Π-sigmoid κατανομών, έχει υιοθετηθεί ειδική τεχνική καθορισμού των παραμέτρων που συνιστούν τον πίνακα περιστροφής W (2.21). Αναλυτικότερα, η τεχνική αυτή υλοποιείται εμβόλιμα εντός της επαναληπτικής διαδικασίας του GEM, και

ταυτόχρονα έξω από το Μ-βήμα, στην οποία υπάρχει η εξασφάλιση ότι μετά το πέρας της συνεισφοράς της, θα πραγματοποιηθεί μείωση της πιθανοφάνειας.

3.2. Μέγιστη Πιθανοφάνεια

Πριν προχωρήσουμε στην διατύπωση ενός μικτού μοντέλου Π-sigmoid κατανομών κρίνεται απαραίτητη η παρουσίαση της εκτίμησης Μέγιστης Πιθανοφάνειας, των παραμέτρων μιας Π-sigmoid κατανομής. Όπως έχει ήδη αναφερθεί και αντιλαμβανόμαστε και από το όνομα της, ο στόχος αυτής της διαδικασίας είναι η βελτιστοποίηση της συνάρτησης της Πιθανοφάνειας, έτσι όπως αυτή ορίστηκε στην σχέση (1.16). Η βελτιστοποίηση γίνεται είτε με αναλυτικό τρόπο, παίρνοντας δηλαδή την παράγωγο της συνάρτησης και εξισώνοντάς την με το μηδέν, είτε με αριθμητικό. Μετά το πέρας αυτής της διαδικασίας, οι παράμετροι που προκύπτουν ως βέλτιστη λύση, οδηγούν στην καλύτερη δυνατή περιγραφή του συνόλου δεδομένων από την υπό εξέταση κατανομή. Για να εφαρμόσουμε αυτήν μέθοδο και στην Π-sigmoid κατανομή θα πρέπει να ορίσουμε την συνάρτηση της Πιθανοφάνειας. Έτσι αν έχουμε ένα σύνολο δεδομένων $X=(x_1, x_2,\ldots,x_N)^T$, $x_i \in ¡^D$, i=1,…,N, στο οποίο οι παρατηρήσεις $\{x_i\}$ υποθέτουμε ότι παράχθηκαν ανεξάρτητα από μια πολυδιάστατη Π-sigmoid κατανομή Πs(x;θ), τότε σύμφωνα με τη σχέση (1.16) έχουμε:

$$L(\mathrm{X}\,|\,\theta)=\prod_{i=1}^{N}\Pi s(x_i\,|\,\theta)=\prod_{i=1}^{N}\prod_{d=1}^{D}\frac{\dfrac{1}{1+e^{-\lambda_d(W_d^T x_i-a_d)}}-\dfrac{1}{1+e^{-\lambda_d(W_d^T x_i-b_d)}}}{b_d-a_d} \tag{3.1}$$

όπου λ, α, b∈ $¡^D$ και W∈ $¡^{D,D}$. Η παραπάνω σχέση είναι ένα γινόμενο και είναι προφανές ότι είναι αρκετά δύσκολο βρεθεί αναλυτικά η παράγωγος του, γι’ αυτό καταφεύγουμε στην λύση της λογαριθμικής Πιθανοφάνειας που διατυπώνεται ως εξής:

$$LL(\mathrm{X}|\theta)=\sum_{i-1}^{N}\ln(\Pi s(x_i\,|\theta))=\sum_{i=1}^{N}\sum_{d=1}^{D}\ln\frac{\dfrac{1}{1+e^{-\lambda_d(W_d^T x_i-a_d)}}-\dfrac{1}{1+e^{-\lambda_d(W_d^T x_i-b_d)}}}{b_d-a_d} \tag{3.2}$$

Όπως φαίνεται και από την σχέση (3.2), με την χρήση του λογαρίθμου, το γινόμενο έγινε άθροισμα και έτσι η παραγώγιση αυτής της έκφρασης γίνεται σαφώς ευκολότερη. Για να βελτιστοποιήσουμε τώρα την παραπάνω ποσότητα, με βάσει πάντα την θεωρία Βελτιστοποίησης, θα πρέπει να υπολογίσουμε τις μερικές παραγώγους ως προς κάθε παράμετρο και να τις εξισώσουμε με το μηδέν. Η επίλυση αυτών των εξισώσεων θα παράγει τις βέλτιστες τιμές των παραμέτρων, οι οποίες θα οδηγήσουν στην μεγιστοποίηση της συνάρτησης της πιθανοφάνειας. Θα πρέπει σε αυτό το σημείο να τονίσουμε, ότι οι παράμετροι λ, α, b είναι διανύσματα που ανήκουν στον χώρο ¡ ^{D}, άρα οφείλουμε να αντιμετωπίσουμε την κάθε μια συνιστώσα τους, ως ξεχωριστή παράμετρο και να παραγωγίσουμε ως προς κάθε μία από αυτές. Ακολουθώντας, λοιπόν, την παραπάνω λογική, υπολογίζουμε τις μερικές παραγώγους ως ακολούθως:

Μερική παράγωγος ως προς την παράμετρο λ_d

$$
\begin{aligned}
&\frac{\partial LL(\mathrm{X}|\lambda,a,b)}{\partial\lambda_d} = \frac{\partial}{\partial\lambda_d}\sum_{i=1}^{N}\sum_{d=1}^{D}\ln\frac{\frac{1}{1+e^{-\lambda_d(W_d^T x_i - a_d)}} - \frac{1}{1+e^{-\lambda_d(W_d^T x_i - b_d)}}}{b_d - a_d} \\
&= \frac{\partial}{\partial\lambda_d}\left\{\sum_{i=1}^{N}\left[\ln\left(\frac{1}{1+e^{-\lambda_d(W_d^T x_i - a_d)}} - \frac{1}{1+e^{-\lambda_d(W_d^T x_i - b_d)}}\right)\right] - N\ln(b_d - a_d)\right\} \\
&\overset{(2.5)}{=} \sum_{i=1}^{N}\left[\frac{\sigma(W_d^T x_i;\lambda_d,a_d)(1-\sigma(W_d^T x_i;\lambda_d,a_d))(W_d^T x_i - a_d)}{\sigma(W_d^T x_i;\lambda_d,a_d) - \sigma(W_d^T x_i;\lambda_d,b_d)} - \right. \\
&\qquad \left. \frac{\sigma(W_d^T x_i;\lambda_d,b_d)(1-\sigma(W_d^T x_i;\lambda_d,b_d))(W_d^T x_i - b_d)}{\sigma(W_d^T x_i;\lambda_d,a_d) - \sigma(W_d^T x_i;\lambda_d,b_d)}\right]
\end{aligned}
\tag{3.3}
$$

Μερική παράγωγος ως προς την παράμετρο a_d

$$\frac{\partial LL(\mathrm{X};\lambda,a,b)}{\partial a_d} = \frac{\partial}{\partial a_d}\sum_{i=1}^{N}\sum_{d=1}^{D}\ln\frac{\frac{1}{1+e^{-\lambda_d(W_d^T x_i - a_d)}} - \frac{1}{1+e^{-\lambda_d(W_d^T x_i - b_d)}}}{b_d - a_d}$$

$$= \frac{\partial}{\partial a_d}\left\{\sum_{i=1}^{N}\left[\ln\left(\frac{1}{1+e^{-\lambda_d(W_d^T x_i - a_d)}} - \frac{1}{1+e^{-\lambda_d(W_d^T x_i - b_d)}}\right)\right] - N\ln(b_d - a_d)\right\}$$

$$\overset{(2.5)}{=} \sum_{i=1}^{N}\left[-\lambda_d\frac{\sigma(W_d^T x_i;\lambda_d,a_d)(1-\sigma(W_d^T x_i;\lambda_d,a_d))}{\sigma(W_d^T x_i;\lambda_d,a_d) - \sigma(W_d^T x_i;\lambda_d,b_d)}\right] + \frac{N}{b_d - a_d} \tag{3.4}$$

Μερική παράγωγος ως προς την παράμετρο b_d

$$\frac{\partial LL(\mathrm{X}|\lambda,a,b)}{\partial b_d} = \frac{\partial}{\partial b_d}\sum_{i=1}^{N}\sum_{d=1}^{D}\ln\frac{\frac{1}{1+e^{-\lambda_d(W_d^T x_i - a_d)}} - \frac{1}{1+e^{-\lambda_d(W_d^T x_i - b_d)}}}{b_d - a_d}$$

$$= \frac{\partial}{\partial b_d}\left\{\sum_{i=1}^{N}\left[\ln\left(\frac{1}{1+e^{-\lambda_d(W_d^T x_i - a_d)}} - \frac{1}{1+e^{-\lambda_d(W_d^T x_i - b_d)}}\right)\right] - N\ln(b_d - a_d)\right\}$$

$$\overset{(2.5)}{=} \sum_{i=1}^{N}\left[\lambda_d\frac{\sigma(W_d^T x_i;\lambda_d,b_d)(1-\sigma(W_d^T x_i;\lambda_d,b_d))}{\sigma(W_d^T x_i;\lambda_d,a_d) - \sigma(W_d^T x_i;\lambda_d,b_d)}\right] - \frac{N}{b_d - a_d} \tag{3.5}$$

Όπως παρατηρούμε από τις παραπάνω σχέσεις, η μορφή που έχουν οι μερικές παράγωγοι ως προς κάθε παράμετρο, είναι μάλλον αποτρεπτική στο να επιχειρήσουμε να τις εξισώσουμε με το μηδέν και να βρούμε αναλυτικά τη λύση που χρειαζόμαστε. Εγκαταλείπουμε, λοιπόν, τον αναλυτικό τρόπο βελτιστοποίησης και καταφεύγουμε σε αριθμητικές μεθόδους. Συγκεκριμένα, η μέθοδος που θα μας απασχολήσει είναι η BFGS η οποία ανήκει στην οικογένεια των Quasi-Newton μεθόδων. Πληροφορίες σχετικά με την μέθοδο μπορούν να αντληθούν πληροφορίες από το βιβλίο των *Nocedal & Wright* [7].

Η μέθοδος BFGS απαιτεί το gradient της αντικειμενικής συνάρτησης καθώς επίσης και κάποια αρχική τιμή για τις παραμέτρους, για να μπορέσει να ξεκινήσει τη διαδικασία βελτιστοποίησης. Στο πρόβλημα μας, το gradient της λογαριθμικής συνάρτησης της Πιθανοφάνειας, ισούται με το διάνυσμα των μερικών παραγώγων έτσι όπως αυτές υπολογίστηκαν στις προηγούμενες σχέσεις (3.3)-(3.5). Δηλαδή ισχύει:

$$\nabla f = \left[\frac{\partial LL(\mathrm{X}|\lambda, a, b)}{\partial \lambda} \quad \frac{\partial LL(\mathrm{X}|\lambda, a, b)}{\partial a} \quad \frac{\partial LL(\mathrm{X}|\lambda, a, b)}{\partial b} \right]^T \tag{3.6}$$

Είναι πολύ σημαντικό στη διαδικασία αυτή της αριθμητικής βελτιστοποίησης να λάβουμε υπόψιν μας τους περιορισμούς που θέτει η ίδια η κατανομή και να εξασφαλίσουμε ότι η λύση που θα παραχθεί θα τους ικανοποιεί. Δηλαδή μετά την ολοκλήρωση της βελτιστοποίησης, θα πρέπει να ισχύει:

$$[a_d]_{ML} < [b_d]_{ML}, \quad d = 1, \ldots, D \tag{3.7}$$

$$[\lambda_d]_{ML} > 0, \qquad d = 1, \ldots, D \tag{3.8}$$

Για να επιτύχουμε την ικανοποίηση των παραπάνω σχέσεων, πρέπει να καταφύγουμε στη λύση της βελτιστοποίησης με περιορισμούς. Στην τελευταία, οι περιορισμοί θα πρέπει πλέον να εμπλακούν στην όλη διαδικασία της βελτιστοποίησης, όπου σε κάθε απόπειρα βελτίωσης των παραμέτρων θα πρέπει αυτοί να ικανοποιούνται.

Όπως γίνεται αντιληπτό, αυτό θα προσέθετε ένα αρκετά μεγάλο κόστος στην χρονική πολυπλοκότητα του αλγορίθμου, κάνοντας την σύγκλιση και την παραγωγή της τελικής λύσης αργή και δυσχερή. Για να αντιμετωπίσουμε αυτό το πρόβλημα καταφεύγουμε στο μετασχηματισμό της μορφής τη Π-sigmoid κατανομής, έτσι ώστε από τον ορισμό και μόνο να ικανοποιούνται οι δεδομένοι περιορισμοί. Οι αλλαγές που πρέπει να πραγματοποιηθούν είναι απλές και συνοψίζονται στις εξής δύο αντικαταστάσεις:

$$\lambda_d = g_d^2 \tag{3.9}$$

$$b_d = a_d + r_d^2 \tag{3.10}$$

Όπως φαίνεται και από την σχέση (3.9), για να εξασφαλίσουμε ότι η κλίση λ_d δεν θα πάρει ποτέ αρνητική τιμή, αντικαταστάθηκε από την ποσότητα g_d^2, $g_d \in$ ¡ η οποία είναι πάντα θετική. Επίσης, έγινε αντικατάσταση της παραμέτρου b_d με την παράσταση $\alpha_d + r_d^2$, $r_d \in$ ¡ , για την οποία ισχύει προφανώς η ανισότητα $b_d = \alpha_d + r_d^2 \geq$

α_d. Με αυτές τις δύο τροποποιήσεις ικανοποιούνται οι περιορισμοί (3.7) και (3.8), πράγμα που οδηγεί στην αποφυγή της βελτιστοποίησης με περιορισμούς, απλουστεύοντας έτσι την όλη διαδικασία. Οι παραπάνω μετασχηματισμοί της Π-sigmoid κατανομής, συνοψίζονται στην παρακάτω φόρμουλα:

$$\Pi s(x;g,a,r,W)=\prod_{d=1}^{D}\frac{\frac{1}{1+e^{-g_d^2(W_d^T x_i-a_d)}}-\frac{1}{1+e^{-g_d^2(W_d^T x_i-a_d-r_d^2)}}}{r_d^2} \tag{3.11}$$

Ένα πολύ σημαντικό ζήτημα που προκύπτει μέσα από αυτή τη νέα διατύπωση, είναι και η αλλαγή της μορφής τόσο της λογαριθμικής πιθανοφάνειας όσο και των εκφράσεων των μερικών παραγώγων της. Τροποποιώντας, λοιπόν, τις σχέσεις (3.3)-(3.5) με βάση την νέα μορφή της Π-sigmoid κατανομής, προκύπτουν οι παρακάτω νέες σχέσεις:

Η νέα μορφή της λογαριθμικής Πιθανοφάνειας

$$LL(\mathrm{X};\theta)=\sum_{i=1}^{N}\sum_{d=1}^{D}\ln\frac{\frac{1}{1+e^{-g_d^2(W_d^T x_i-a_d)}}-\frac{1}{1+e^{-g_d^2(W_d^T x_i-a_d-r_d^2)}}}{r_d^2}$$

$$=\sum_{i=1}^{N}\sum_{d=1}^{D}\left[\ln\left[\frac{1}{1+e^{-g_d^2(W_d^T x_i-a_d)}}-\frac{1}{1+e^{-g_d^2(W_d^T x_i-a_d-r_d^2)}}\right]-2\ln r_d\right] \tag{3.12}$$

Παρακάτω φαίνονται και οι μερικές παράγωγοι ως προς τις νέες παραμέτρους g_d, α_d, r_d.

Παρατήρηση 1

Να σημειωθεί ότι για λόγους συντομίας θα συμβολίζουμε με $\sigma(\alpha_d)$ την σιγμοειδή συνάρτηση με κέντρο α_d, κλίση g_d^2 και βάρη W_d. Κάνουμε αυτή την παραδοχή επειδή το κέντρο είναι αυτό που διακρίνει τις σιγμοειδείς συναρτήσεις μεταξύ τους. Έτσι, οι υπόλοιπες παράμετροι που είναι κοινές, θα υπονοούνται . Δηλαδή θα ισχύει:

$$\sigma(a_d)=\frac{1}{1+e^{-g_d^2(W_d^T x-a_d)}} \quad \text{και} \quad \sigma(a_d+r_d^2)=\frac{1}{1+e^{-g_d^2(W_d^T x-a_d-r_d^2)}}$$

Μερική παράγωγος ως προς την παράμετρο g_d

$$\frac{\partial LL(\mathrm{X};g,a,r)}{\partial g_d}=\sum_{i=1}^{N}\left[\frac{2\sigma(a_d)(1-\sigma(a_d))(W_d^T x_i-a_d)}{\sigma(a_d)-\sigma(a_d+r_d^2)}-\right.$$
$$\left.-\frac{2\sigma(a_d+r_d^2)(1-\sigma(a_d-r_d^2))(W_d^T x_i-a_d-r_d^2)}{\sigma(a_d)-\sigma(a_d+r_d^2)}\right] \tag{3.13}$$

Μερική παράγωγος ως προς την παράμετρο α_d

$$\frac{\partial LL(\mathrm{X}|g,a,r)}{\partial a_d}=\sum_{i=1}^{N}\frac{-\sigma(a_d)(1-\sigma(a_d))g_d^2+\sigma(a_d+r_d^2)(1-\sigma(a_d+r_d^2))g_d^2}{\sigma(a_d)-\sigma(a_d+r_d^2)} \tag{3.14}$$

Μερική παράγωγος ως προς την παράμετρο r_d

$$\frac{\partial LL(\mathrm{X}|g,a,r)}{\partial r_d}=\sum_{i=1}^{N}\left[\frac{\sigma(a_d+r_d^2)(1-\sigma(a_d-r_d^2))g_d^2 2r_d}{\sigma(a_d)-\sigma(a_d+r_d^2)}-\frac{2}{r_d}\right] \tag{3.15}$$

Τοποθετώντας τώρα τις τροποποιημένες μερικές παραγώγους των σχέσεων (3.13)-(3.15) στον ορισμό της σχέσης (3.6) παίρνουμε το νέο gradient της λογαριθμικής πιθανοφάνειας.

Αυτό που εκκρεμεί για να ολοκληρώσουμε την περιγραφή της διαδικασίας εκτίμησης μέγιστης πιθανοφάνειας (MLE), είναι η αρχικοποίηση των παραμέτρων της λογαριθμικής πιθανοφάνειας. Είναι προφανές, ότι όσο πιο κοντά βρίσκονται στην βέλτιστη λύση, τόσο πιο γρήγορα και αποδοτικά θα εξελιχθεί η όλη διαδικασία. Η διαπραγμάτευση αυτού του θέματος επιδέχεται ποικίλες αντιμετωπίσεις, εντούτοις όμως, εδώ θα περιγράψουμε μια από αυτές, που θεωρείται αξιόπιστη και απλή.

Η γενική ιδέα είναι ότι αρχικά εκτελούμε την διαδικασία MLE για την κανονική κατανομή (βλέπε παρ. 1.3.1) με στόχο την ανίχνευση των βέλτιστων παραμέτρων (κέντρο μ_{ML}, πίνακας συμμεταβλητότητας Σ_{ML}) για το σύνολο δεδομένων που εξετάζουμε. Με δεδομένες τώρα τις παραμέτρους μ_{ML}, Σ_{ML} που προκύπτουν αναλυτικά, θα προσπαθήσουμε να προσδιορίσουμε κάποιους συσχετισμούς με τις παραμέτρους της Π-sigmoid κατανομής έτσι ώστε οι τελευταίες να είναι όσο δυνατόν πιο κοντά στην βέλτιστη λύση. Οι συσχετισμοί των

παραμέτρων προκύπτουν από τις γεωμετρικές ιδιότητες του κέντρου μ_{ML} και των χαρακτηριστικών μεγεθών του πίνακα Σ_{ML}. Πιο συγκεκριμένα, θα θέλαμε τα υπερεπίπεδα $W_d^T x-\alpha_d$ και $W_d^T x-\alpha_d-r_d^2$, να είναι παράλληλα με το d-οστό ιδιοδιάνυσμα u_d του πίνακα Σ_{ML}. Ταυτόχρονα, πρέπει να είναι τοποθετημένα εκατέρωθεν του κέντρου μ_{ML} και να επέχουν από αυτό απόσταση ίση με $0.5*L_d^{1/2}$ (όπου L_d είναι η d-οστή ιδιοδιοτιμή του πίνακα Σ_{ML}) Σχήμα 3.2. Όσον αφορά την παράμετρο g_d, βάζουμε αυθαίρετα κάποια αρχική τιμή που συνήθως επιδιώκουμε να είναι σχετικά μικρή, όπως θα γίνει κατανοητό σε επόμενη παράγραφο.

Έχοντας λοιπόν περιγράψει την λογική της διαδικασίας αρχικοποίησης, θα συνοψίσουμε στην συνέχεια την διαδικασία αυτή. Έστω μ_{ML} και Σ_{ML} που οι βέλτιστες ML παράμετροι της αντίστοιχης κανονικής κατανομής και U ο πίνακας που έχει ως στήλες του τα ιδιοδιανύσματα του Σ_{ML} και L_d η d-οστή ιδιοτιμή του.

Αρχικοποίηση του πίνακα βαρών W

$$W:=U \tag{3.16}$$

Αρχικοποίηση παραμέτρου gd (*όπως έχει αναφερθεί η τιμή 2 είναι σχεδόν αυθαίρετη)*

$$g_d:=2\ ,\quad d=1,..,D \tag{3.17}$$

Αρχικοποίηση της παραμέτρου a_d *(όπου W_d είναι η d-οστή στήλη του πίνακα W)*

$$a_d = W_d^T \mu_{ML} - \|W_d\|_2 \sqrt{L_d} \tag{3.18}$$

Αρχικοποίηση της παραμέτρου r_d *(όπου W_d είναι η d-οστή στήλη του πίνακα W)*

$$r_d = \sqrt{2\|W_d\|_2 \sqrt{L_d}} \tag{3.19}$$

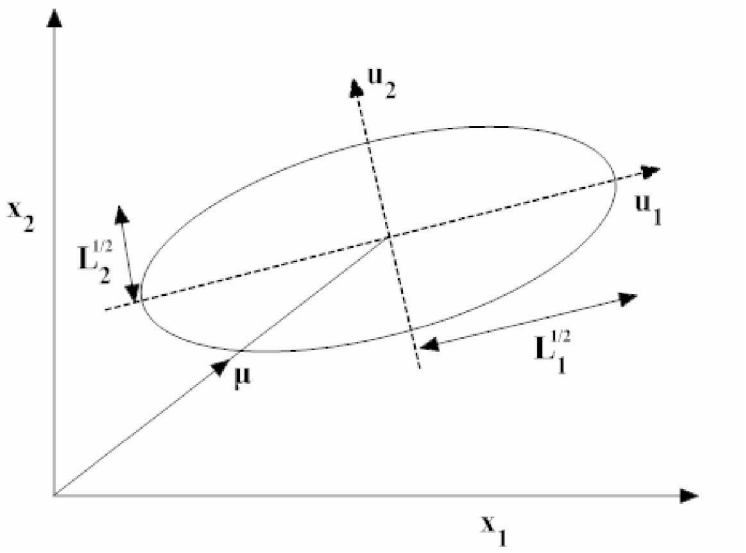
Οι δύο τελευταίες αναθέσεις προκύπτουν από τον πολύ γνωστό τύπο υπολογισμού απόστασης μεταξύ δύο υπέρ-επιπέδων $\varepsilon_1=w^Tx-a_1$ και $\varepsilon_2=w^Tx-a_2$ που φαίνεται παρακάτω:

$$d(\varepsilon_1,\varepsilon_2) = \frac{|a_2 - a_1|}{\|w\|_2} \tag{3.20}$$

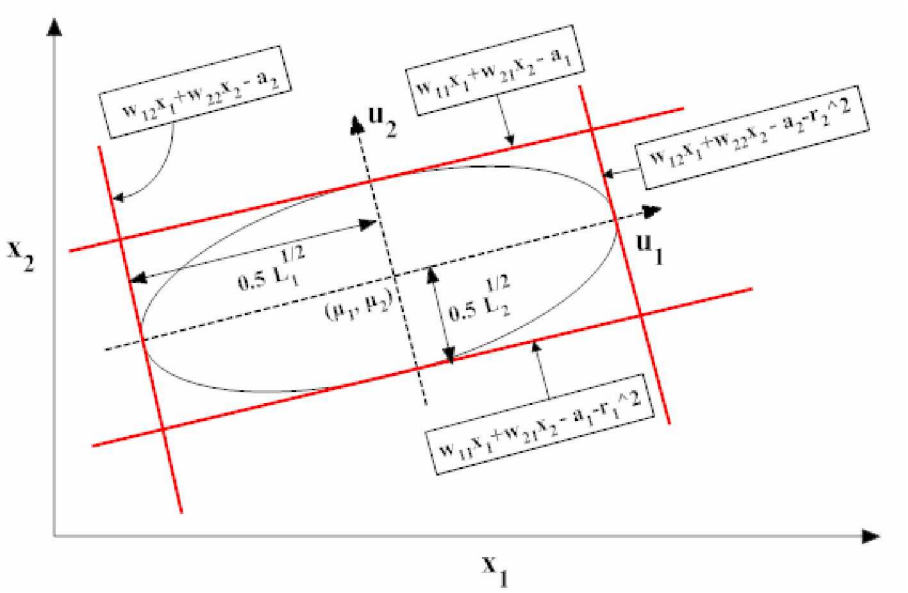
Πιο συγκεκριμένα η λογική του προσδιορισμού των παραμέτρων της d-οστής διάστασης, στηρίζονται στον υπολογισμό του υπερπεπιπέδου που είναι παράλληλο με το d-οστό διάνυσμα u_d του πίνακα συμμεταβλητότητας Σ_{ML} και ταυτόχρονα περνάει από το κέντρο μ_{ML}. Αν συμβολίσουμε με ε_μ αυτό το υπερεπίπεδο, τότε αυτό δίνεται από την παρακάτω σχέση:

$$\varepsilon_\mu = W_d^T x - W_d^T \mu_{ML} \tag{3.21}$$

Για να υπολογίσω τις παραμέτρους α_d, b_d και W_d, που υπάρχουν στον ορισμό (2.20) της πολυδιάστατης Π-sigmoid, αρκεί να ανιχνεύσω εκείνα τα υπερεπίπεδα που είναι παράλληλα με το ε_μ και βρίσκονται εκατέρωθεν αυτού, σε απόσταση $\sqrt{L_d}$ το καθένα. Εκμεταλλευόμενοι την σχέση (3.20), προκύπτουν οι σχέσεις (3.18) και (3.19)



Σχήμα 3.1 Ένα περίγραμμα της gaussian κατανομής στις δύο διαστάσεις η οποία χαρακτηρίζεται από το κέντρο μ και πίνακα συμμεταβλητότητας του οποίου τα ιδιοδιανύσματα είναι u_1 και u_2, με αντίστοιχες ιδιοτιμές L_1 και L_2.



Σχήμα 3.2 Ο τρόπος αρχικοποίησης των παραμέτρων της Π-sigmoid και η γραφική αναπαράσταση του συσχετισμού τους με την βέλτιστη λύση που προκύπτει από την μέθοδο MLE για την κανονική κατανομή.

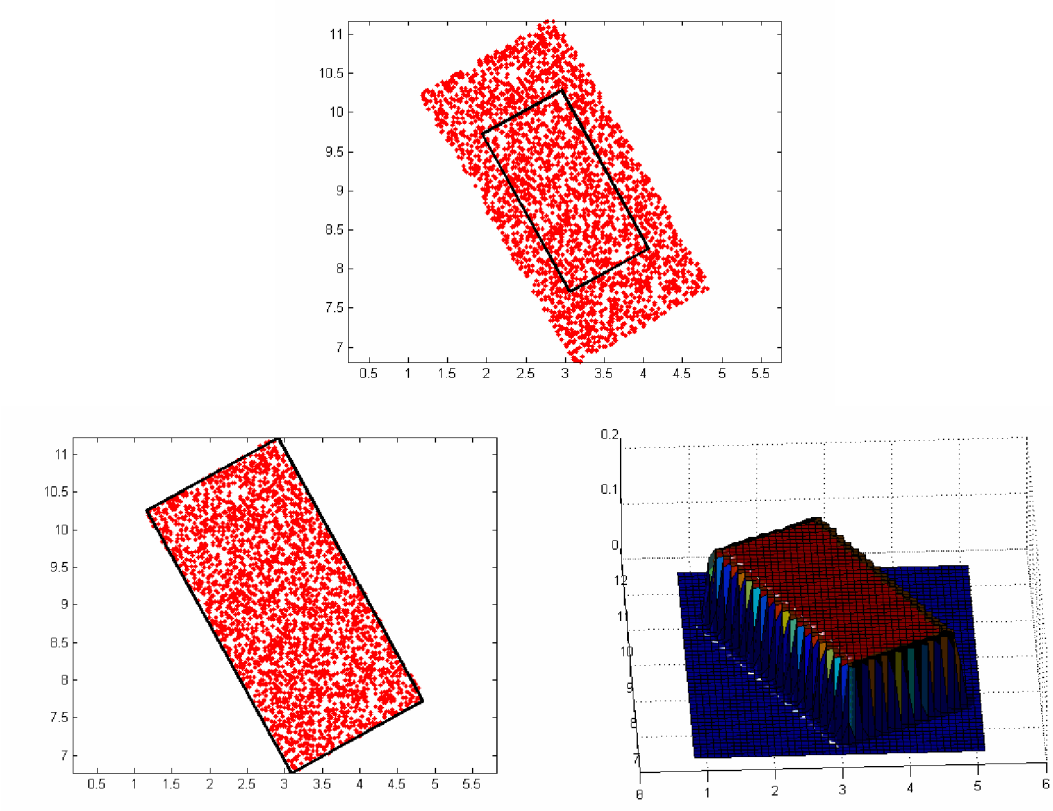
Όπως φαίνεται και στο Σχήμα 3.2, η λογική αρχικοποίησης που περιγράφηκε παραπάνω δημιουργεί στις D διαστάσεις ένα περιγεγραμμένο υπερ-ορθογώνιο ή ταυτόσημα, η υπέρ-έλλειψη που παράγεται από την κανονική κατανομή γίνεται εγγεγραμμένη του υπερ-ορθογωνίου που δημιουργείται από τα υπέρ-επίπεδα της μορφής (2.19) που υπάρχουν στον ορισμό της Π-sigmoid.

Αφού τώρα έχουμε ολοκληρώσει την παρουσίαση του τρόπου υπολογισμού του gradient της λογαριθμικής συνάρτησης πιθανοφάνειας αλλά και του τρόπου αρχικοποίησης των παραμέτρων, μπορούμε πλέον να εφαρμόσουμε την διαδικασία βελτιστοποίησης με τη μέθοδο BFGS. Να σημειωθεί ότι όταν αναφέρουμε τον όρο βελτιστοποίηση στην διαδικασία MLE, ουσιαστικά υπονοούμε μεγιστοποίηση της πιθανοφάνειας (ή ισοδύναμα της λογαριθμικής πιθανοφάνειας). Εντούτοις, αρκετές φορές και για τεχνικούς κυρίως λόγους μας βολεύει το να ελαχιστοποιήσουμε μια συνάρτηση.

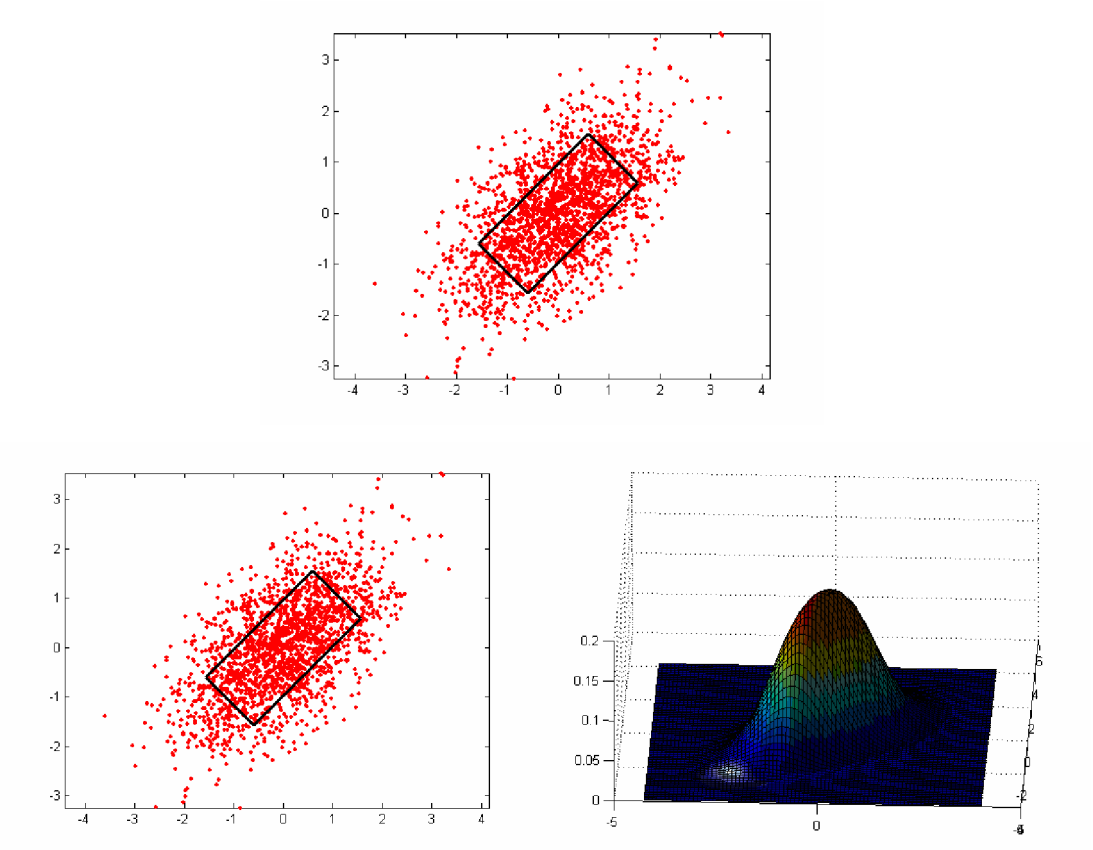
Θεωρώντας λοιπόν την αρνητική λογαριθμική Πιθανοφάνεια με την βοήθεια της σχέσης (3.12) αλλά και το αρνητικό του gradient αυτής, ανάγουμε το πρόβλημα μας,

σε πρόβλημα ελαχιστοποίησης, θεωρώντας πλέον την συνάρτηση -LL(Θ|X) ως αντικειμενική συνάρτηση την οποία πρέπει να ελαχιστοποιήσουμε.

Παρακάτω παρουσιάζουμε το αποτέλεσμα της εφαρμογής της μεθόδου MLE στην κατανομή Π-sigmoid που την διαδικασία που περιγράφηκε παραπάνω. Η μέθοδος εφαρμόζεται σε δύο σύνολα δεδομένων με ορθογώνια ομοιόμορφα και γκαουσιανά δεδομένα αντίστοιχα. Πειραματικά αποτελέσματα σχετικά με την ποιότητα της λύσης περιγράφονται στο επόμενο κεφάλαιο.



Σχήμα 3.3 Εφαρμογή της μεθόδου MLE σε μια ομοιόμορφη ορθογώνια ομάδα. Πάνω βλέπουμε το contour της αρχικής λύσης και κάτω το contour της τελικής λύσης και το τρισδιάστατο plot της κατανομής.



Σχήμα 3.4 Εφαρμογή της μεθόδου MLE σε μια γκαουσιανή ομάδα. Πάνω βλέπουμε το contour της αρχικής λύσης και κάτω το contour της τελικής λύσης και το τρισδιάστατο plot της κατανομής.

3.3. Μικτά μοντέλα Π-sigmoid κατανομών (ΠsMM)

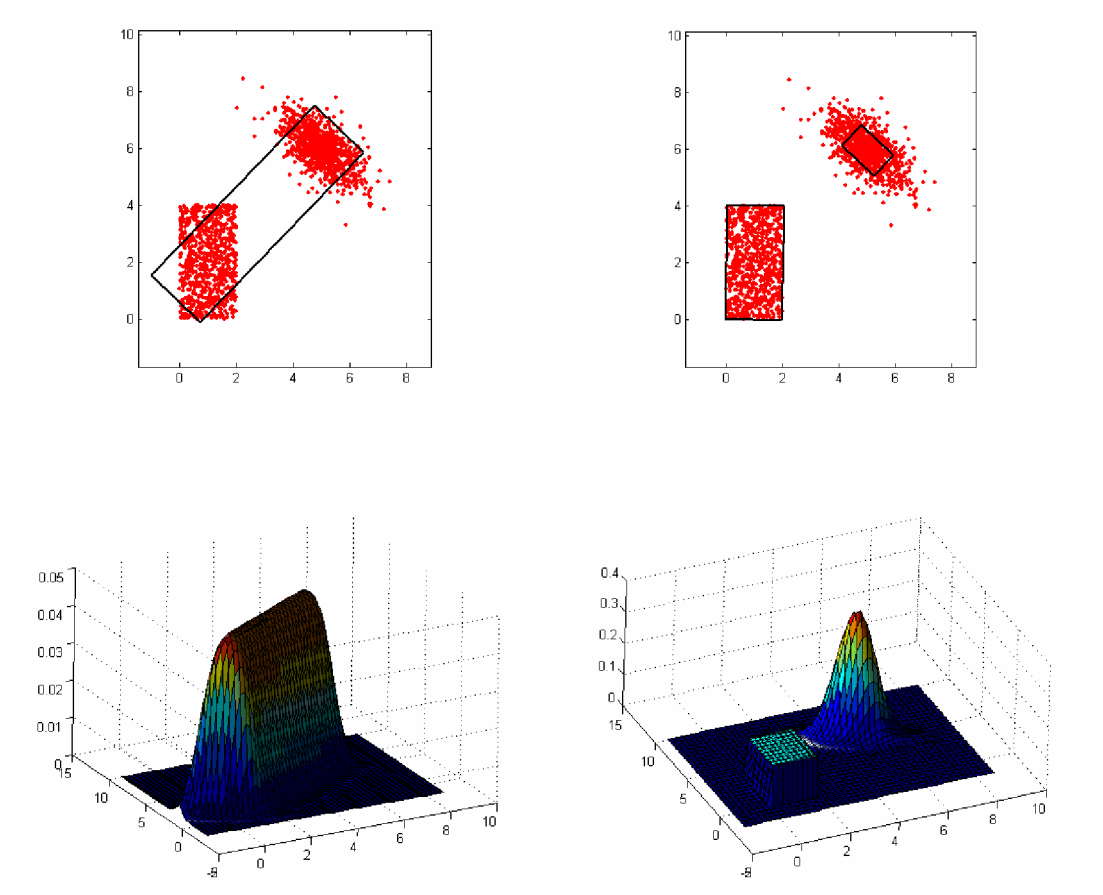
Παρόλο που η Π-sigmoid κατανομή έχει αρκετά μεγάλη ευελιξία στην περιγραφή πολλών τύπων δεδομένων, εντούτοις, όπως και στην περίπτωση της gaussian κατανομής, έχουμε κάποια σημαντικά προβλήματα στην περιγραφή πραγματικών δεδομένων. Όπως φαίνεται και στο Σχήμα 3.5(α), υπάρχει ένα πρόβλημα όταν τα δεδομένα σχηματίζουν περισσότερα του ενός “clusters”. Παρατηρήστε ότι η Π-sigmoid αδυνατεί να ανταποκριθεί ικανοποιητικά, και η λύση που παράγει χαρακτηρίζεται ως μη επαρκής και ανακριβής. Γι’ αυτό το λόγο καταφεύγουμε στην λύση ενός μικτού μοντέλου. Ανακαλούμε τη σχέση (1.11) για να διατυπώσουμε πλέον, την μορφή που θα έχει ένα ΠsMM (Π-sigmoid Mixture Model) το οποίο είναι ένας γραμμικός συνδυασμός πεπερασμένων Π-sigmoid κατανομών, σταθμισμένων με βάρη. Το μοντέλο για K κατανομές θα είναι της μορφής:

$$p(x) = \sum_{k=1}^{K} \pi_k \Pi s(x; g_k, a_k, r_k, W_k) \tag{3.22}$$

Η κάθε μια Π-sigmoid κατανομή $\Pi s(x; g_k, a_k, r_k, W_k)$ θα αποκαλείται συνιστώσα του μικτού μοντέλου και θα έχει τις δικές της ξεχωριστές παραμέτρους g_k, α_k, r_k και W_k. Οι παράμετροι π_k της σχέσης (3.22), αποκαλούνται συντελεστές μίξης και πρέπει να ισχύει:

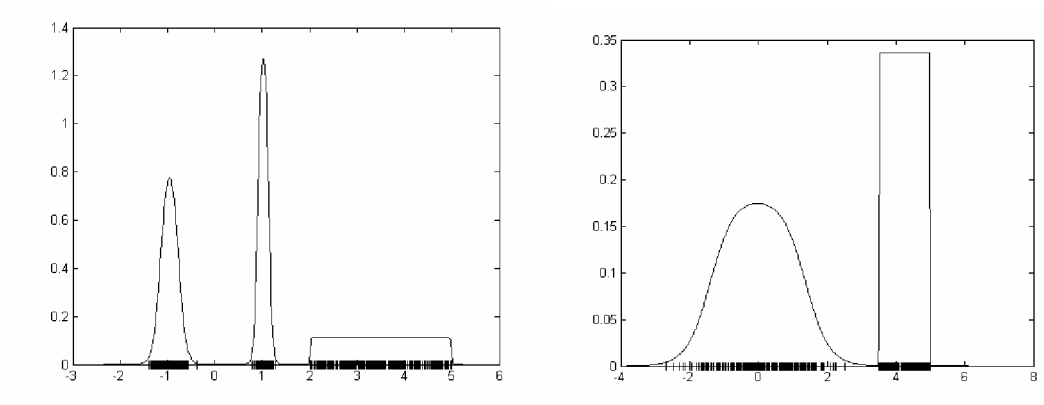
$$0 \le \pi_k \le 1 \tag{3.23}$$

$$\sum_{k=1}^{K} \pi_k = 1 \tag{3.24}$$



(α) (β)

Σχήμα 3.5 Αριστερά βλέπουμε την Π-sigmoid κατανομή να προσπαθεί να περιγράψει ανεπιτυχώς, 2 cluster δεδομένων, ενώ δεξιά βλέπουμε ένα μικτό μοντέλο με 2 Π-sigmoid κατανομές με σαφώς καλύτερη απόδοση.



Σχήμα 3.6 Παραδείγματα μικτών μοντέλων Π-sigmoid κατανομών σε μονοδιάστατα τεχνητά δεδομένα. Ο τρόπος εκπαίδευσης αναφέρεται σε επόμενη παράγραφο.

Συνοψίζοντας, λοιπόν θα λέγαμε ότι ένα μικτό μοντέλο Π-digmoid κατανομών χαρακτηρίζεται από τις παραμέτρους π, g, α, r, W για τα οποία έχουμε χρησιμοποιήσει για να συμβολίσουμε τις παραμέτρους των επιμέρους συνιστωσών, δηλαδή ισχύει:

$$\pi \text{ ø}\{\pi_1, \ldots, \pi_K\}$$

$$g \text{ ø}\{g_1, \ldots, g_K\}$$

$$\alpha \text{ ø}\{\alpha_1, \ldots, \alpha_K\}$$

$$r \text{ ø}\{r_1, \ldots, r_K\}$$

$$W\text{ø}\{W_1, \ldots, W_K\}$$

Το ερώτημα που τίθεται είναι, το πώς μπορούμε να προσδιορίσουμε τις παραπάνω παραμέτρους έτσι ώστε να προσαρμόσουμε με τον βέλτιστο δυνατό τρόπο το μοντέλο μας στα δεδομένα που έχουμε. Μια πρώτη απάντηση θα ήταν με την βοήθεια της μεθόδου Μέγιστης Πιθανοφάνειας που περιγράφηκε σε προηγούμενη παράγραφο. Χρησιμοποιώντας την σχέση (3.22), η λογαριθμική συνάρτηση της πιθανοφάνειας δίνεται από την σχέση:

$$\ln p(X \mid \pi, g, a, r, W) = \sum_{i=1}^{N} \ln \left\{ \sum_{k=1}^{K} \pi_\kappa \Pi s(x_n \mid g_k, a_k, r_k, W_k) \right\} \tag{3.25}$$

όπου $X=\{x_1, \ldots, x_N\}$. Το πρώτο πράγμα που μας κάνει εντύπωση είναι ότι η σχέση (3.25) είναι πιο πολύπλοκη από την λογαριθμική πιθανοφάνεια μιας μοναδικής Π-sigmoid κατανομής λόγω και του αθροίσματος που υπάρχει στο λογάριθμο. Έτσι αποκλείουμε το ενδεχόμενο να βρούμε αναλυτικά το μέγιστο της συνάρτησης αυτής. Θα καταφύγουμε σε άλλες μεθόδους βελτιστοποίησης που και συγκεκριμένα τον αλγόριθμο GEM.

3.4. Εκπαίδευση ενός ΠsMM μέσω του GEM

Σε αυτή τη ενότητα θα παρουσιάσουμε τον τρόπο με τον οποίο θα μπορέσουμε να προσδιορίσουμε κάποιες καλές (ιδανικά τις βέλτιστες) εκτιμήσεις για τις παραμέτρους ενός μικτού μοντέλου Π-sigmoid κατανομών της μορφής (3.22). Όπως έχει ήδη αναφερθεί, η επίτευξη του τελευταίου πραγματοποιείται μέσα από τη βελτιστοποίηση της συνάρτησης της πιθανοφάνειας, ή ισοδύναμα της λογαριθμικής πιθανοφάνειας έτσι όπως διατυπώθηκε και στην σχέση (3.25). Ο αλγόριθμος βελτιστοποίησης που επιλέχτηκε, είναι ο GEM (generalized expectation maximization) ο οποίος έχει παρουσιαστεί στο $1^{ο}$ κεφάλαιο. Ο λόγος που μας οδήγησε στη επιλογή αυτής της μεθόδου είναι ότι ο GEM εκμεταλλεύεται την δομή του μοντέλου με απώτερο στόχο το “σπάσιμο” της διαδικασίας βελτιστοποίησης σε περισσότερα και ευδιαχείριστα κομμάτια. Να σημειωθεί ότι η τελευταία αυτή σημαντική ιδιότητα θα μας δώσει το δικαίωμα να εκπαιδεύσουμε σταδιακά και τους πίνακες περιστροφής W_k οι οποίοι, όπως θα δούμε παρακάτω, δεν εμπλέκονται στην διαδικασία βελτιστοποίησης που πραγματοποιείται στο Μ-βήμα, αλλά πραγματοποιείται έξω από αυτό, με χρήση ειδικής τεχνικής.

Ο αλγόριθμος GEM, ως μια υποπερίπτωση του αλγορίθμου ΕΜ, έχει παρουσιαστεί αναλυτικά στο πρώτο κεφάλαιο. Σε αυτή την παράγραφο, θα επικεντρώσουμε την παρουσίαση μας στον τρόπο με τον οποίο μπορούμε να υλοποιήσουμε τον συγκεκριμένο αλγόριθμο για την εκπαίδευση ενός ΠsMM. Θεωρούμε ότι ο αριθμός Κ των συνιστωσών είναι δεδομένος εξαρχής. Τα σημεία στα οποία θα εστιάσουμε την προσοχή μας και αφορούν την ειδικά την μορφή ενός ΠsMM είναι τρία και είναι τα εξής:

- Αρχικοποίηση του GEM
- Ε-βήμα
- Περιγραφή του τρόπου βελτιστοποίησης στο Μ-βήμα
- Τρόπος καθορισμού των πινάκων περιστροφής W_k

3.5. Αρχικοποίηση GEM

Όπως έχουμε ξαναπεί η τελική λύση που θα παραχθεί από τον αλγόριθμο ΕΜ (αντίστοιχα GEM) εξαρτάται πάρα πολύ από τις αρχικές τιμές που θα δώσουμε. Είναι πολύ σημαντικό λοιπόν, να χρησιμοποιήσουμε έναν αποδοτικό τρόπο αρχικοποίησης των παραμέτρων για να εξασφαλίσουμε ένα καλής ποιότητας αποτέλεσμα. Η τεχνική που έχουμε χρησιμοποιήσει είναι σε ένα μεγάλο βαθμό παρεμφερής με αυτή που πραγματοποιήθηκε και στην μέθοδο MLE. Η γενική ιδέα είναι ότι εκπαιδεύουμε πρώτα ένα GMM με ακριβώς τον ίδιο αριθμό συνιστωσών που επιθυμούμε (ή/και έχουμε ανιχνεύσει) να έχει το αντίστοιχο ΠsΜΜ. Στην συνέχεια, προσπαθούμε να βρούμε συσχετισμούς μεταξύ των παραμέτρων της κάθε συνιστώσας κατανομής του GMM με αυτές του ΠsΜΜ. Ένα εύλογο ερώτημα που προκύπτει από την παραπάνω διαδικασία είναι το πως μπορούμε να εκπαιδεύσουμε αποδοτικά ένα GMM. Αν και δεν υπάρχει αντικειμενική και σαφής απάντηση σε αυτό το ερώτημα, εντούτοις, ένας καλός και αποδοτικός τρόπος είναι ο αλγόριθμος Greedy EM στον οποίο έχουμε αναφερθεί και σε προηγούμενο κεφάλαιο. Με δεδομένη λοιπόν το εκπαιδευμένο GMM, οι συσχετισμοί που υπάρχουν μεταξύ των παραμέτρων των δύο μοντέλων φαίνονται παρακάτω:

Έστω Σ_k και μ_k οι παράμετροι της k-οστής συνιστώσας του GMM και g_π_k η αντίστοιχη prior πιθανότητα, που προέκυψαν μετά την εκπαίδευση, τότε οι παράμετροι W_k, g_k, α_k, r_k της k-οστής συνιστώσας του ΠsΜΜ προκύπτουν ως εξής:

Βρίσκουμε τον πίνακα U_k που έχει ως στήλες του τα ιδιοδιανύσματα του πίνακα συμμεταβλητότητας Σ_k και τον αναθέτουμε ως τιμή, στον πίνακα W_k.

$$W_k := U_k \tag{3.26}$$

Η λογική για να κάνουμε αυτό στηρίζεται στο γεγονός ότι η μεγαλύτερη διασπορά των δεδομένων υπάρχει κατά την ίδια κατεύθυνση που υπαγορεύουν τα ιδιοδιανύσματα του πίνακα συμμεταβλητότητας [9]. Έτσι αντιλαμβανόμαστε ότι τα τελευταία μας δίνουν μια πολύ καλή εκτίμηση για τους πίνακες περιστροφής W.

Για τις prior πιθανότητες π_k του ΠsMM κάνουμε απλά την ανάθεση

$$\pi_k := g\_\pi_k \tag{3.27}$$

Αρχικοποίηση της παραμέτρου g_{kd}

$$g_{kd} := 2 , \quad d=1,..,D \tag{3.28}$$

Όπως έχει αναφερθεί επιδιώκουμε να βάζουμε μικρές τιμές στην παράμετρο της κλίσης g_{kd} αυτό γιατί στα πρώτα βήματα του GEM θέλουμε η κατανομή μας να είναι σχετικά “απλωμένη” για να υπάρχει σχετική ευχέρεια στην αλλαγή του σχήματος της, ανάλογα με τον τρόπο κατανομής των δεδομένων.

Αρχικοποίηση της παραμέτρου α_{kd}

$$a_{kd} = W_{kd}^T \mu_{ML} - \|W_{kd}\|_2 \sqrt{L_{kd}} \tag{3.29}$$

Αρχικοποίηση της παραμέτρου α_{kd}

$$r_{kd} = \sqrt{2 \|W_{kd}\|_2 \sqrt{L_{kd}}} \tag{3.30}$$

όπου W_{kd} είναι η d-οστή στήλη του πίνακα περιστροφής W_k (της k-οστής συνιστώσας).

Οι δύο τελευταίες αναθέσεις προκύπτουν, όπως έχουμε ήδη αναφέρει, προκύπτουν από την πολύ γνωστή σχέση υπολογισμού απόστασης μεταξύ δύο υπέρ-επιπέδων $\varepsilon_1=w^Tx-a_1$ και $\varepsilon_2=w^Tx-a_2$ που φαίνεται παρακάτω:

$$d(\varepsilon_1,\varepsilon_2)=\frac{|a_2-a_1|}{\|w\|_2} \tag{3.31}$$

3.5.1. E-Βήμα

Σε αυτό το βήμα του αλγορίθμου, όπως και στην περίπτωση της του μικτού μοντέλου κανονικών κατανομών, γίνεται υπολογισμός των posterior πιθανοτήτων $p(k|x_i)$ με τον τρόπο που φαίνεται παρακάτω:

$$p(k\mid x_i)=\frac{\pi_k \Pi s(x_i;a_k,r_k,g_k,W_k)}{\sum_{j=1}^{K}\pi_j \Pi s(x_i;a_j,r_j,g_j,W_j)} \tag{3.32}$$

3.5.2. Μ-βήμα: Περιγραφή του τρόπου βελτιστοποίησης

Αντικειμενικός στόχος αυτού του βήματος του αλγορίθμου είναι η μεγιστοποίηση της πλήρους πιθανοφάνειας, ως προς τις παραμέτρους του μικτού μοντέλου π, g, a, r.

$$\begin{aligned} LLc &= \sum_{i=1}^{N}\sum_{k=1}^{K} p(k\mid x_i)\ln[\pi_\kappa \Pi s(x_i\mid k)] \\ &= \sum_{i=1}^{N}\sum_{k=1}^{K} p(k\mid x_i)\ln \pi_\kappa + \sum_{i=1}^{N}\sum_{k=1}^{K} p(k\mid x_i)\ln \Pi s(x_i\mid k) \end{aligned} \tag{3.33}$$

Όπως παρατηρούμε, οι δύο τελευταίοι όροι της σχέσης (3.33), είναι ανεξάρτητοι μεταξύ τους αφού ο δεύτερος εμπλέκει την k-οστή συνιστώσα κατανομή του μοντέλου, ενώ ο πρώτος την prior πιθανότητα της τελευταίας. Αυτό σημαίνει ότι για να βρούμε τις βέλτιστες τιμές των παραμέτρων $\pi ø\{\pi_1,\ldots,\pi_K\}$ θα πρέπει να μεγιστοποιήσουμε τον πρώτο όρο, ενώ για τις παραμέτρους g, a, r αρκεί η μεγιστοποίηση του δευτέρου.

Εδικά για την περίπτωση των παραμέτρων $\pi \emptyset \{\pi_1,\ldots,\pi_K\}$, όπως έχει δειχτεί και στο πρώτο κεφάλαιο, η βελτιστοποίηση πραγματοποιείται αναλυτικά με την εξής κλειστή μορφή:

$$\pi_k = \frac{1}{N}\sum_{i=1}^{N} p(k \mid x_i), \forall k \tag{3.34}$$

όπου $p(k \mid x_i)$ είναι η posterior πιθανότητα που υπολογίστηκε στο E-βήμα. Αναφορικά τώρα με την βελτιστοποίηση των άλλων παραμέτρων, θα πρέπει μεγιστοποιήσουμε το δεύτερο όρο της σχέσης (3.33). Το τελευταίο, όπως είναι εύκολα αντιληπτό, είναι αδύνατο να πραγματοποιηθεί αναλυτικά, γι' αυτό καταφεύγουμε και πάλι στην λύση της αριθμητικής βελτιστοποίησης και συγκεκριμένα στην μέθοδο BFGS. Όπως και στην περίπτωση της μιας κατανομής(προηγούμενη παράγραφος), εγείρονται τα ίδια ζητήματα σχετικά με τις απαιτήσεις της μεθόδου που θα την βοηθήσουν να δουλέψει ικανοποιητικά (αρχικές εκτιμήσεις, gradient). Η λογική είναι ακριβώς αντίστοιχη. Οι αρχικές τιμές με τις οποίες τροφοδοτούμε την BFGS για να ξεκινήσει η διαδικασία βελτιστοποίησης, είναι οι παλιές εκτιμήσεις των παραμέτρων g, α, r που παράχθηκαν στο ακριβώς προηγούμενο βήμα του GEM, δηλαδή ισχύει:

$$\text{BFGS\_initial\_values: } \{ g^{initial}, \alpha^{initial}, r^{initial} \} = \{ g^{old}, \alpha^{old}, r^{old} \} \tag{3.35}$$

Να σημειωθεί ότι οι παλιές εκτιμήσεις των παραμέτρων $[g^{old}, \alpha^{old}, r^{old}]$ δεν αποτελούν πλέον το μέγιστο (προσέγγιση μεγίστου, ακριβέστερα) της πλήρους πιθανοφάνειας γιατί έχει μεσολαβήσει το E-βήμα, στο οποίο έχουν αλλάξει οι ποσότητες $p(k|x_i)$ που υπάρχουν στην σχέση (3.33). Άρα σίγουρα αποτελούν μια καλή αρχική εκτίμηση, η οποία μάλιστα καθώς θα προχωρούν τα βήματα του GEM θα γίνεται ολοένα και καλύτερη μιας και θα συγκλίνουν σε κάποιο τοπικό μέγιστο της πιθανοφάνειας.

Αναφορικά τώρα με το gradient της σχέσης (3.33), υπολογίζεται κάνοντας ακριβώς τις ίδιές παραδοχές με αυτές που έγιναν στην μέθοδο MLE. Δηλαδή θα χρησιμοποιήσουμε και εδώ την τροποποιημένη μορφή της Π-sigmoid κατανομής για την εξασφάλιση της ικανοποίησης των απαραίτητων περιορισμών.

Οι σχέσεις που θα μας δώσουν τελικά το gradient φαίνονται στην συνέχεια:

Μερική παράγωγος ως προς την παράμετρο g_{kd}

$$\frac{\partial LLc}{\partial g_{kd}} = \sum_{i=1}^{N} p(k \mid x_i)\left[g_{kd}^2 - \frac{g_{kd}^2}{1+e^{g_{kd}^2 (W^T x_i - a_{kd})}} - \frac{g_{kd}^2}{1+e^{g_{kd}^2 (W^T x_i - a_{kd} - r_{kd}^2)}} \right] \tag{3.36}$$

Μερική παράγωγος ως προς την παράμετρο a_{kd}

$$\frac{\partial LLc}{\partial a_{kd}} = \sum_{i=1}^{N} p(k \mid x_i)\left[\frac{2g_{kd}^2 r_{kd}}{1-e^{-g_{kd}^2 r_{kd}^2}} - \frac{2g_{kd}^2 r_{kd}}{1+e^{g_{kd}^2 (W_{kd}^T x_i - a_{kd} - r_{kd}^2)}} - \frac{2}{r_{kd}} \right] \tag{3.37}$$

Μερική παράγωγος ως προς την παράμετρο r_{kd}

$$\begin{aligned} \frac{\partial LLc}{\partial r_{kd}} = \sum_{i=1}^{N} p(k \mid x_i)\Bigg[& -2g_{kd} W_{kd}^T x_i + 2g_{kd} a_{kd} + \frac{2g_{kd} r_{kd}^2}{1-e^{-g_{kd}^2 r_{kd}^2}} + \\ & + \frac{2g_{kd} (W_{kd}^T x_i - a_{kd})}{1+e^{g_{kd}^2 (W_{kd}^T x_i - a_{kd})}} + \frac{2g_{kd} (W_{kd}^T x_i - a_{kd} - r_{kd}^2)}{1+e^{g_{kd}^2 (W_{kd}^T x_i - a_{kd} - r_{kd}^2)}} \Bigg] \end{aligned} \tag{3.38}$$

Άρα συνολικά το gradient με βάση τις παραπάνω σχέσεις ορίζεται ως εξής:

$$\nabla LLc = \left[\frac{\partial LLc}{\partial r} \quad \frac{\partial LLc}{\partial a} \quad \frac{\partial LLc}{\partial r} \right] \tag{3.39}$$

Όπως παρατηρούμε από τις σχέσεις (3.36)–(3.38) οι μερικές παράγωγοι υπολογίζονται με βάση τις παραμέτρους g_{kd}, $\alpha_{kd,}$ r_{kd} πράγμα που σημαίνει ότι η καθεμία από τις τρεις συνιστώσες του gradient (3.39) είναι ένας πίνακας διάστασης $K \times D$.

Αυτό που ίσως προξένησε αίσθηση, στην παραπάνω περιγραφή, είναι η απουσία της παραμέτρου W από την διαδικασία βελτιστοποίησης. Φυσικά, αυτό έγινε σκόπιμα για δύο βασικούς λόγους. Πρώτον, ο πίνακας αυτός έχουμε την αυστηρή απαίτηση να είναι ορθογώνιος δηλαδή να ισχύει $W^T W = I$, διαφορετικά η σταθερά κανονικοποίησης δεν θα υπολογίζεται σωστά. Αυτό πρακτικά σημαίνει ότι για να βελτιστοποιήσουμε ως προς τα στοιχεία του $W=\{w_{ij,}\ i,j=1\ldots,D\}$, θα πρέπει να βρούμε μια κλειστή μορφή που θα αναπαριστά πάντα έναν ορθογώνιο πίνακα $D \times D$ και να βελτιστοποιήσουμε ως προς τα στοιχεία του, έτσι ώστε μετά το πέρας της

βελτιστοποίησης, αυτός να παραμείνει ορθογώνιος. Για να γίνουν κατανοητά τα προηγούμενα αναφέρουμε ένα παράδειγμα ορθογώνιου πίνακα στις δύο διαστάσεις:

$$W_{orthogonal} = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \tag{3.40}$$

Όπως βλέπουμε ο πίνακας $W_{orthogonal}$ είναι ένας πίνακας περιστροφής στον οποίο φαίνεται ξεκάθαρα η εξάρτηση των στοιχείων του, όπως επίσης και το γεγονός ότι για οποιαδήποτε τιμή του θ αυτός παραμένει ορθογώνιος. Θα θέλαμε μια τέτοια μορφή του W, γενικά για πίνακες $D \times D$, αλλά δυστυχώς είναι πολύ δύσκολο, αφού ιδανικά επιδιώκουμε όχι μόνο να είναι ορθογώνιος, αλλά και η παράγωγος της LLc (3.33) ως προς τα στοιχεία του να έχει μια κλειστή μορφή για κάθε διάσταση $D \in ¥^*$

Ο δεύτερος λόγος, αφορά τον πλήθος τον παραμέτρων ως προς τις οποίες βελτιστοποιούμε, που γενικά επιθυμούμε να είναι μικρό. Η εισαγωγή και των στοιχείων του W στην όλη διαδικασία, θα απαιτούσε περισσότερη υπολογιστική ισχύ και χρόνο για την παραγωγή της τελικής λύσης, πράγμα ανεπιθύμητο. Έτσι αποκλείουμε την συμμετοχή του W από την διαδικασία της αριθμητικής βελτιστοποίησης και υιοθετούμε μια νέα τεχνική, που περιγράφεται στην συνέχεια.

3.5.3. Καθορισμός πινάκων περιστροφής W_k

Ο αλγόριθμος GEM μας επιτρέπει να εφαρμόσουμε οποιαδήποτε τεχνική εντός της επαναληπτικής διαδικασίας του, με την προϋπόθεση ότι θα υπάρχει εγγύηση για την αύξηση της πιθανοφάνειας μετά την εφαρμογή της. Εκμεταλλευόμενοι αυτό το δικαίωμα, ενσωματώνουμε στην όλη διαδικασία μια τεχνική για την εύρεση κάποιων καλών εκτιμήσεων για τους πίνακες βαρών W_k k=1,…,K. Η λογική της τεχνικής αυτής στηρίζεται στο γεγονός της ταυτόχρονης εκτίμησης, εντός της επαναληπτικής διαδικασίας του GEM, των παραμέτρων μ_k και Σ_k (κ=1,…,K) ενός GMM. Αυτό πραγματοποιείται εκτελώντας το αντίστοιχο M-βήμα για GMM, το οποίο συνίσταται από τις φόρμουλες ενημέρωσης των σχέσεων (1.39)-(1.40).

Συνεπώς, η νέα μορφή του αλγορίθμου GEM μετά την προσθήκη της παραπάνω τεχνικής συνοψίζεται ως εξής:

Loop until convergence

E-step *(for ΠsMM)*

M-step1 *(run the update formulas of GMM and compute W_k)*

M-step2 *(maximize complete likelihood of ΠsMM through BFGS and update the parameters a,r,g,π)*

End loop

Παρατηρώντας τον παραπάνω αλγόριθμο εύκολα βγάζουμε το συμπέρασμα ότι σε κάθε βήμα του GEM έχουμε στη διάθεση μας τους πίνακες συμμεταβλητότητας Σ_k και τα κέντρα μ_k για κάθε μια συνιστώσα κατανομή του GMM (k=1,…,K). Όπως φαίνεται και από το Σχήμα 3.1, τα ιδιοδιανύσματα u_{kd} (d=1,…,D) του πίνακα συμμεταβλητότητας Σ_k μας δείχνουν τις κατευθύνσεις στις οποίες υπάρχει η μεγαλύτερη διασπορά των δεδομένων. Συνεπώς, ο πίνακας U_k, ο οποίος έχει ως στήλες του τα ιδιοδιανύσματα u_{kd}, αποτελεί μια καλή εκτίμηση του πινάκα περιστροφής W_k που υπάρχει στον ορισμό της πολυδιάστατης Π-sigmoid κατανομής.

Ένα πολύ σημαντικό ερώτημα που προκύπτει από την όλη διαδικασία είναι το πώς μπορούμε να γνωρίζουμε ποια αντιστοιχία υπάρχει μεταξύ των συνιστωσών κατανομών του GMM με αυτές του ΠsMM , έτσι ώστε να είμαστε βέβαιοι ότι θα γίνουν οι σωστές ανανεώσεις σε κάθε βήμα του GEM. Απάντηση σε αυτό το ερώτημα έρχεται να δώσει η posterior πιθανότητα $p(k| x_i)$ η οποία υπολογίστηκε στο Ε-βήμα του ΠsMM. Με λίγα λόγια, όχι μόνο δεν είναι απαραίτητη η εκτέλεση ενός ξεχωριστού Ε-βήματος για το GMM, αλλά απαιτούμε να χρησιμοποιηθούν τα αποτελέσματα που παράχθηκαν από το Ε-βήμα για λογαριασμό του ΠsMM, βάσει των οποίων θα μπορέσουν να γίνουν οι ανανεώσεις των παραμέτρων μ_k, Σ_k, k=1,..,K. Έχοντας λοιπόν την posterior πιθανότητα $p(k| x_i)$ ως κοινό σημείο αναφοράς, ορίζονται αυτόματα και οι αντιστοιχίες που θέλουμε μεταξύ των συνιστωσών των δύο μοντέλων.

Η παραπάνω διαδικασία σε μορφή ψευδοκώδικα φαίνεται παρακάτω:

loop until likelihood convergence

1. **E_step** *(compute posterior probabilities [$P_{\Pi sMM}(k/x_i)$] for ΠsMM)*
2. **Run the update formulas of a GMM weighted by the posterior probabilities found at step 1.**

$$\mu_k^{new} = \frac{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i) x_i}{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i)}$$

$$\Sigma_k^{new} = \frac{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i)(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T}{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i)}$$

4. **Compute the eigenvectors of every covariance matrix Σ_k and form a matrix U_k having them as columns.**
5. **For every k make the assignment $W_k=U_k$**
6. **M_step of ΠsMM: update g, a, r, π**

End_loop

Το πρόβλημα που προκύπτει από την παραπάνω τεχνική, έτσι όπως αναλυτικά παρουσιάστηκε με την βοήθεια ψευδοκώδικα, είναι ότι μετά από κάθε ανανέωση των πινάκων περιστροφής W_k, οι υπόλοιπες παράμετροι (α_k, r_k, g_k) παύουν είναι πλέον έγκυρες, αφού η περιστροφή μετατοπίζει την λύση μας. Αυτό συμβαίνει επειδή η βελτιστοποίηση τόσο των παραμέτρων α_k, r_k και g_k από την μία, όσο και των πινάκων W_k από την άλλη γίνονται ανεξάρτητα, το οποίο πρακτικά σημαίνει ότι δεν συνυπολογίζονται οι αλλαγές της μιας κατηγορίας παραμέτρων στην διαδικασία βελτιστοποίησης της άλλης. Για να επιλύσουμε αυτό το πρόβλημα, βρίσκουμε κάποιες καλές εκτιμήσεις των παραμέτρων α_k, r_k και g_k συνυπολογίζοντας τους ανανεωμένους πίνακες περιστροφής W_k. Πρακτικά, για να το πετύχουμε αυτό, εκτελούμε πάλι την διαδικασία αρχικοποίησης του GEM χρησιμοποιώντας τις σχέσεις (3.26)-(3.30). Για τον υπολογισμό των τελευταίων γίνεται χρήση του πίνακα συμμεταβλητότητας Σ_k και του κέντρου μ_k, που υπολογίσαμε στο βήμα 2 του ψευδοκώδικα. Με αυτόν τον τρόπο όλοι οι παράμετροι ανανεώνονται έχοντας ως

κοινό σημείο αναφοράς τις ποσότητες Σ_k και μ_k του GMM, τα οποία ανανεώνονται σε κάθε βήμα του αλγορίθμου. Για την καλύτερη απόδοση της τεχνικής αυτής, επιδιώκουμε η ανανέωση των W_k να μην γίνεται σε κάθε βήμα του GEM, αλλά κάθε Η βήματα (συνήθως Η=5), για να εξασφαλίσουμε ότι ο αλγόριθμος θα προλάβει να συγκλίνει αρκετά μέχρι την επόμενη ανανέωση. Αυτό θα μας δώσει αρκετά πιο ακριβείς τιμές για τις posterior πιθανότητες που υπολογίζονται στο Ε-βήμα, οι οποίες θα προκαλέσουν με τη σειρά τους καλύτερη εκτίμηση για τις παραμέτρους Σ_k, μ_k του GMM, συνεπώς και για τις α_k, r_k και g_k.

Ένα πρόβλημα που προκύπτει μέσα από αυτή την τεχνική, είναι ότι ανανεώνοντας τους πίνακες W_k, υπάρχει το ενδεχόμενο να προκληθεί αύξηση της αρνητικής λογαριθμικής πιθανοφάνειας. Το γεγονός αυτό είναι λογικό και δεν θα πρέπει να μας ανησυχεί, αφού οι εκτιμήσεις των παραμέτρων αρχικά είναι "χονδροειδείς", και χρειάζονται μερικά βήματα του GEM για να ρυθμιστούν όπως πρέπει.

Ας υποθέσουμε τώρα ότι $W^{(t)}$ και $W^{(t-1)}$ είναι η τρέχουσα και η προηγούμενη εκτίμηση των πινάκων W και $LL_{best}(W^{(t)})$, $LL_{best}(W^{(t-1)})$ είναι οι βέλτιστες τιμές της αρνητικής λογαριθμικής πιθανοφάνειας που έχει επιτύχει ο GEM με δεδομένους τους πίνακες $W^{(t)}$ και $W^{(t-1)}$ αντίστοιχα. Για να διαπιστώσουμε αν μας συμφέρει να επιχειρήσουμε μια εκ νέου ανανέωση των πινάκων W, η οποία θα δημιουργήσει την εκτίμηση $W^{(t+1)}$, αρκεί να ελέγξουμε αν επαληθεύεται η παρακάτω συνθήκη:

$$LL_{best}(W^{(t)}) < LL_{best}(W^{(t-1)}) \tag{3.41}$$

Εάν αυτή δεν ικανοποιείται, τότε επαναφέρουμε τις εκτιμήσεις των παραμέτρων W, α, r αλλά και των posterior πιθανοτήτων $p(k|\, x_i)$, στις τιμές που υπολογίστηκαν στο βήμα του αλγορίθμου GEM κατά το οποίο επιτεύχθηκε η βέλτιστη τιμή της αρνητικής λογαριθμικής πιθανοφάνειας $LL_{best}(W^{(t-1)})$ (δηλαδή, στο Η-οστό βήμα μετά την παραγωγή της εκτίμησης $W^{(t-1)}$). Από αυτό το σημείο και έπειτα, συνεχίζουμε εκτελώντας τον αλγόριθμο GEM, χωρίς να ανανεώνουμε τους πίνακες W_k, έως ότου ο αλγόριθμος συγκλίνει. Με την παραπάνω τεχνική μπορούμε να εγγυηθούμε ότι στο Η-οστό βήμα μετά από κάθε ανανέωση των πινάκων W, η αρνητική λογαριθμική

πιθανοφάνεια θα είναι η ελάχιστη που θα έχει επιτευχθεί μέχρι εκείνη τη στιγμή το οποίο συνεπάγεται ότι θα έχουμε σύγκλιση του αλγορίθμου.

Πριν παρουσιάσουμε εκ νέου τον νέο ψευδοκώδικα με τις αλλαγές που προτάθηκαν, θα πρέπει να σημειώσουμε ότι κατά την ανανέωση των W_k ΔΕΝ αλλάζουν οι παράμετροι π_k, μιας και τις θεωρούμε ήδη καλές εκτιμήσεις που δεν επηρεάζονται σε μεγάλο βαθμό από την αλλαγή των πινάκων περιστροφής.

steps:=1

continue_updating_W_k:=TRUE

previous_best_L := Infinity

loop until likelihood convergence

1. **L_1 := negative log-likelihood**
2. **E_step** *(compute posterior probabilities [$P_{\Pi sMM}(k|\theta)$] for ΠsMM)*
3. **if steps modulus H == 0 AND continue_updating_W_k==TRUE**

   if L_1 worst than previous_best_L

   continue_updating_W_k := FALSE

   current (α_k, r_k, g_k, W_k) := previous_best (α_k, r_k, g_k, W_k)

   E_step *(recompute posterior probabilities w.r.t. new assignment made above)*

   else

   a. **previous_best_L := L_1**
   b. **previous_best(α_k, r_k, g_k, W_k) := current (α_k, r_k, g_k, W_k)**
   c. **run the update formulas of a GMM weighted by the posterior probabilities found at step 2, as shown below**

$$\mu_k^{new} = \frac{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i) x_i}{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i)} \qquad \Sigma_k^{new} = \frac{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i)(x_i - \mu_k^{new})(x_i - \mu_k^{new})^T}{\sum_{i=1}^{N} p_{\Pi sMM}(k \mid x_i)}$$

   d. **Compute the eigenvectors of every covariance matrix Σ_k^{new} and form a matrix U_k having them as columns.**
   e. **For every k=1,…,K make the assignment $W_k=U_k$ and find appropriate values for the parameters α_k, r_k, g_k using the (3.26)-(3.30) formulas w.r.t. μ_k^{new} and Σ_k^{new} computed above.**
   f. **E_step** *(recompute posterior probabilities w.r.t. new assignment made above)*

   End if

   End if

4. **M_step**
5. **Steps:=steps+1**

End_loop

Παρακάτω βλέπουμε ένα παράδειγμα με τον τρόπο που λειτουργεί η επαναληπτική διαδικασία του GEM με ενσωματωμένη την παραπάνω τεχνική για H=5.

Set initial values for parameters W, α, r, g
Previous_best_L = Infinity
(1): -Log-Likelihood = 6872.889 **parameters($W^{(0)}$, $\alpha^{(0)}$, $r^{(0)}$, $g^{(0)}$)**
(2): -Log-Likelihood = 4823.453 **parameters($W^{(0)}$, $\alpha^{(1)}$, $r^{(1)}$, $g^{(1)}$)**
(3): -Log-Likelihood = 4817.607 **parameters($W^{(0)}$, $\alpha^{(2)}$, $r^{(2)}$, $g^{(2)}$)**
(4): -Log-Likelihood = 4817.177 **parameters($W^{(0)}$, $\alpha^{(3)}$, $r^{(3)}$, $g^{(3)}$)**
(5): -Log-Likelihood = 4817.141 **parameters($W^{(0)}$, $\alpha^{(4)}$, $r^{(4)}$, $g^{(4)}$)**

Current_best_L = 4817.141 < Previous_best_L = Infinity è Update W and α, r, g
Previous_best_L := 4817.141
(6): -Log-Likelihood = 3928.108 **parameters($W^{(1)}$, $\alpha^{(5)}$, $r^{(5)}$, $g^{(5)}$)**
(7): -Log-Likelihood = 3810.314 **parameters($W^{(1)}$, $\alpha^{(6)}$, $r^{(6)}$, $g^{(6)}$)**
(8): -Log-Likelihood = 3799.986 **parameters($W^{(1)}$, $\alpha^{(7)}$, $r^{(7)}$, $g^{(7)}$)**
(9): -Log-Likelihood = 3787.307 **parameters($W^{(1)}$, $\alpha^{(8)}$, $r^{(8)}$, $g^{(8)}$)**
(10): -Log-Likelihood = 3782.270 **parameters($W^{(1)}$, $\alpha^{(9)}$, $r^{(9)}$, $g^{(9)}$)**

Current_best_L = 3782.270 < Previous_best_L = 4817.141 è Update W and α, r, g
Previous_best_L := 3782.270
(11): -Log-Likelihood = 3836.644 **parameters($W^{(2)}$, $\alpha^{(10)}$, $r^{(10)}$, $g^{(10)}$)**
(12): -Log-Likelihood = 3814.401 **parameters($W^{(2)}$, $\alpha^{(11)}$, $r^{(11)}$, $g^{(11)}$)**
(13): -Log-Likelihood = 3804.046 **parameters($W^{(2)}$, $\alpha^{(12)}$, $r^{(12)}$, $g^{(12)}$)**
(14): -Log-Likelihood = 3798.749 **parameters($W^{(2)}$, $\alpha^{(13)}$, $r^{(13)}$, $g^{(13)}$)**
(15): -Log-Likelihood = 3793.536 **parameters($W^{(2)}$, $\alpha^{(14)}$, $r^{(14)}$, $g^{(14)}$)**

Current_best_L = 3793.536 > Previous_best_L = 3782.270 è Do not update W
Set parameters back to ($W^{(1)}$, $\alpha^{(9)}$, $r^{(9)}$, $g^{(9)}$) which achieved the best likelihood value.
Re-execute E-step based on new parameters.
Continue GEM without updating W any more, until convergence.
(16): -Log-Likelihood = 3781.644 **parameters($W^{(1)}$, $\alpha^{(10)}$, $r^{(10)}$, $g^{(10)}$)**
(17): -Log-Likelihood = 3781.045 **parameters($W^{(1)}$, $\alpha^{(11)}$, $r^{(11)}$, $g^{(11)}$)**
(18): -Log-Likelihood = 3780.705 **parameters($W^{(1)}$, $\alpha^{(12)}$, $r^{(12)}$, $g^{(12)}$)**
(19): -Log-Likelihood = 3780.341 **parameters($W^{(1)}$, $\alpha^{(13)}$, $r^{(13)}$, $g^{(13)}$)**

…

3.6. Αντιμετώπιση θορύβου

Σε πολλά σύνολα δεδομένων παρατηρείται συχνά το φαινόμενο της ύπαρξης μη τυπικών τιμών σε ορισμένα πρότυπα (outliers [8]), καθώς επίσης και της παρουσίας θορύβου (noise=ανεπιθύμητα δεδομένα). Και οι δύο αυτές περιπτώσεις δεδομένων, υποβαθμίζουν την απόδοση των αλγορίθμων ομαδοποίησης, αφού προκαλούν αλλοιώσεις στο τελικό αποτέλεσμα. Για την αντιμετώπιση αυτού του φαινομένου έχουμε υιοθετήσει ειδική τεχνική, η οποία περιγράφεται στην συνέχεια.

Η λογική της τεχνικής αυτής είναι ότι εισάγουμε στο μικτό μοντέλο, ΠsMM, μια επιπλέον συνιστώσα κατανομή (θα την αποκαλούμε “background κατανομή” ή k_{BG}), η οποία καλύπτει όλο το πεδίο τιμών (domain) των προτύπων του συνόλου εκπαίδευσης. Με αυτό τον τρόπο, όλες οι ανεπιθύμητες και “ακραίες” τιμές, θα περιγράφονται από αυτή, επιτρέποντας τις υπόλοιπες να εκπαιδεύονται από τα ωφέλημα δεδομένα. Θα πρέπει να σημειωθεί ότι η background κατανομή είναι επίσης Π-sigmoid και δεν εμπλέκεται στην διαδικασία βελτιστοποίησης που πραγματοποιείται στο M-βήμα του αλγορίθμου GEM. Αντιθέτως, την λαμβάνουμε υπ’ όψιν στο E-βήμα του αλγορίθμου, υπολογίζοντας τις posterior πιθανότητες $p(x_i| k_{BG})$ και με την βοήθεια αυτών ανανεώνουμε και την prior πιθανότητα $p(k_{BG})$.

Οι αρχικές τιμές που δίνουμε στην background κατανομή, οι οποίες διατηρούνται σταθερές σε όλη τη διάρκεια της εκτέλεσης του αλγορίθμου GEM, φαίνονται παρακάτω:

Έστω ότι έχουμε ένα σύνολο δεδομένων $X = \{x_1, ..., x_N\}, x_i \in ¡^D$ τότε:

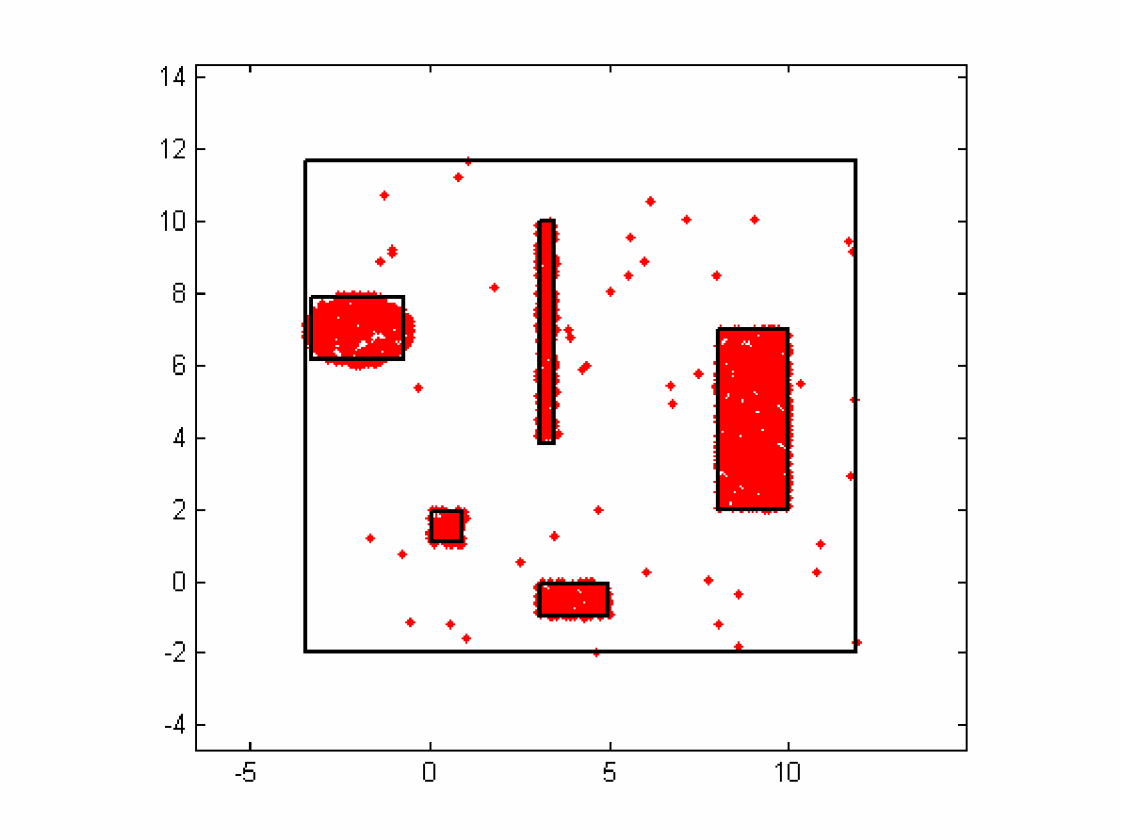
$$W_{BG} = I_D, \quad (I_D \text{ είναι ο μοναδιαίος } D \times D \text{ πίνακας}) \tag{3.42}$$

$$a_d = \min\{x_{id}\}, \ d = 1, ..., D \tag{3.43}$$

$$r_d = \sqrt{\max\{x_{id}\} - \min\{x_{id}\}}, \ d = 1, ..., D \tag{3.44}$$

$$g_d = 10, \quad d = 1, ..., D \tag{3.45}$$

Θα πρέπει να διευκρινίσουμε ότι η συνεισφορά της background κατανομής στο μικτό μοντέλο είναι πολύ μικρή. Αυτό συμβαίνει επειδή αποκτά πολύ γρήγορά μικρή prior και η τιμή που δίνει για ένα δεδομένο πρότυπο x, συνήθως είναι αρκετά μικρή. Συνεπώς, δεν τίθεται ζήτημα για την αλλοίωση του τελικού αποτελέσματος.



Σχήμα 3.7 Βλέπουμε στιγμιότυπο από την εφαρμογή του αλγορίθμου GEM σε ένα θορυβώδες dataset. Παρατηρείστε τον τρόπο τοποθέτησης της background κατανομής (εξωτερικό μαύρο ορθογώνιο).

ΚΕΦΑΛΑΙΟ 4. ΠΕΙΡΑΜΑΤΙΚΗ ΑΞΙΟΛΟΓΗΣΗ

4.1. Εισαγωγή

Σε αυτό το κεφάλαιο παρουσιάζουμε τα πειραματικά αποτελέσματα τα οποία προέκυψαν από την χρήση της κατανομής Π-sigmoid για διάφορα σύνολα δεδομένων. Ειδικότερα ο προσανατολισμός και στόχος των πειραμάτων αυτών είναι η σύγκριση της απόδοσης και συμπεριφοράς ενός μικτού μοντέλου Π-sigmoid κατανομών εν σχέσει με ένα αντίστοιχο GMM. Τα πεδία σύγκρισης των δύο μοντέλων αφορούν του εξής τρεις τομείς:

- **Σύγκριση πιθανοφάνειας σε τεχνητά δεδομένα.** Σε αυτή τη κατηγορία επιχειρούμε να εκπαιδεύσουμε τα δύο μοντέλα (ΠsMM και GMM) μέσω των αλγορίθμων GEM και EM αντίστοιχα, έχοντας στη διάθεση μας τεχνητά δεδομένα. Τα τελευταία απαρτίζονται είτε μόνο από ομοιόμορφα clusters είτε μόνο από gaussian ή από μια μίξη gaussian και ομοιόμορφων clusters. Αφού ολοκληρωθεί η εκπαίδευση των μικτών κατανομών, ως μέτρο σύγκρισης της απόδοσης και αποτελεσματικότητας χρησιμοποιούμε την τιμή της αρνητικής λογαριθμικής πιθανοφάνειας σε ένα σύνολο ελέγχου. Όσο πιο μικρή είναι η τελευταία τόσο πιο αποδοτικά γίνεται η περιγραφή του δεδομένου dataset από το εκάστοτε μοντέλο.

- **Πραγματικά δεδομένα και Ταξινόμηση.** Με χρήση πραγματικών δεδομένων [3], τα οποία είναι ήδη κατηγοριοποιημένα, προσπαθούμε να εκπαιδεύσουμε τα μοντέλα μας (supervised learning [8], ένα μικτό μοντέλο ανά κατηγορία) και να τα συγκρίνουμε βάσει της επίδοσης ταξινόμησης, όταν εφαρμοστούν σε κάποιο σύνολο ελέγχου. Γίνεται επίσης σύγκριση και της αρνητικής λογαριθμικής πιθανοφάνειας, όταν τα δεδομένα χρησιμοποιούνται χωρίς την πληροφορία της κατηγορίας.

- **Κατάτμηση Εικόνας.** Τα δεδομένα σε αυτή την κατηγορία πειραμάτων είναι τα pixel μιας grey-scale εικόνας. Στόχος αυτής της διαδικασίας είναι η κατάτμηση της εικόνας σε περιοχές, κάθε μια από τις οποίες θα έχει μια ομογενή οπτική εμφάνιση. Η απόδοση δεν μετριέται με την χρήση κάποιου μέτρου, αλλά εκ του τελικού οπτικού αποτελέσματος της κατάτμησης.

Στην συνέχεια θα παρουσιάσουμε τα αναλυτικά αποτελέσματα των πειραμάτων για κάθε μια κατηγορία ξεχωριστά.

4.2. Τεχνητά Δεδομένα

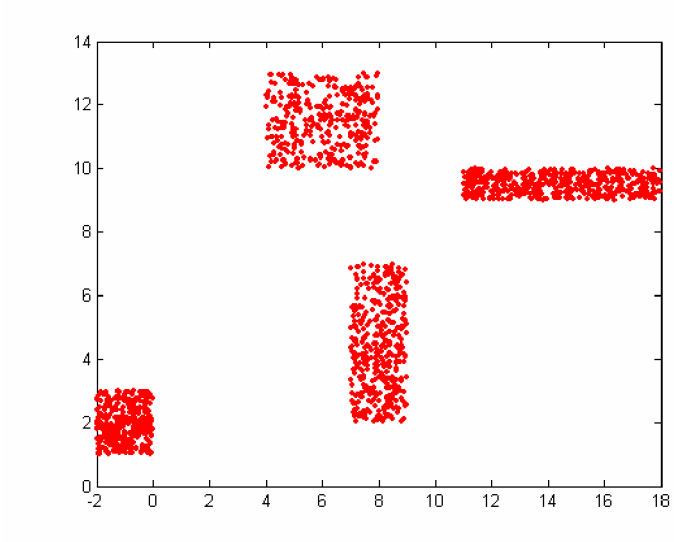
Όπως έχει ήδη αναφερθεί σε αυτή την παράγραφο θα παραθέσουμε τα πειραματικά αποτελέσματα που έγιναν με χρήση τεχνητών δεδομένων. Πιο συγκεκριμένα εκπαιδεύουμε ένα ΠsΜΜ και ένα GMM με ακριβώς των ίδιο αριθμό πυρήνων, ο οποίος να σημειωθεί ότι είναι γνωστός εκ των προτέρων. Η εκπαίδευση του GMM γίνεται μέσω του αλγορίθμου Greedy EM [2] ενώ του ΠsΜΜ γίνεται με την βοήθεια του GEM χρησιμοποιώντας ως αρχικές τιμές τις παραμέτρους του εκπαιδευμένου GMM (βλ. παρ. 3.4). Οι τύποι των τεχνητών δεδομένων που χρησιμοποιήθηκαν είναι τρεις.

Uniform. Σε αυτή την κατηγορία έχουν κατασκευαστεί σύνολα δεδομένων τα οποία αποτελούνται από μη επικαλυπτόμενα ομοιόμορφα clusters. Να σημειωθεί ότι όταν αναφερόμαστε σε ομοιόμορφα τεχνητά cluster υπονοείται ότι το σχήμα τους είναι γενικά στις D διαστάσεις ένα υπέρ-ορθογώνιο με ποικίλες διαστάσεις.

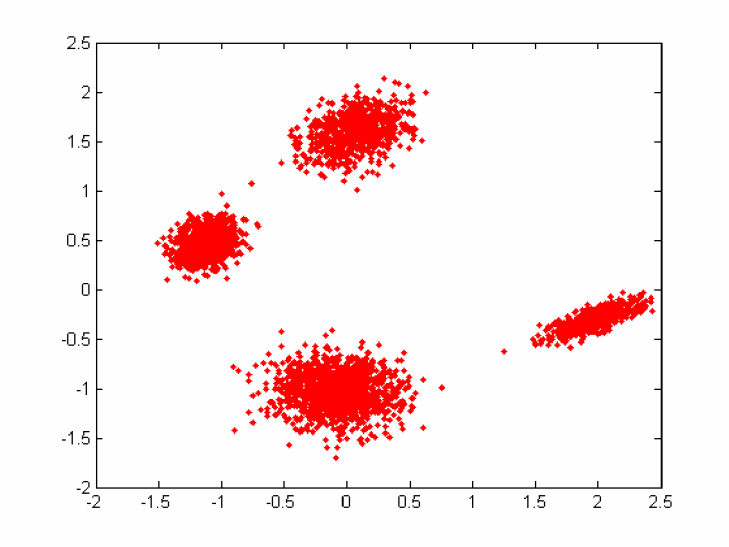
Gaussian. Σε αυτή την κατηγορία τα δεδομένα μας παράχθηκαν από την gaussian κατανομή και συγκροτούν επίσης μη επικαλυπτόμενα clusters.

Mixed. Η τελευταία αυτή κατηγόρια αποτελεί ένα συγκερασμό των δύο παραπάνω και ουσιαστικά χαρακτηρίζει ένα σύνολο δεδομένων που απαρτίζεται από ένα μίγμα uniform και gaussian clusters. Να σημειωθεί ότι για να υπάρχει δικαιοσύνη στην σύγκριση, τα ομοιόμορφα και τα gaussian clusters είναι ίσα στον αριθμό και κάθε μια από τις δύο κατηγορίες περιέχει περίπου τον ίδιο αριθμό προτύπων.

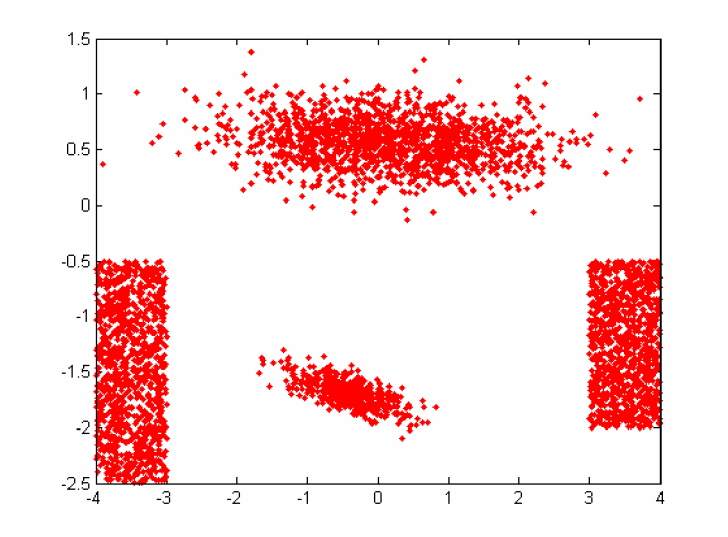
Παρακάτω βλέπουμε ένα δείγμα από κάθε μια από τις τρεις διαφορετικές κατηγορίες δεδομένων που παρουσιάστηκαν παραπάνω. Εμφανίζονται με την σειρά ομοιόμορφα, gaussian και μικτά cluster στις δύο διαστάσεις.



Σχήμα 4.1 Ομοιόμορφα δεδομένα που σχηματίζουν 4 ομάδες ορθογώνιου σχήματος.



Σχήμα 4.2 Τέσσερις ομάδες γκαουσιανών δεδομένων.



Σχήμα 4.3 Μικτές ομάδες γκαουσιανών και ομοιόμορφων δεδομένων.

Οι πίνακες με τα αποτελέσματα φαίνονται παρακάτω. Στο πάνω μέρος κάθε πίνακα αναγράφεται ο τύπός και ο αριθμός των cluster που σχηματίζουν τα δεδομένα καθώς επίσης και η διάστασή τους. Θα πρέπει να σημειωθεί ότι η απόδοση των δύο μοντέλων μετρήθηκε μέσω της αρνητικής λογαριθμικής πιθανοφάνειας τόσο στο σύνολο εκπαίδευσης όσο και σε ένα σύνολο ελέγχου. Επιθυμούμε η ποσότητα **-Log-Likelihood** που εμφανίζεται στους πίνακες να είναι όσο το δυνατόν μικρότερη.

Πίνακας 4.1 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=2, K=1.

N	**Gaussian -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=3000 TEST=3000	8544.748825 8461.348276	8551.783742 8469.122968
TRAIN=3000 TEST=3000	8079.499994 7990.747137	8087.747908 7999.669409
TRAIN=3500 TEST=3500	9715.593344 9854.417684	9727.185865 9860.078856
TRAIN=3500 TEST=3500	7621.365331 7661.740241	7628.045305 7665.860414
TRAIN=3000 TEST=3000	8369.053572 8401.785352	8377.099861 8411.393357

Πίνακας 4.2 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=3, K=1.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=4000 TEST=4000	16418.414103 16358.737562	16746.454106 16740.068878
TRAIN=4000 TEST=4000	15364.521148 15434.273201	15762.163410 15870.031079
TRAIN=3500 TEST=3500	13717.873061 13697.013105	14449.302390 14447.485123
TRAIN=4500 TEST=4500	17084.459035 17179.557548	18845.074905 18913.786844
TRAIN=4500 TEST=4500	18919.845419 18918.365886	19127.274480 19122.420870

Πίνακας 4.3 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=5, K=1.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=5500 TEST=5500	38035.818599 37995.246486	38809.059508 38748.180475
TRAIN=5500 TEST=5500	33494.525811 33547.776030	37763.593534 37589.115078
TRAIN=5000 TEST=5000	27534.942739 27667.539602	33885.874872 34108.732384
TRAIN=4500 TEST=4500	29715.289609 29794.964420	31739.491493 31711.616665
TRAIN=4500 TEST=4500	27512.090883 27318.151958	30911.823350 30618.991774

Πίνακας 4.4 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Uniform, D=2, K=1.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=3000 TEST=3000	4370.698643 4341.987882	3473.135648 3466.188617
TRAIN=2500 TEST=2500	4370.698643 4341.987882	3473.135648 3466.188617
TRAIN=2500 TEST=2500	6057.556324 6013.953912	5235.978237 5237.983913
TRAIN=2000 TEST=2000	5972.767138 5971.949124	5311.180694 5318.683431
TRAIN=2250 TEST=2250	4889.219547 4787.858290	4135.586037 4132.140314

Πίνακας 4.5 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Uniform, D=3, K=1.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=3000 TEST=3000	11632.369822 11606.975480	11025.957626 10989.313874
TRAIN=3000 TEST=3000	10732.663688 10710.310787	9216.959563 9236.013764
TRAIN=3500 TEST=3500	10641.942192 10566.309587	8712.263789 8707.284079
TRAIN=3500 TEST=3500	14780.706812 14791.813656	13192.827841 13183.468692
TRAIN=2250 TEST=2250	8345.287402 8338.912070	7839.482847 7839.230729

Πίνακας 4.6 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Uniform, D=5, K=1.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=5000 TEST=5000	29168.415085 29286.988045	28215.241251 28311.280888
TRAIN=5000 TEST=5000	30017.602242 30015.456961	28010.573970 27936.705285
TRAIN=4750 TEST=4750	30220.576492 30296.226867	29211.227524 29190.396330
TRAIN=4500 TEST=4500	25552.068493 25534.784373	23623.714954 23636.812920
TRAIN=5500 TEST=5500	36784.612228 36889.224389	37428.190370 37479.273910

Πίνακας 4.7 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=2, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=4001 TEST=3999	3144.011558 3172.283792	3152.092343 3185.275085
TRAIN=4501 TEST=4499	4524.412060 4568.390303	4531.081266 4576.565689
TRAIN=4501 TEST=4499	9143.198068 9153.617989	9152.767795 9167.475634
TRAIN=4501 TEST=4499	5094.469254 5072.443805	5094.396658 5093.511996
TRAIN=5001 TEST=4999	4870.057674 4622.388452	4875.705822 4629.182457

Πίνακας 4.8 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=3, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=5001 TEST=4999	2339.333611 2265.208443	2369.525886 2300.351008
TRAIN=5001 TEST=4999	11103.357161 11117.078101	11124.332009 11122.832329
TRAIN=6001 TEST=5999	6856.016869 6867.547052	6908.548684 6883.636926
TRAIN=6001 TEST=5999	27.826558 242.156450	76.486142 293.312345
TRAIN=5501 TEST=5499	5616.190008 5428.897228	5657.416999 5472.481584

Πίνακας 4.9 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=5, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=6751 TEST=6749	15092.009286 15404.482618	15154.024487 15469.343439
TRAIN=6751 TEST=6749	10832.259351 11030.184720	10878.717660 11071.584693
TRAIN=7751 TEST=7749	-971.667112 -1108.117024	-896.037671 -1053.224405
TRAIN=7001 TEST=6999	1698.396378 1409.446955	1815.587089 1507.856030
TRAIN=8501 TEST=8499	-9748.499238 -9585.282394	-9693.015001 -9542.284594

Πίνακας 4.10 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Uniform, D=2, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=2500 TEST=2500	7771.808281 7665.562388	6883.015647 6876.978193
TRAIN=2500 TEST=2500	8643.847986 8621.271026	7766.967100 7759.071955
TRAIN=2750 TEST=2750	10289.813058 10234.783265	9730.178462 9727.579673
TRAIN=2500 TEST=2500	6536.245354 6570.719102	5745.670990 5752.493932
TRAIN=2500 TEST=2500	8040.575690 8145.874973	7290.962282 7315.652896

Πίνακας 4.11 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Uniform, D=3, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=2750 TEST=2750	11417.187448 11436.645048	10582.225669 10578.800227
TRAIN=3250 TEST=3250	14979.821906 14989.342636	13494.987930 13486.908888
TRAIN=3250 TEST=3250	12047.978526 11978.906264	10518.911932 10506.930422
TRAIN=3000 TEST=3000	13164.458982 13141.176657	12040.200386 12024.353198
TRAIN=2750 TEST=2750	11034.322202 11016.211759	10079.185671 10022.396399

Πίνακας 4.12 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Uniform, D=5, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=6750 TEST=6750	46960.491194 47030.380675	44606.928932 44738.403894
TRAIN=7250 TEST=7250	47320.261137 47470.805964	43440.393155 43536.887441
TRAIN=7750 TEST=7750	52514.667714 52542.175957	49853.467551 49942.384210
TRAIN=7000 TEST=7000	48279.607477 48326.965765	44950.597033 44974.657535
TRAIN=8500 TEST=8500	61658.959495 61601.603219	58340.535779 58384.374511

Πίνακας 4.13 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=2, K=4.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=5001 TEST=4999	-1915.763515 -2080.768630	-1881.165979 -2037.865227
TRAIN=5001 TEST=4999	-175.360243 -243.496933	-139.501296 -218.717097
TRAIN=5501 TEST=5499	2543.450125 2707.164456	2553.507910 2722.184426
TRAIN=5501 TEST=5499	1826.930916 2018.231266	1840.122294 2042.831350
TRAIN=5501 TEST=5499	3891.398492 3817.018202	3919.081583 3853.890051

Πίνακας 4.14 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=3, K=4.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=5501 TEST=5499	-5641.849796 -5621.753686	-5607.229050 -5593.240522
TRAIN=5401 TEST=5499	-185.969790 -150.697376	-176.850672 -120.989303
TRAIN=6501 TEST=6499	1500.475648 1615.270122	1515.748198 1635.932717
TRAIN=6001 TEST=5999	2967.448283 2907.806051	2993.224897 2942.398150
TRAIN=6001 TEST=5999	-6837.326194 -6787.099362	-6808.278690 -6732.739904

Πίνακας 4.15 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: Gaussian, D=5, K=4.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=6501 TEST=6499	-13728.007697 -13549.551447	-13666.407533 -13489.467235
TRAIN=6401 TEST=6499	-4154.373025 -3836.355749	-4118.022613 -3761.836411
TRAIN=7001 TEST=6999	-17738.669929 -17722.744161	-17690.126472 -17655.441994
TRAIN=7001 TEST=6999	-15636.802675 -15381.608582	-15590.202874 -15319.349902
TRAIN=7001 TEST=6999	-21477.346530 -21337.230472	-21422.260375 -21263.204054

Πίνακας 4.16 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: mixed, D=2, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=3000 TEST=3000	9489.151278 9765.170424	9038.791871 9270.542632
TRAIN=3000 TEST=3000	8967.060691 8894.131111	8467.029009 8409.182820
TRAIN=3000 TEST=3000	7219.557409 7242.407937	6717.590579 6762.815155
TRAIN=3000 TEST=3000	9932.879015 9967.684083	9378.599665 9446.572687
TRAIN=3000 TEST=3000	10304.590247 10443.227199	9818.327860 9999.225085

Πίνακας 4.17 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: mixed, D=3, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=4000 TEST=4000	18427.971059 18623.266115	17894.763667 18118.067998
TRAIN=4000 TEST=4000	18574.010329 18631.793671	17586.020310 17714.095618
TRAIN=4000 TEST=4000	17452.840943 17527.545476	16499.670285 16512.355486
TRAIN=4000 TEST=4000	15973.112717 16068.227005	14909.121520 15072.518924
TRAIN=4000 TEST=4000	16540.830409 16519.853346	16214.532877 16184.873731

Πίνακας 4.18 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: mixed, D=5, K=2.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=4000 TEST=4000	28228.491312 28383.317075	27292.066151 27479.484845
TRAIN=4000 TEST=4000	28751.688295 28868.617478	27754.061235 27827.555682
TRAIN=4000 TEST=4000	28842.808777 29153.369845	27771.445776 27982.218284
TRAIN=4000 TEST=4000	26071.020117 25826.558160	25126.249812 24964.577528
TRAIN=4000 TEST=4000	24430.998247 24352.793639	23619.850545 23492.441247

Πίνακας 4.19 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: mixed, D=2, K=4.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=5501 TEST=5500	11239.811823 11330.637156	10223.641922 10320.381648
TRAIN=5501 TEST=5500	9147.238276 9281.667256	8212.739461 8330.646952
TRAIN=5501 TEST=5500	9911.832231 10011.768014	8991.242089 9009.221191
TRAIN=55001 TEST=5500	13212.631062 13487.038957	12266.984022 12533.911526
TRAIN=5501 TEST=5500	12629.827828 12653.825936	11706.236616 11658.104711

Πίνακας 4.20 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: mixed, D=3, K=4.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=6001 TEST=6000	19093.677351 19252.781506	18606.178544 18790.019159
TRAIN=6001 TEST=6000	22338.389971 22299.490320	20841.104120 20846.070441
TRAIN=6001 TEST=6000	26165.758408 26318.241997	22909.134783 23012.608072
TRAIN=6001 TEST=6000	45053.242927 45051.149577	41323.281855 41290.093846
TRAIN=6001 TEST=6000	22205.880684 22276.428857	21203.380269 21277.744435

Πίνακας 4.21 Η τιμή της αρνητικής λογαριθμικής πιθανοφάνειας μετά την εκτίμηση παραμέτρων για τα GMM και ΠsMM. Data Type: mixed, D=5, K=4.

N	**GMM -Log-Likelihood**	**ΠsMM -Log-Likelihood**
TRAIN=7001 TEST=7000	29897.725482 29951.207013	28603.889243 28605.220667
TRAIN=7001 TEST=7000	42041.328965 41934.800445	38878.731757 39143.885077
TRAIN=7001 TEST=7000	46189.073250 46251.109387	43803.169032 43868.499414
TRAIN=7001 TEST=7000	38275.933398 38452.625216	35114.152968 35204.839557
TRAIN=7001 TEST=7000	49055.596310 49044.795497	46634.811940 46616.578420

Παρατηρώντας τα αποτελέσματα των πειραμάτων, μπορούμε να βγάλουμε κάποια συμπεράσματα. Καταρχάς είναι εμφανές ότι η αρνητική λογαριθμική πιθανοφάνεια που δίνει το ΠsMM στην περίπτωση των unifom δεδομένων είναι σαφώς μικρότερη από την αντίστοιχη του GMM. Ακριβώς το ίδιο ισχύει και για τα mixed δεδομένα. Η εξήγηση που πρακτικά μπορούμε να δώσουμε για τεκμηριώσουμε την διαφορετική συμπεριφορά των δύο αυτών μοντέλων, έγκειται στις ιδιότητες των Π-sigmoid και Gaussian κατανομών, που συνιστούν τα αντίστοιχα μοντέλα. Η καλή γενικευτική ικανότητα που έχει η Π-sigmoid κατανομή μας δίνει το δικαίωμα της ικανοποιητικής περιγραφής τόσο των ομοιόμορφων όσο και των gaussian δεδομένων. Από την άλλη, τα πειράματα αναδεικνύουν τα μειονεκτήματα της κανονικής κατανομής, η οποία αδυνατεί να ανταποκριθεί ικανοποιητικά σε περιπτώσεις όπου υπάρχουν ομοιόμορφα clusters Έτσι παρατηρούμε μια σαφώς καλύτερη επίδοση του ΠsMM, μιας και η ικανότητα να περιγραφή ομοιόμορφα δεδομένα είναι αρκετά υποβοηθητική.

Η σύγκριση των δύο μοντέλων μπορεί να συνεχιστεί και στα gaussian δεδομένα, τα οποία η Gaussian κατανομή, κατ' επέκταση ένα GMM, προφανώς μπορεί να περιγράψει με βέλτιστο τρόπο. Όπως ήταν αναμενόμενο το GMM αποδίδει καλύτερα

σε όλα τα πειράματα με αυτού του είδους τα δεδομένα. Παρόλα αυτά όμως, ακόμα και σε αυτή την περίπτωση, η αρνητική λογαριθμική πιθανοφάνεια των δύο μοντέλων αρκετές φορές συμβαδίζει χωρίς να υπάρχουν εντυπωσιακές διαφορές ανάμεσά τους. Αυτό πρακτικά αναδεικνύει με ένα ακόμα τρόπο, την γενικευτική ικανότητα του ΠsMM έναντι του αντίστοιχου GMM, καθιστώντας το κατάλληλο για περιγραφή δεδομένων με άγνωστες στατιστικές ιδιότητες.

4.3. Πραγματικά δεδομένα και classification

Σε αυτή την κατηγορία των πειραμάτων θα συγκρίνουμε τα δύο μοντέλα (ΠsMM, GMM) σε πραγματικά δεδομένα και ειδικότερα στην διαδικασία της κατηγοριοποίησης. Τα δεδομένα που χρησιμοποιήθηκαν αντλήθηκαν από την βάση UCI [3] που υπάρχει στο διαδίκτυο. Να σημειωθεί ότι, τα δεδομένα προφανώς ήταν ταξινομημένα εκ των προτέρων, συνεπώς σε αυτή την περίπτωση έχουμε μάθηση με επίβλεψη. Για να εκπαιδεύσουμε και στην συνέχεια να συγκρίνουμε τα δύο μοντέλα μας, ήταν απαραίτητος ο διαχωρισμός του κάθε συνόλου δεδομένων σε δύο βασικά υποσύνολα, το σύνολο εκπαίδευσης X_{Tr} και το σύνολο ελέγχου X_{Ts}. Η εκπαίδευση των μοντέλων γίνεται μόνο με βάση το σύνολο εκπαίδευσης και πραγματοποιείται με τον παρακάτω τρόπο.

Χρησιμοποιώντας την ετικέτα κατηγορίας (class label) του κάθε προτύπου, που μας πληροφορεί για την κατηγορία στην οποία ανήκει, δημιουργούμε C υποσύνολα $X_{Tr}=\{X_{Tr\_1},X_{Tr\_2},\ldots,X_{Tr\_C}\}$ του συνόλου εκπαίδευσης, όσες δηλαδή είναι συνολικά οι κατηγορίες των δεδομένων. Για κάθε ένα από τα υποσύνολα X_{Tr_c}, c=1,…,C δημιουργούμε δύο μικτά μοντέλα, ένα ΠsMM και ένα GMM, και τα εκπαιδεύουμε ξεχωριστά με κάποιο αυθαίρετο αριθμό πυρήνων K_c. Να σημειωθεί ότι πλέον η εκπαίδευση στα νέα υποσύνολα X_{Tr_c}, γίνεται χωρίς επίβλεψη, επειδή δεν έχουμε κάποια εκ των προτέρων γνώση για τα δεδομένα της κάθε κατηγορίας. Αφού δημιουργήσουμε τα 2*C εκπαιδευμένα μικτά μοντέλα (C ΠsMMs και C GMMs) χρησιμοποιούμε το σύνολο ελέγχου για να ανιχνεύσουμε την απόδοση του κάθε συστήματος. Προτού αναφέρουμε τον τρόπο με τον οποίο έγινε η κατηγοριοποίηση κρίνεται απαραίτητη η αναφορά στο γεγονός της αραιότητας των δεδομένων λόγω

του μικρού πλήθους και της μεγάλης διάστασης που έχουν. Για να είναι ακριβή τα αποτελέσματα καταφύγαμε στην τεχνική K-fold cross validation [8] όπου στο K δώσαμε τις τιμές 3, 5 και 10 ανάλογα με μέγεθος του προβλήματος. Θα παρουσιάσουμε στην συνέχεια τον τρόπο με τον οποίο έγινε η λήψη απόφασης για τα δεδομένα του συνόλου ελέγχου. Ορίζουμε ως ΠsMM_c και GMM_c τα μικτά μοντέλα που εκπαιδεύσαμε με τα δεδομένα του συνόλου X_{Tr_c}, τότε με την βοήθεια του Bayes rule [8], ένα πρότυπο $x \in X_{Ts}$ τοποθετείται στην κατηγορία που μεγιστοποιεί την παρακάτω ποσότητα

p(c)*ΠsMM_c(x), όσον αφορά τα μικτά μοντέλα Π-sigmoid κατανομών.

και

p(c)*GMM_c(x), όσον αφορά τα μικτά μοντέλα gaussian κατανομών.

όπου p(c) είναι η prior πιθανότητα της κατηγορίας c και ορίζεται ως $p(c)= |X_{Tr\_c}|/|X_{Tr}|$. Τα ΠsMM_c(x) και GMM_c(x) είναι οι τιμές που επιστρέφουν τα αντίστοιχα εκπαιδευμένα μικτά μοντέλα της κατηγορίας c, για δεδομένη είσοδο x προερχόμενη από το σύνολο ελέγχου.

Τα αποτελέσματα των πειραμάτων φαίνονται στους παρακάτω πίνακες.

Πίνακας 4.22 Σύγκριση της απόδοσης των μοντέλων GMM και ΠsMM στην κατηγοριοποίηση πραγματικών δεδομένων.

DataSet	**N**	**D**	**K**	**GMM Classifier (score %)**	**ΠsMM Classifier (score %)**
segment	*5-fold cross validation*	18	2	82.51 %	85.58 %
			3	84.89 %	87.05 %

iris	*10-fold cross validation*	4	2	96.66 %	94.66 %
			3	95.33 %	92.66 %
			4	95.33 %	94.66 %
new_thyroid	*10-fold cross validation*	5	2	97.14 %	95.23 %
			3	96.66 %	92.38 %
			4	96.19 %	92.38 %
vehicles	*5-fold cross validation*		2	65.24 %	65.00 %
			4	69.58 %	68.26 %
phoneme	*3-fold cross validation*	5	2	77.81 %	79.12 %
			3	77.70 %	77.25 %
			4	77.79 %	77.81 %
waveform	*3-fold cross validation*	21	3	85.61 %	85.83 %
			4	86.05 %	86.09 %
			5	85.99 %	85.71 %
satimage	*3-fold cross validation*	36	2	82.05 %	85.46 %
			3	82.55 %	85.06 %
			4	84.42 %	85.93 %
bupa	*10-fold cross validation*	6	2	58.23 %	58.82 %
			3	64.11 %	59.11 %
			4	62.94 %	62.64 %
wdbc	*5-fold cross validation*	30	2	94.69 %	95.93 %
			3	93.63 %	95.04 %
			4	95.22 %	94.51 %

Από τα παραπάνω αποτελέσματα παρατηρούμε ότι τα αποτελέσματα είναι σχεδόν μοιρασμένα για τα δύο μοντέλα. Είναι αξιοσημείωτο το γεγονός ότι η ο ταξινομητής ΠsMM αποδίδει καλύτερα σε πυκνά σύνολα με μεγάλο πλήθος δεδομένων.

Στην συνέχεια παρουσιάζουμε τα συγκριτικά αποτελέσματα αναφορικά με την αρνητική λογαριθμική πιθανοφάνεια πάντα σε πειράματα που πραγματοποιήθηκαν σε πραγματικά δεδομένα, χωρίς να χρησιμοποιηθεί η ετικέτα της κατηγορίας κάθε προτύπου. Ακολουθεί ο πίνακας των αποτελεσμάτων.

Πίνακας 4.23 Σύγκριση των μοντέλων GMM και ΠsMM χρησιμοποιώντας ως μέτρο την αρνητική λογαριθμική πιθανοφάνεια.

DataSet	**N**	**D**	**K**	**GMM Log-Likelihood**	**ΠsMM Log-Likelihood**
bupa	345	6	2	2298.940311	2271.875669
			4	2060.877723	2124.401742
			6	1965.393179	2033.033080
cleveland	297	13	2	4762.890461	4483.444856
diabetes	768	8	2	18099.179114	16835.801037
			4	6519.058297	6217.526886
			6	6364.727873	5837.024420
Iris	150	4	3	302.976060	295.697915
			5	228.457641	229.861335
			7	210.974570	201.681522
new_thyroid	215	5	3	438.232594	423.321941
			5	383.177149	369.238736
			7	342.546645	318.710524
nimegen	120	7	2	937.792154	918.759800
wine	178	13	3	2102.825267	2026.895518
			5	1964.994488	1910.273839
			7	1659.344153	1594.969707

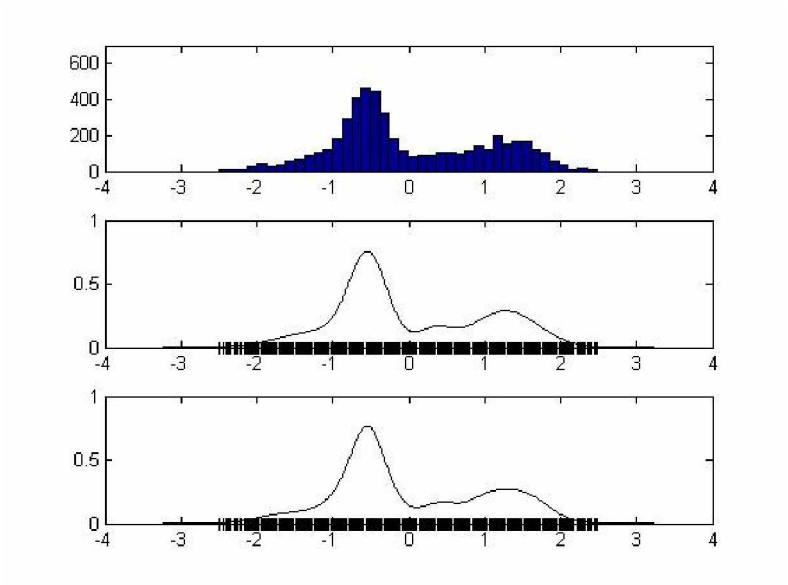
vowel	990	10	11	8843.395248	8573.390498
			13	8140.268436	7872.409010
			15	7846.727014	7544.483090
Satimage	6435	36	6	-42452.401095	-43811.572334
	TR= 3217	36	6	-22238.453664	-22999.499974
	TST=3217			-16482.286488	-17256.508297
	TR= 3217	36	8	-23780.026324	-24512.337880
	TST=3217			-14947.751092	-15996.032063
	TR= 3217	36	10	-25680.481166	-26333.854304
	TST=3217			-15621.906201	-16436.652624
waveform	5000	21	3	118173.650817	117694.297444
	TR= 2500	21	3	58916.758901	58811560392
	TST=2500			59744.978338	59668.023600
	TR= 2500	21	5	58692.499832	58552.647787
	TST=2500			60453.866750	60397.746374
	TR= 2500	21	7	57827.478167	57692.607487
	TST=2500			60066.191543	60006.957259

Τα μεγάλα σύνολα δεδομένων waveform, Satimage μας δόθηκε η ευκαιρία να τα χωρίσουμε σε σύνολο εκπαίδευσης και σύνολο έλεγχου. Με αυτό τον τρόπο μπορούμε να έχουμε μια πιο σφαιρική εικόνα για την γενικευτική ικανότητα των μοντέλων αυτών. Παρατηρούμε ότι σε ελάχιστες περιπτώσεις υπερτερεί το GMM μοντέλο έναντι του ΠsMM. Όπως είδαμε και στα τεχνητά δεδομένα η επαυξημένη ικανότητα μοντελοποίησης της Π-sigmoid κατανομής είναι εμφανής και σε πραγματικά δεδομένα.

4.4. Ομαδοποίηση Εικονοστοιχείων

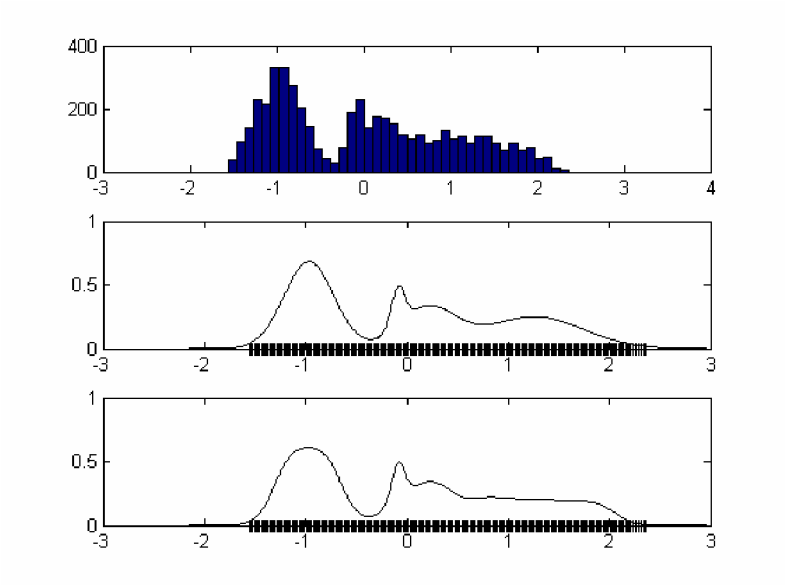
Η τελευταία αυτή κατηγορία πειραμάτων αφορά την εφαρμογή ενός ΠsΜΜ στην διαδικασία κατάτμησης μιας εικόνας. Αντικειμενικός στόχος αυτής της εφαρμογής είναι η ανίχνευση ομογενών περιοχών, με βάση πάντα την απόχρωση του γκρι του κάθε εικονοστοιχείου (pixel). Όπως είναι εύκολα αντιληπτό, τα δεδομένα μας πλέον είναι μονοδιάστατα και το πλήθος του είναι W*H, όπου H και W είναι οι διαστάσεις της εικόνας. Η διαδικασία που ακολουθείται, είναι ότι εκπαιδεύουμε πάλι τα δύο μοντέλα, ΠsΜΜ και GMM, με απώτερο στόχο την ομαδοποίηση των pixels σε K ομάδες (η τιμή του K στα πειράματα επιλέγεται αυθαίρετα). Επειδή σε μια τυπική εικόνα με ανάλυση 256×256, τα pixels είναι 65536, το οποίο θεωρείται μεγάλος αριθμός για δεδομένα εκπαίδευσης, πραγματοποιούμε μια ομοιόμορφη δειγματοληψία και περιορίζουμε έτσι τον όγκο των δεδομένων. Ένας τυπικός αριθμός για την ποσότητα των pixels που χρησιμοποιήθηκαν για την εκπαίδευση των μοντέλων είναι το 5000.

Από την στιγμή που η εκπαίδευση ολοκληρωθεί, ομαδοποιούμε τα pixels με βάση την posterior πιθανότητα που υπολογίστηκε στο E-βήμα των αλγορίθμων GEM και GrEM, για το ΠsΜΜ και GMM αντίστοιχα. Τα αποτελέσματα της κατάτμησης μαζί με το ιστόγραμμα της κάθε εικόνας και τις γραφικές αναπαραστάσεις των δύο μοντέλων φαίνονται παρακάτω.

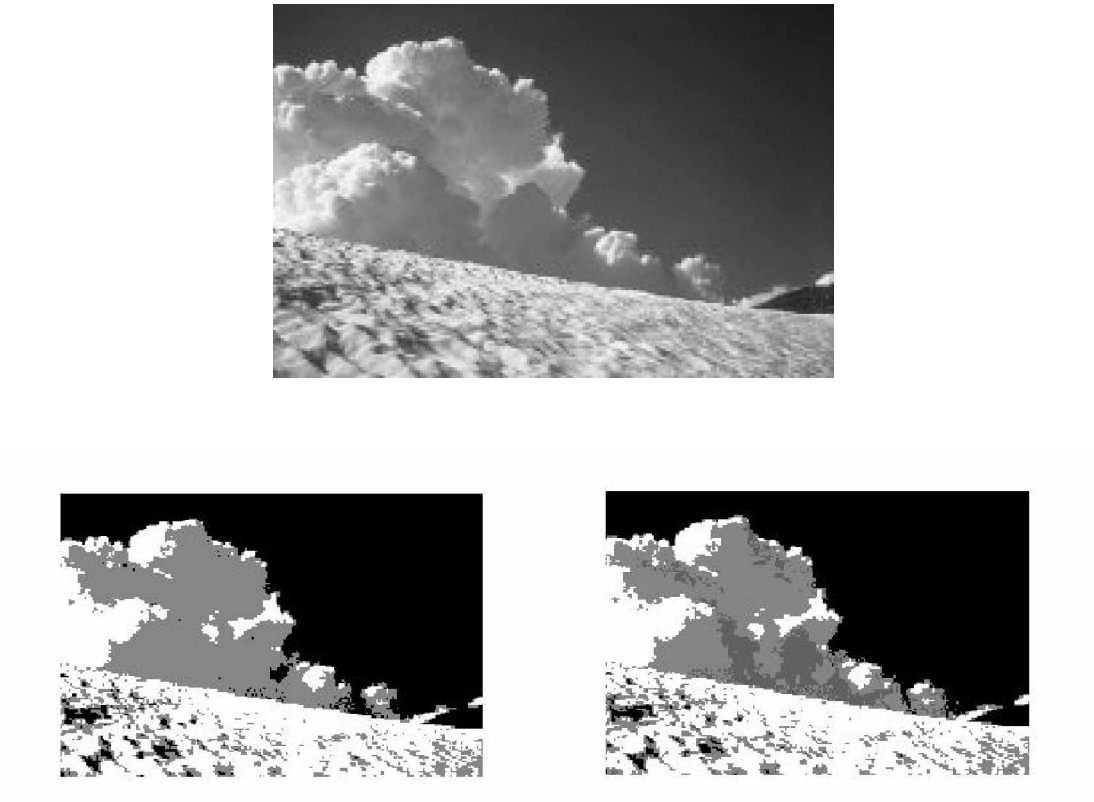


Σχήμα 4.4 Πάνω, το ιστόγραμμα της εικόνας “amakses.jpg”. Η δεύτερη παράσταση απεικονίζει το γράφημα της GMM λύσης. Και η τελευταία, το γράφημα της ΠsΜΜ λύσης. Ο αριθμός των συνιστωσών κατανομών είναι 4.

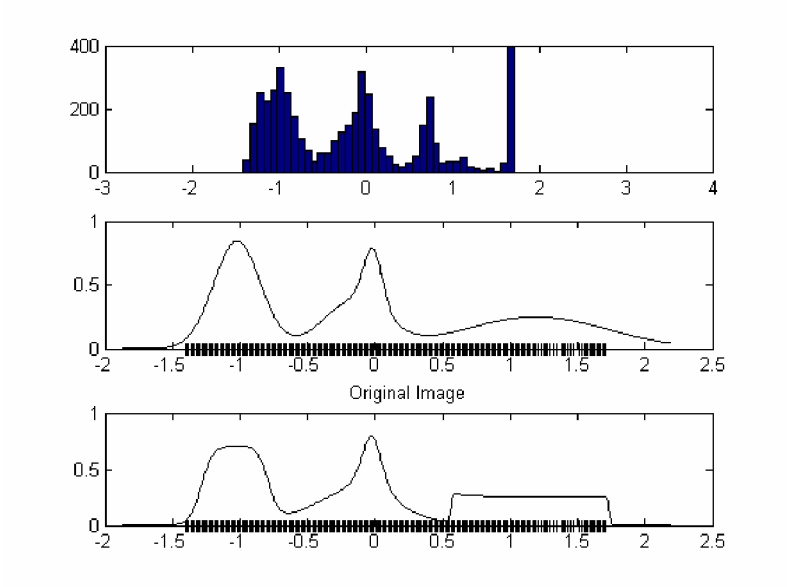
Σχήμα 4.5 Πάνω η αρχική εικόνα “amakses.jpg”. Κάτω αριστερά το αποτέλεσμα της κατάτμησης από το GMM. Κάτω δεξιά, το αποτέλεσμα της κατάτμησης από το ΠsΜΜ. Ο αριθμός των συνιστωσών κατανομών είναι 4.



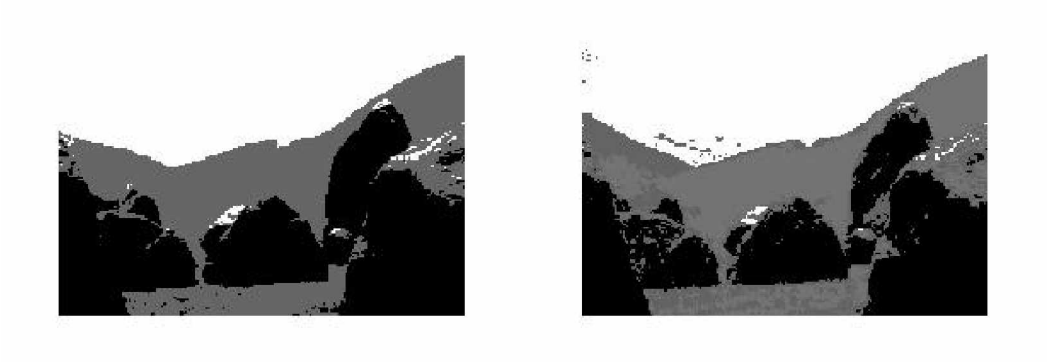
Σχήμα 4.6 Πάνω, το ιστόγραμμα της εικόνας “clouds.jpg”. Η δεύτερη παράσταση απεικονίζει το γράφημα της GMM λύσης. Και η τελευταία, το γράφημα της ΠsMM λύσης. Ο αριθμός των συνιστωσών κατανομών είναι 4.



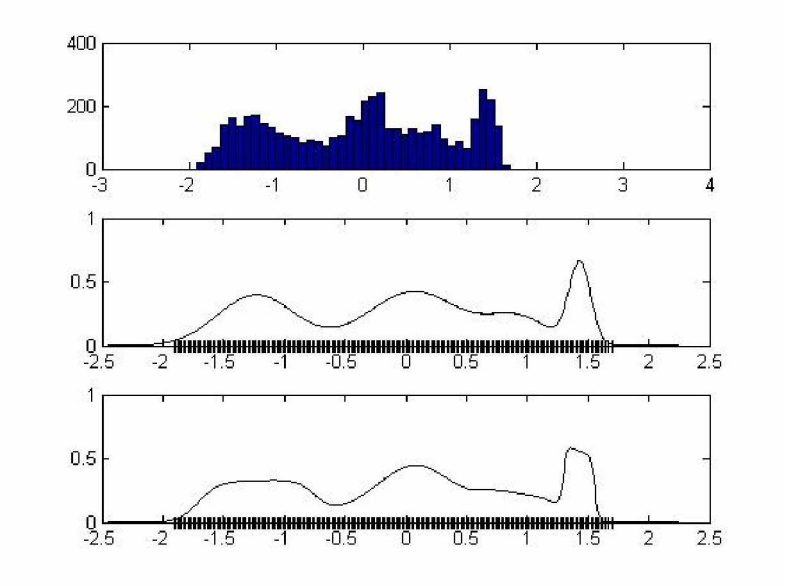
Σχήμα 4.7 Πάνω η αρχική εικόνα “clouds.jpg”. Κάτω αριστερά το αποτέλεσμα της κατάτμησης από το GMM. Κάτω δεξιά, το αποτέλεσμα της κατάτμησης από το ΠsMM Ο αριθμός των συνιστωσών κατανομών είναι 4.



Σχήμα 4.8 Πάνω, το ιστόγραμμα της εικόνας “rocks.jpg”. Η δεύτερη παράσταση απεικονίζει το γράφημα της GMM λύσης. Και η τελευταία, το γράφημα της ΠsMM λύσης. Ο αριθμός των συνιστωσών κατανομών είναι 4.

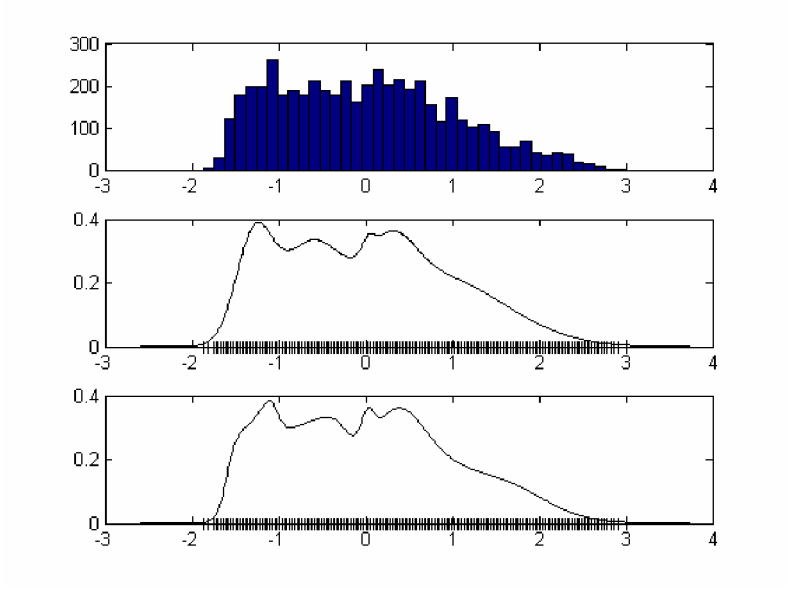


Σχήμα 4.9 Πάνω η αρχική εικόνα “rocks.jpg”. Κάτω αριστερά το αποτέλεσμα της κατάτμησης από το GMM. Κάτω δεξιά, το αποτέλεσμα της κατάτμησης από το ΠsMM. Ο αριθμός των συνιστωσών κατανομών είναι 4.



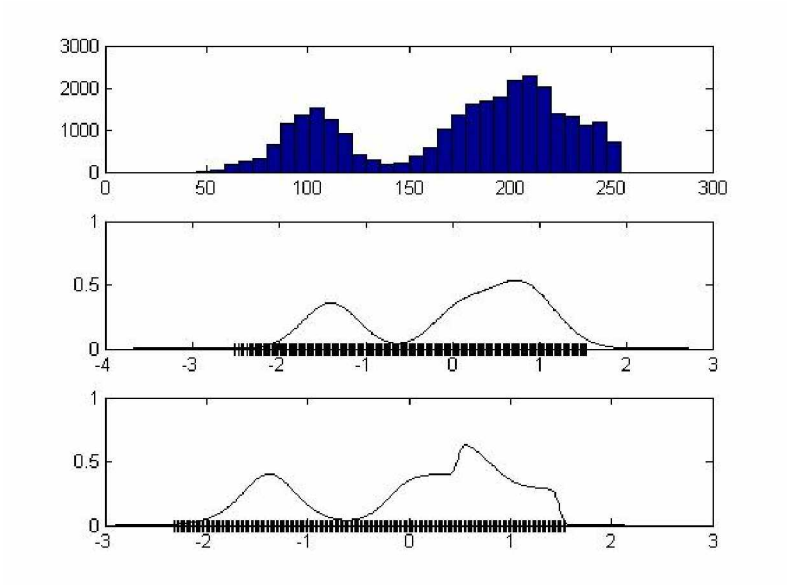
Σχήμα 4.10 Πάνω, το ιστόγραμμα της εικόνας “woman.jpg”. Η δεύτερη παράσταση απεικονίζει το γράφημα της GMM λύσης. Και η τελευταία, το γράφημα της ΠsMM λύσης. Ο αριθμός των συνιστωσών κατανομών είναι 4

Σχήμα 4.11 Αριστερά, η αρχική εικόνα “woman.jpg”. Στην μέση, το αποτέλεσμα της κατάτμησης από το GMM. Δεξιά, το αποτέλεσμα της κατάτμησης από το ΠsMM. Ο αριθμός των συνιστωσών κατανομών είναι 4.

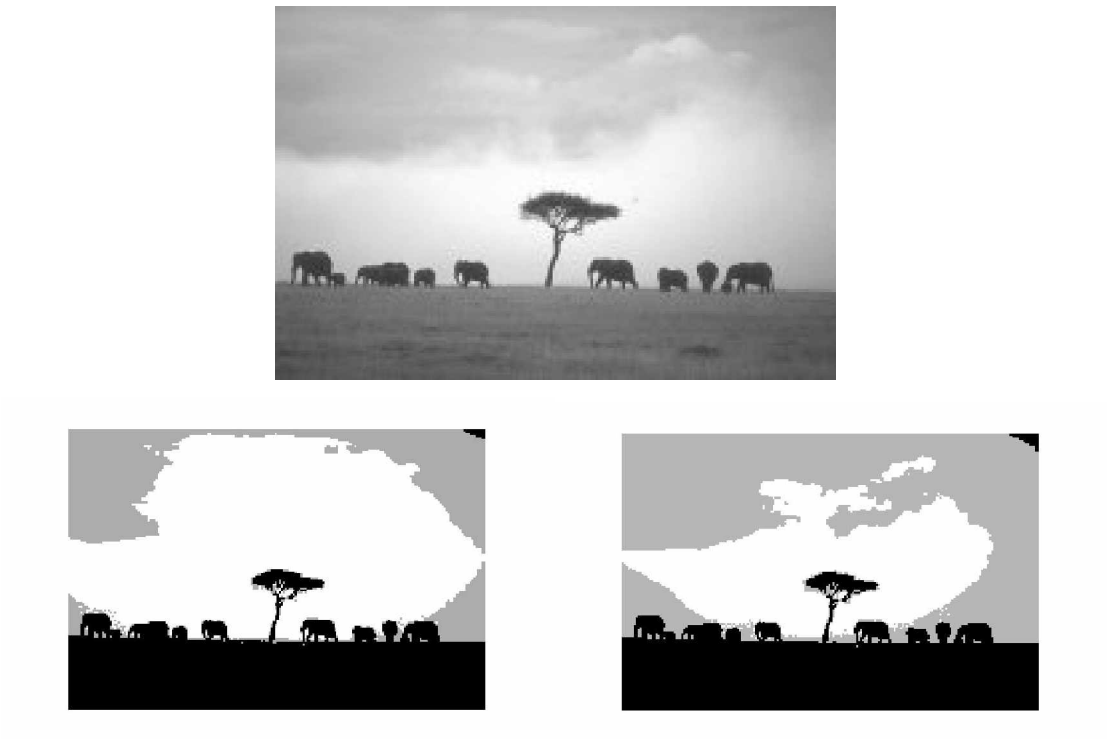


Σχήμα 4.12 Πάνω, το ιστόγραμμα της εικόνας “rocks-tree.jpg”. Η δεύτερη παράσταση απεικονίζει το γράφημα της GMM λύσης. Και η τελευταία, το γράφημα της ΠsMM λύσης. Ο αριθμός των συνιστωσών κατανομών είναι 5.

Σχήμα 4.13 Πάνω η αρχική εικόνα “rocks-tree.jpg”. Κάτω αριστερά το αποτέλεσμα της κατάτμησης από το GMM. Κάτω δεξιά, το αποτέλεσμα της κατάτμησης από το ΠsMM Ο αριθμός των συνιστωσών κατανομών είναι 5.



Σχήμα 4.14 Πάνω, το ιστόγραμμα της εικόνας “elephants.jpg”. Η δεύτερη παράσταση απεικονίζει το γράφημα της GMM λύσης. Και η τελευταία, το γράφημα της ΠsMM λύσης. Ο αριθμός των συνιστωσών κατανομών είναι 4



Σχήμα 4.15 Πάνω η αρχική εικόνα “elephants.jpg”. Κάτω αριστερά το αποτέλεσμα της κατάτμησης από το GMM. Κάτω δεξιά, το αποτέλεσμα της κατάτμησης από το ΠsMM. Ο αριθμός των συνιστωσών κατανομών είναι 4.

Τα αποτελέσματα της κατάτμησης είναι σαφώς ενθαρρυντικά αφού παρατηρούμε ότι σε όλες σχεδόν τις περιπτώσεις το ΠsMM αποδίδει καλύτερα από το αντίστοιχο GMM. Είναι πολύ σημαντικό να παρατηρήσουμε τη γραφική παράσταση που παράγει το κάθε ένα μοντέλο μετά την εκπαίδευση του. Βλέπουμε λόγου χάρη, στην εικόνα "elephants.jpg" ότι το GMM, ενώ σταματούν στα αριστερά να υφίστανται δεδομένα (λεπτές μαύρες κατακόρυφες γραμμές επί του άξονα) εκείνο συνεχίζει να "απλώνεται" στο κενό, δίνοντας μια κακή περιγραφή των δεδομένων. Αντιθέτως, στην ίδια εικόνα, η παράσταση του ΠsMM, κατέρχεται με ένα απότομο τρόπο στο τέλος των δεδομένων ταιριάζοντας ικανοποιητικά σε αυτά Σχήμα 4.14.

Μέσα από το προηγούμενη παρατήρηση, η οποία υφίσταται σε κάποιο βαθμό και στις υπόλοιπες εικόνες των πειραμάτων μας, αναδεικνύεται η πολύ σπουδαία ιδιότητα της Π-sigmoid κατανομής να περιγράφει ομοιόμορφα κατανεμημένα δεδομένα, πράγμα που την κάνει ευέλικτη και αποτελεσματική.

Στον παρακάτω πίνακα φαίνεται και η αρνητική λογαριθμική πιθανοφάνεια που παράχθηκε από το κάθε μοντέλο, σε κάθε εικόνα ξεχωριστά. Να σημειωθεί ότι εξαιτίας της δειγματοληψίας που έγινε στα pixels της κάθε εικόνας, μας δόθηκε η ευκαιρία να χρησιμοποιήσουμε τα εναπομείναντα pixels ως σύνολο ελέγχου. Έτσι έχουμε μια πλήρη εικόνα για την απόδοση των δύο κατανομών.

Πίνακας 4.24 Σύγκριση των μοντέλων GMM και ΠsMM σε gray-scale εικόνες, χρησιμοποιώντας ως μέτρο την αρνητική λογαριθμική πιθανοφάνεια

Image	**Clusters**	**N**	**GMM -Log-ikelihood**	**ΠsMM -Log-Likelihood**
amakses.jpg	4	TRAIN=5000 TEST=24400	6218.650587 30229.161367	6217.433933 30220.665190
woman.jpg	4	TRAIN=5000 TEST=24400	5978.541501 29027.006550	5935.415767 28788.607483
clouds.jpg	4	TRAIN=5000 TEST=24400	5972.983983 29169.309661	5941.711410 28983.716357

rocks-tree.jpg	5	TRAIN=5000 TEST=24400	6724.609454 32946.739065	6702.920664 32885.594371
elephants.jpg	4	TRAIN=5000 TEST=24400	5779.865100 28124.055712	5763.366375 27953.237875
rocks.jpg	4	TRAIN=5000 TEST=24400	5416.115651 26576.471606	4770.070877 23487.026916

Από τον προηγούμενο πίνακα βλέπουμε πάλι μια ελαφρά υπεροχή του ΠsMM έναντι του GMM αναφορικά με την αρνητική λογαριθμική πιθανοφάνεια.

Γενικότερα, παρατηρούμε ότι τα πειράματα, που σχετίζονται με την σύγκριση της αρνητικής λογαριθμικής πιθανοφάνειας, δείχνουν μια γενικότερη υπεροχή του μικτού μοντέλου Π-sigmoid κατανομών έναντι του GMM. Το γεγονός αυτό φανερώνει την γενικότερη ευελιξία του μοντέλου αυτού στην περιγραφή διαφόρων ειδών δεδομένων. Από τα αποτελέσματα των υπολοίπων πειραμάτων, παρατηρούμε επίσης την αξιοσημείωτη θετική απόδοση του ΠsMM, γεγονός που το αναδεικνύει ως ένα γενικό και αποδοτικό μοντέλο για την επίλυση ποικίλων προβλημάτων.

ΚΕΦΑΛΑΙΟ 5. ΕΠΙΛΟΓΟΣ-ΣΥΜΠΕΡΑΣΜΑΤΑ

5.1. Γενικά

Σε αυτή την εργασία προτάθηκε η νέα συνάρτηση πυκνότητας πιθανότητας Π-sigmoid, ως ένα εργαλείο διαχείρισης δεδομένων με άγνωστες στατιστικές ιδιότητες. Παρουσιάστηκε, η καλή ικανότητα μοντελοποίησης που παρέχει, η οποία εν γένει οφείλεται και στην δυνατότητα περιγραφής ομοιόμορφων δεδομένων. Στο κεφάλαιο 2 δείξαμε αναλυτικά της ιδιότητές της, καθώς και τη μορφή που αυτή παίρνει για τις διάφορες τιμές των παραμέτρων της. Έγινε αναφορά στην πολυδιάστατη εκδοχή της, όπως επίσης και στο μετασχηματισμό που πραγματοποιήθηκε για να μπορεί να περιγράφει περιστραμμένα δεδομένα. Από την σχετική παρουσίαση, με γνώμονα πάντα τις γραφικές παραστάσεις που παρατέθηκαν, έγινε εμφανής η ικανότητα της Π-sigmoid κατανομής να περιγράφει ομοιόμορφα δεδομένα. Το σχήμα της στην μία διάσταση, για λ>>1, παίρνει το σχήμα “Π” (με αφορμή αυτό, προκύπτει και το όνομα της), ενώ στις D διαστάσεις, η γραφική της παράσταση παίρνει το σχήμα ενός υπερ-ορθογωνίου.

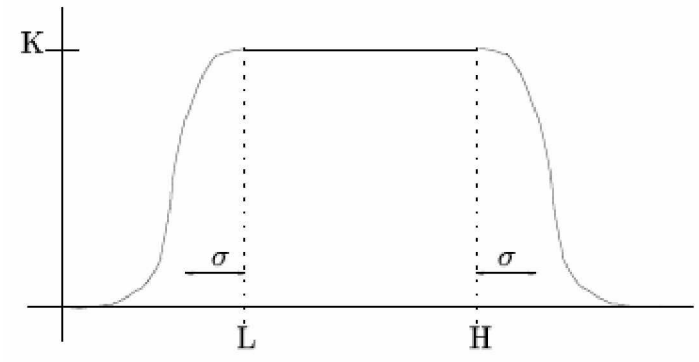
Στο κεφάλαιο 3 ορίσαμε το μικτό μοντέλο Π-sigmoid κατανομών(ΠsMM), και παρουσιάσαμε τον τρόπο εκπαίδευσης του, μέσω του αλγορίθμου GEM. Ειδικότερα, περιγράφτηκε ο τρόπος αρχικοποίησης των παραμέτρων του ΠsMM, μέσω αντιστοιχιών με τις παραμέτρους ενός εκπαιδευμένου GMM. Επίσης, δείξαμε τον τρόπο υλοποίησης του Μ-βήματος του αλγορίθμου GEM, με τη χρήση της αριθμητικής μεθόδου βελτιστοποίησης BFGS. Τέλος, έγινε αναφορά στην τεχνική

ρύθμισης των παραμέτρων W_k η οποία επιτεύχθηκε μέσω της ταυτόχρονης εκπαίδευσης ενός μικτού μοντέλου κανονικών κατανομών, από το οποίο εκμεταλλευτήκαμε τα ιδιοδιανύσματα $U_k=\{u_{k1},\ldots,u_{kD}\}$ των πινάκων συμμεταβλητότητας Σ_k, θεωρώντας τα ως μια καλή προσέγγιση για τις παραμέτρους W.

Στο κεφάλαιο 4 παρουσιάσαμε τα πειραματικά αποτελέσματα από την εφαρμογή ενός μικτού μοντέλου Π-sigmoid κατανομών (ΠsMM) σε διάφορους τύπους δεδομένων. Το συμπέρασμα που προέκυψε ήταν ότι το ΠsMM αποτελεί ένα πολύ καλό εργαλείο μοντελοποίησης δεδομένων, ισάξιο και σε πολλές περιπτώσεις πιο ευέλικτο και αποτελεσματικό από ένα αντίστοιχο μοντέλο κανονικών κατανομών.

5.2. Σχετική εργασία

Θα πρέπει να επισημάνουμε για το συγκεκριμένο ζήτημα το οποίο διαπραγματευτήκαμε δεν έχουν γίνει κάποιες αντίστοιχες ερευνητικές προσπάθειες και γι' αυτό η βιβλιογραφία είναι πενιχρή. Η μόνη σχετική δουλειά έγινε από τους D. Pelleg και A. Moore [1] οι οποίοι κατασκεύασαν μια κατανομή για την παραγωγή ερμηνεύσιμων κανόνων με ορθογώνιο σχήμα. Η κατανομή που προτείνουν συνίσταται από μια ομοιόμορφη κατανομή πλαισιωμένη εκατέρωθεν από γκαουσιανές ουρές με τυπική απόκλιση σ. Σχήμα 5.1.



Σχήμα 5.1 Η γραφική παράσταση της προτεινόμενης κατανομής από τους Moore και Pelleg [1]

Ουσιαστικά πρόκειται για την γενίκευση της κανονικής κατανομής, στην οποία διαφέρει ο τρόπος υπολογισμού της απόστασης των προτύπων από το κέντρο. Η συνάρτηση που ορίζει το σημείο που ανήκει στο διάστημα [L, H] (Σχήμα 5.1) και είναι το κοντινότερο στο πρότυπο x είναι:

$$closest(x, L, H) = \begin{cases} L, & if\ x < L \\ x, & if\ L \le x < H \\ H, & if\ x > H \end{cases} \tag{5.1}$$

Η διατύπωση της κατανομής είναι η εξής:

$$p(x) = K \exp\left(-\frac{1}{2} \sum_{d=1}^{D} \left(\frac{x_d - closest(x_d, L_d, H_d)}{\sigma_d} \right)^2 \right) \tag{5.2}$$

(όπου K η σταθερά κανονικοποίησης)

Εύκολα παρατηρούμε ότι η κατανομή από το ορισμό της είναι ασυνεχής και μη παραγωγίσιμη. Συνεπώς, η εκπαίδευση ενός μικτού μοντέλου από κατανομές της μορφής (5.2), δε μπορεί να γίνει με κάποια αποδοτική αριθμητική μέθοδο, που χρησιμοποιεί το gradient της λογαριθμικής πιθανοφάνειας. Επίσης, δεν υπάρχει η έννοια της περιστροφής στον ορισμό της. Τα δύο αυτά μειονεκτήματα, όπως έχουμε δείξει σε προηγούμενα κεφάλαια, δεν υφίστανται για την Π-sigmoid κατανομή.

5.3. Μελλοντική δουλειά

Ως κατευθύνσεις μελλοντικής έρευνας προτείνουμε τα ακόλουθα:

1. **Εφαρμογή του αλγορίθμου Greedy EM [2] για την εκπαίδευση ενός ΠsMM.** Όπως ειπώθηκε και στο κεφάλαιο 3 για να ξεκινήσει ο αλγόριθμος GEM τη διαδικασία βελτιστοποίησης της λογαριθμικής πιθανοφάνειας χρειάζεται να δώσουμε κάποιες αρχικές τιμές στις παραμέτρους του μικτού μοντέλου Π-sigmoid κατανομών. Αυτό που είχαμε υποδείξει ως λύση σε αυτό το ζήτημα, ήταν η εκπαίδευση ενός GMM με τη βοήθεια του αλγορίθμου

greedy EM και η ακόλουθη αρχικοποίηση των παραμέτρων του ΠsΜΜ μέσω κάποιων συσχετισμών με τις παραμέτρους του πρώτου. Το γεγονός αυτό όμως, περιορίζει σημαντικά τις δυνατότητες του ΠsΜΜ, αφού λαμβάνει μια αρχική λύση και καλείται απλά να τη βελτιώσει. Ιδανικά επιθυμούμε αυτή η αρχική λύση να παραχθεί, εκμεταλλευόμενοι τις ιδιότητες του ΠsΜΜ, το οποίο ενδεχομένως να δώσει μια αρκετά διαφορετική λύση.

2. **Ανίχνευση του βέλτιστου αριθμού συνιστωσών για το ΠsΜΜ.** Ένα άλλο πολύ σημαντικό ζήτημα, είναι ο προσδιορισμός του βέλτιστου αριθμού συνιστωσών για το μοντέλο ΠsΜΜ. Έχουν προταθεί διάφορες τεχνικές για την επίλυση αυτού του ζητήματος, αλλά αφορούν κυρίως τα μικτά μοντέλα κανονικών κατανομών. Στόχος μας είναι είτε να τροποποιήσουμε τις παρούσες τεχνικές είτε να δημιουργήσουμε μια νέα, που ανταποκρίνεται στις ιδιότητες του μικτού μοντέλου Π-sigmoid κατανομών

ΑΝΑΦΟΡΕΣ

[1] Pelleg D. and Moore A., Mixture of rectangles: Interpretablesoft clustering, Proc ICML 2001

[2] Vlassis, N. and Likas, A., A greedy EM algorithm for Gaussian mixture learning, Neural Processing Letters , vol. 15, pp. 77-87, 2002.

[3] C. Blake and Merz C. (1998). UCI repository of machine learning databases. http://www.ics.uci.edu/~mlearn/databases.

[4] G. J. McLachlan and T. Krishnan, Finite Mixture Models, Wiley, 2000.

[5] G. J. McLachlan and T. Krishnan, The EM algorithm and extensions, Marcel Dekker, 1997.

[6] C. M. Bishop, Pattern recognition and machine learning, Springer 2006.

[7] J. Nocedal and S. J. Wright, Numerical Optimization, Springer 1999.

[8] P. N. Tan, M. Steinbach and V. Kumar, Introduction to Data Mining, Pearson 2005.

[9] I.T. Jollife, Principal Component Analysis, Spinger, 2002

ΣΥΝΤΟΜΟ ΒΙΟΓΡΑΦΙΚΟ

Ο Αλιβάνογλου Αναστάσιος γεννήθηκε το 1982, στη Λάρισα. Μεγάλωσε στο δημοτικό διαμέρισμα Καππαδοκικού του δήμου Σοφάδων-Καρδίτσας και τελείωσε το Γενικό Λύκειο Σοφάδων με βαθμό 18,5. Το 2000 ξεκίνησε τις ανώτατες σπουδές του στο τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων, από όπου αποφοίτησε το Φεβρουάριο του 2005, με βαθμό 6,60. Κατά τη διάρκεια των προπτυχιακών σπουδών συμμετείχε στην υλοποίηση project έχοντας άρτια γνώση σε γλώσσες και περιβάλλοντα όπως C, C++, Java, PHP, HTML, OpenGL, SRGP, CORBA και Matlab. Για την λήψη του πτυχίου και κατά την διάρκεια του τέταρτου έτους σπουδών, εκπόνησε πτυχιακή εργασία με θέμα την “ομαδοποίηση ιστοσελίδων με βάση το περιεχόμενο” (web content mining). Από το Φεβρουάριο του 2005 είναι μεταπτυχιακός φοιτητής στο τμήμα Πληροφορικής του Πανεπιστημίου Ιωαννίνων και μέλος του εργαστηρίου IPAN (Image Processing and Analysis) του ιδίου τμήματος.