**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

# Μη-παραμετρικεσ Μπεϋζιανεσ Μεθοδοι

Διπλωματική Εργασία

της

Λυμπέρη Χρυσοβαλάντης

**Επιβλέπων:** Δημήτριος Φουσκάκης
Καθηγητής

Αθήνα, Σεπτέμβριος 2024

**ΕΘΝΙΚΟ ΜΕΤΣΟΒΙΟ ΠΟΛΥΤΕΧΝΕΙΟ**

ΣΧΟΛΗ ΕΦΑΡΜΟΣΜΕΝΩΝ ΜΑΘΗΜΑΤΙΚΩΝ ΚΑΙ ΦΥΣΙΚΩΝ ΕΠΙΣΤΗΜΩΝ

ΤΟΜΕΑΣ ΜΑΘΗΜΑΤΙΚΩΝ

# Μη-παραμετρικεσ Μπεϋζιανεσ Μεθοδοι

Διπλωματική Εργασία

της

Λυμπέρη Χρυσοβαλάντης

**Επιβλέπων:** Δημήτριος Φουσκάκης
Καθηγητής

**Τριμελής Επιτροπή**

| *(Υπογραφή)* | *(Υπογραφή)* | *(Υπογραφή)* |
|---|---|---|
| ..... | ..... | ..... |
| Δημήτριος Φουσκάκης | Ιωάννης Ντζούφρας | Μιχαήλ Λουλάκης |
| Καθηγητής, ΕΜΠ | Καθηγητής, ΟΠΑ | Καθηγητής, ΕΜΠ |

Αθήνα, Σεπτέμβριος 2024

# Abstract

This thesis explores the broad topic of Bayesian non-parametric methods, focusing on both their theoretical foundation and practical application. Initially, the reader is introduced to the Bayesian framework and fundamental concepts such as the definition of a Bayesian model, the prior and posterior distributions. Essential computational methods, including Markov Chain Monte Carlo (MCMC), the Metropolis-Hastings algorithm, and Gibbs Sampling, are also presented.

The Dirichlet Process is examined as a crucial component of Bayesian Nonparametrics. Beyond its definition and key properties, its connection to the well known clustering problem is highlighted. The analysis then extends to the Hierarchical Dirichlet Process and its construction using the Stick-Breaking method, analogous to the Dirichlet Process. Applications in Machine Learning, such as in Topic Modeling problems, are discussed, emphasizing their flexibility in modeling complex data structures. This thesis also includes a brief introduction to Dependent Dirichlet Processes, a useful tool for problems where data evolve over time or space.

Subsequently, the focus shifts to the Gaussian Process, a stochastic process that can be viewed as an infinite-dimensional distribution over functions. The application of Gaussian Processes to non-linear regression and classification problems is described, along with related model selection methods.

Throughout the thesis, visual aids and examples with simulated data are employed to make the concepts more accessible. Additionally, two applications on real-world datasets are included.

## Keywords

Non-parametric Bayesian methods, Dirichlet Process (DP), Hierarchical Dirichlet Process (HDP), Dependent Dirichlet Process (DDP), Markov Chain Monte Carlo (MCMC), Metropolis-Hastings algorithm, Gibbs sampling, Expectation-Maximization (EM) algorithm, clustering, topic modeling, Gaussian Processes (GP), non-linear regression, classification.

# Περίληψη

Η παρούσα διπλωματική εργασία εξετάζει το ευρύ αντικείμενο των μη-παραμετρικών Μπεϋζιανών μεθόδων, εστιάζοντας τόσο στην θεωρητική θεμελίωση όσο και στην πρακτική εφαρμογή τους. Αρχικά, ο αναγνώστης εισάγεται στο Μπεϋζιανό πλαίσιο αναφοράς και σε θεμελιώδεις σε αυτό έννοιες όπως ο ορισμός του Μπεϋζιανού μοντέλου, η πρότερη και η ύστερη κατανομή. Ακόμα παρουσιάζονται απαραίτητες υπολογιστικές μέθοδοι όπως το Markov Chain Monte Carlo (MCMC), ο αλγόριθμος Metropolis-Hastings και η Δειγματοληψία κατά Gibbs.

Η Διαδικασία Dirichlet παρουσιάζεται ως το επίκεντρο της μη-παραμετρικής Μπεϋζιανής στατιστικής. Εκτός από τον ορισμό της και τις κύριες ιδιότητες της, αναδυκνείεται ο τρόπος που αυτή συνδέεται με το γνωστό πρόβλημα της συσταδοποίησης. Η ανάλυση επεκτείνεται στην Ιεραρχική Διαδικασία Dirichlet και την κατασκευή της μέσω της μεθόδου Stick-Breaking, σε αναλογία με την Διαδικασία Dirichlet. Συζητούνται εφαρμογές αυτών των διαδικασιών στη Μηχανική Μάθηση, όπως σε προβλήματα Θεματικής Μοντελοποίησης, υπογραμμίζοντας την ευελιξία τους στη μοντελοποίηση σύνθετων δομών δεδομένων. Επίσης, περιλαμβάνεται μια σύντομη εισαγωγή στις Εξαρτημένες Διαδικασίες Dirichlet, οι οποίες είναι χρήσιμες σε προβλήματα όπου τα δεδομένα εξελίσσονται χρονικά ή χωρικά.

Έπειτα, η προσοχή στρέφεται στην Γκαουσιανή Διαδικασία, μια στοχαστική διαδικασία που μπορεί κανείς να φανταστεί ως μια απειροδιάστατη κατανομή πάνω σε συναρτήσεις. Περιγράφεται ο τρόπος που αυτή εφαρμόζεται σε προβλήματα μη-γραμμικής παλινδρόμησης και ταξινόμησης, ενώ γίνεται αναφορά και στις μεθόδους επιλογής μοντέλου.

Καθ'όλη τη διάρκεια της εργασίας επιστρατεύονται οπτικά μέσα και παραδείγματα σε προσομοιωμένα δεδομένα, με σκοπό να καταστήσουν πιο κατανοητές τις έννοιες που παρουσιάζονται. Ακόμα, περιλαμβάνονται δύο εφαρμογές σε πραγματικά σύνολα δεδομένων.

## Λέξεις-Κλειδιά

Μη-παραμετρικές Μπεϋζιανές Μέθοδοι, Διαδικασία Dirichlet (DP), Ιεραρχική Διαδικασία Dirichlet (HDP), Εξαρτημένη Διαδικασία Dirichlet (DDP), Markov Chain Monte Carlo (MCMC), Αλγόριθμος Metropolis-Hastings, Δειγματοληψία Gibbs, Expectation-Maximization (EM), συσταδοποίηση, θεματική μοντελοποίηση, Γκαουσιανές Διαδικασίες (GP), Μη-γραμμική Παλινδρόμηση, ταξινόμηση.

# Ευχαριστίες

Θα ήθελα να εκφράσω τις ειλικρινείς μου ευχαριστίες σε όλους όσους στάθηκαν δίπλα μου κατά την διάρκεια των προπτυχιακών μου σπουδών αλλά και της εκπόνησης της παρούσας εργασίας.

Πρώτα απ' όλα, ευχαριστώ θερμά τον καθηγητή μου, κ. Φουσκάκη, για την καθοδήγηση και την εμπιστοσύνη που μου έδειξε καθ' όλη τη διάρκεια της συγγραφής της εργασίας.

Επίσης, θα ήθελα να ευχαριστήσω τους φίλους μου για τη συνεχή υποστήριξη και την ενθάρρυνσή τους σε κάθε βήμα αυτής της πορείας. Η βαθιά κατανόηση που έδειξαν αποτέλεσαν σημαντική πηγή δύναμης και έμπνευσης.

Τέλος, ευχαριστώ από καρδιάς την οικογένειά μου: τους γονείς μου, Πετρούλα και Μάκη και τον αδερφό μου, Βασίλη που στάθηκαν δίπλα μου με υπομονή, αγάπη και αδιάλειπτη στήριξη.

Βαλεντίνα Λυμπέρη
Αθήνα, Σεπτέμβριος 2024

# Περιεχόμενα

# Συντομογραφίες

| | |
|---|---|
| τ.μ | τυχαία μεταβλητή |
| σ.μ.π | συνάρτηση μάζας πιθανότητας |
| σ.π.π | συνάρτηση πυκνότητας πιθανότητας |
| MCMC | Markov Chain Monte Carlo |
| EM | Expectation Maximization |
| DP | Dirichlet Process |
| DPMM | Dirichlet Process Mixture Model |
| HDP | Hierarchical Dirichlet Process |
| HMM | Hidden Markov Model |
| DDP | Dependent Dirichlet Process |
| GP | Gaussian Process |

# Κατάλογος Διαγραμμάτων

# Κεφάλαιο 1

# Το Μπεϋζιανό πλαίσιο αναφοράς

## 1.1 Ορισμός του Μπεϋζιανού Στατιστικού Μοντέλου

Σκοπός κάθε στατιστικής ανάλυσης είναι η εξαγωγή επιστημονικά βάσιμων συμπερασμάτων αναφορικά με τον υπό μελέτη πληθυσμό. Αυτό επιτυγχάνεται με τη μοντελοποίηση του εν λόγω πληθυσμού· η έννοια του στατιστικού μοντέλου είναι κεντρική για την ανάπτυξη οποιασδήποτε μεθόδου ανάλυσης δεδομένων, αφού καθιστά δυνατή την μαθηματική αναπαράσταση της υποκείμενης διαδικασίας που *γεννά* τα δεδομένα αυτά. Το στατιστικό μοντέλο περικλείει τις υποθέσεις και τις σχέσεις που συνδέουν τις υπό μελέτη μεταβλητές, παρέχοντας ένα δομημένο πλαίσιο για την ποσοτικοποίηση της αβεβαιότητας που τις διέπει.

Έστω $\boldsymbol{x} = (x_1, x_2, \ldots, x_n)$ παρατηρήσεις από δοθέν δείγμα $\boldsymbol{X} = (X_1, X_2, \ldots, X_n)$. Υποθέτουμε ότι το δείγμα είναι τυχαίο, δηλαδή οι $X_i, i = 1, \ldots, n$ είναι ανεξάρτητες και ισόνομες τυχαίες μεταβλητές προερχόμενες από τον υπό μελέτη πληθυσμό. Συμβολίζουμε με $p(\boldsymbol{x}|\boldsymbol{\theta})$ την συνάρτηση μάζας πιθανότητας (ή συνάρτηση πυκνότητας πιθανότητας) του χαρακτηριστικού $\boldsymbol{X}$. Το πρόβλημα της Εκτιμητικής περιλαμβάνει τον προσδιορισμό, κατα βέλτιστο τρόπο, της άγνωστης παραμέτρου $\boldsymbol{\theta}$. Περιγράφουμε τον πληθυσμό με το μοντέλο

$$\{\boldsymbol{X}, \mathcal{X}^n, p(\boldsymbol{x}|\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$$

όπου

- $\boldsymbol{X}$: μια τυχαία μεταβλητή που συμβολίζει το χαρακτηρηστικό στο δείγμα που μελετάμε,
- $\mathcal{X}^n \subseteq \mathbb{R}^n$: το πεδίο ορισμού του $\boldsymbol{X}$,
- $p(\boldsymbol{x}|\boldsymbol{\theta})$: η σ.μ.π ή σ.π.π του χαρακτηριστικού $\boldsymbol{X}$ και
- $\boldsymbol{\theta} = (\theta_1, \theta_2, \ldots, \theta_m)^T \in \Theta \subseteq \mathbb{R}^m$: το διάνυσμα των παραμέτρων.

Στην κλασική στατιστική, η παράμετρος $\boldsymbol{\theta}$ αντιμετωπίζεται ως άγνωστη σταθερά, για την εκτίμηση της οποίας χρησιμοποιούνται τα διαθέσιμα δεδομένα (παρατηρήσεις). Στο Μπεϋζιανό πλαίσιο, η παράμετρος $\boldsymbol{\theta}$ θεωρείται ως μία τυχαία μεταβλητή και ο βαθμός της αβεβαιότητας της μοντελοποιείται με χρήση πιθανοτήτων.

Ο πλήρης καθορισμός ενός Μπεϋζιανού παραμετρικού μοντέλου, απαιτεί γνώση δύο κατανομών: της δεσμευμένης κατανομής $p(\boldsymbol{X}|\boldsymbol{\theta})$ καθώς και της κατανομής της $\boldsymbol{\theta}, p(\boldsymbol{\theta})$. Παρατηρούμε ότι η πρώτη κατανομή ταυτίζεται με την συνάρτηση πιθανοφάνειας, αν αυτή θεωρηθεί συνάρτηση στον χώρο της παραμέτρου και όχι συνάρτηση της τυχαίας μεταβλητής $\boldsymbol{X}$. Στη συνέχεια θα "συνδέσουμε" τις δύο αυτές κατανομές, κάνοντας χρήση του Κανόνα του Bayes.

**Κανόνας του Bayes**: Έστω δύο ενδεχόμενα $A, B$ με $\mathbb{P}(B) \neq 0$. Τότε,

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

ενώ κάνοντας χρήση του Θεωρήματος Ολικής Πιθανότητας για την $\mathbb{P}(B)$ λαμβάνουμε

$$\mathbb{P}(A|B) = \frac{\mathbb{P}(B|A)\mathbb{P}(A)}{\mathbb{P}(B|A)\mathbb{P}(A) + \mathbb{P}(B|A^c)\mathbb{P}(A^c)}.$$

Εφαρμόζοντας τον Κανόνα του Bayes για την κατανομή της παραμέτρου $\boldsymbol{\theta}$ δεδομένου του $\mathbf{X}$, δηλαδή την $p(\boldsymbol{\theta}|\boldsymbol{x})$, έχουμε

$$p(\boldsymbol{\theta}|\boldsymbol{X}) = kp(\boldsymbol{\theta})p(\boldsymbol{X}|\boldsymbol{\theta}),$$

όπου

$$k^{-1} = p(\boldsymbol{x})$$

δηλαδή

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta}). \tag{1.1.1}$$

Η $p(\boldsymbol{\theta}|\boldsymbol{x})$ καλείται εκ των υστέρων (*posterior*) κατανομή της $\boldsymbol{\theta}$ δεδομένου του $\boldsymbol{X}$, αφού τα συμπεράσματα σχετικά με την $\boldsymbol{\theta}$ εξάγονται *μετά* την λήψη του τυχαίου δείγματος. Η $p(\boldsymbol{\theta})$ ονομάζεται εκ των προτέρων (*prior*) κατανομή και εκφράζει την υποκειμενική γνώση μας σχετικά με την άγνωστη παράμετρο.

Η (1.1.1) υποδηλώνει μια μορφή επαναληψιμότητας που καθίσταται δυνατή όταν χρησιμοποιείται η Μπεϋζιανή Στατιστική και που συνάδει με την διαίσθηση μας: Ανανεώνουμε την γνώση μας για την άγνωστη παράμετρο $\boldsymbol{\theta}$ συνδυάζοντας την πρότερη, εμπειρική γνώση μας και την γνώση που μας παρέχει το δείγμα. Έτσι, πριν από την τυχαία δειγματοληψία έχουμε το αρχικό, *a priori* μοντέλο

$$\{\boldsymbol{X}, \mathcal{X}^n, p(\boldsymbol{x}|\boldsymbol{\theta}), p(\boldsymbol{\theta}), \boldsymbol{\theta} \in \Theta\}$$

ενώ τελικά, λαμβάνουμε το ύστερο, *a posteriori* μοντέλο

$$\{\boldsymbol{X}, \mathcal{X}^n, p(\boldsymbol{x}|\boldsymbol{\theta}), p(\boldsymbol{\theta}|\boldsymbol{x}), \boldsymbol{\theta} \in \Theta\}$$

το οποίο δύναται να χρησιμοποιηθεί εκ νέου ως αρχικό μοντέλο μιας νέας δειγματοληψίας.

## 1.2 Καθορισμός της εκ των προτέρων κατανομής

Ένα βασικό ερώτημα που ανακύπτει ως φυσική απόρροια των παραπάνω, είναι το πώς ακριβώς θα γίνει ο καθορισμός της εκ των προτέρων κατανομής. Συνήθως, η επιλογή της αντανακλά την προσωπική μας πεποίθηση σχετικά με τα δεδομένα, πριν αυτά καταστούν διαθέσιμα, ή μία πρότερη γνώση που έχουμε για την παράμετρο. Όταν υπάρχουν διαθέσιμες πληροφορίες σχετικά με την κατανομή της παραμέτρου, είτε από προηγούμενες παρατηρήσεις είτε μέσω της τοποθέτησης ενός ειδικού επί του προβλήματος, κάνουμε λόγο για *πληροφοριακή* εκ των προτέρων κατανομή. Στην περίπτωση που τέτοιου είδους πληροφορίες δεν είναι διαθέσιμες, λόγω της φύσης του πειράματος, η εκ των προτέρων κατανομή ονομάζεται μη-πληροφοριακή.

### 1.2.1 Συζυγείς κατανομές

Η εφαρμογή του κανόνα του Bayes οδηγεί συχνά σε δύσκολα στον υπολογισμό ολοκληρώματα. Προκειμένου να αποφευχθεί αυτό, προτιμούνται συνήθως οι *συζυγείς* εκ των προτέρων κατανομές που διευκολύνουν τον υπολογισμό της εκ των υστέρων κατανομής, αφού μετά την χρήση τους, δεν είναι αναγκαίος ο υπολογισμός της συνάρτησης πιθανοφάνειας (του παρονομαστή).

**Ορισμός 1.2.1** (Συζυγής Κατανομή)**.** Έστω οικογένεια $\mathcal{F}$ κατανομών $p(\boldsymbol{x}|\boldsymbol{\theta})$ και $\mathcal{P}$ η οικογένεια των εκ των προτέρων κατανομών της $\boldsymbol{\theta}$. Η $\mathcal{P}$ είναι συζυγής οικογένεια κατανομών για την $\mathcal{F}$ αν

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \in \mathcal{P} \ \forall \ p(\cdot|\boldsymbol{\theta}) \in \mathcal{F} \text{ και } p(\cdot) \in \mathcal{P}.$$

Δηλαδή, μία οικογένεια εκ των προτέρων κατανομών ονομάζεται συζυγής, αν η εκ των υστέρων κατανομή ανήκει στην ίδια οικογένεια με αυτήν.

Μάλιστα, στην περίπτωση που η $\mathcal{F}$ είναι η Εκθετική Οικογένεια Κατανομών (ΕΟΚ), αποδεικνύεται ότι υπάρχει πάντα συζυγής κατανομή για κάθε $p(\boldsymbol{x}|\boldsymbol{\theta}) \in \mathcal{F}$. Μια κατανομή $p(\boldsymbol{x}|\boldsymbol{\theta})$ ανήκει στην ΕΟΚ αν έχει τη μορφή

$$p(\boldsymbol{x}|\boldsymbol{\theta}) = c(\boldsymbol{\theta})h(\boldsymbol{x}) \exp\{\varphi(\boldsymbol{\theta})^T s(\boldsymbol{x})\}.$$

Τότε, αν η πρότερη είναι της μορφής

$$p(\boldsymbol{\theta}) \propto c(\boldsymbol{\theta})^\eta \exp\left\{\varphi(\boldsymbol{\theta})^T \boldsymbol{v}\right\},$$

η εκ των υστέρων κατανομή έχει τη μορφή

$$p(\boldsymbol{\theta}|\boldsymbol{x}) \propto c(\boldsymbol{\theta})^{\eta+n} \exp\left\{\varphi(\boldsymbol{\theta})^T (\boldsymbol{v} + t(\boldsymbol{x}))\right\},$$

όπου $t(\boldsymbol{x}) = \sum_{i=1}^n s(x_i)$ το επαρκές στατιστικο για την $\boldsymbol{\theta}$.

Παραδείγματα συνεχών κατανομών που ανήκουν στην ΕΟΚ είναι η Εκθετική, η Κανονική, η Βήτα και η Γάμμα από τις κατανομές μίας μεταβλητής και η πολυπαραμετρική Κανονική και η Dirichlet από τις πολυμεταβλητές.

**Παράδειγμα 1.2.1.** *Η Βήτα κατανομή είναι συζυγής της Διωνυμικής κατανομής.*

*Απόδειξη*:

Έστω $p(x|\theta) \sim Bin(n, \theta)$ και $p(\theta) \sim Beta(a, b)$. Τότε έχουμε,

$$\begin{aligned} p(\theta|x) &\propto p(x|\theta)p(\theta) \\ \Rightarrow p(\theta|x) &\propto \theta^x(1-\theta)^{n-x}\frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ \Rightarrow p(\theta|x) &\propto \theta^x(1-\theta)^{n-x}\theta^{\alpha-1}(1-\theta)^{\beta-1} \\ \Rightarrow p(\theta|x) &\propto \theta^{x+\alpha-1}(1-\theta)^{n-x+\beta-1} \end{aligned}$$

Δηλαδή, $p(\theta|x) \sim Beta(x+\alpha, n-x+\beta)$. ■

Οι συζυγείς εκ των προτέρων κατανομές, δεν μειώνουν μόνο την υπολογιστική πολυπλοκότητα του προβλήματος εύρεσης της εκ των υστέρων κατανομής, αλλά έχουν και ένα επιπρόσθετο πρακτικό πλεονέκτημα: Μπορούν να ερμηνευτούν κατά περίπτωση ως επιπλέον δεδομένα (Gelman et al. 2013). Στο παράδειγμα (1.2.1) π.χ, η χρήση της $Beta(\alpha, \beta)$ ως πρότερη κατανομή, μπορεί να ερμηνευτεί ως "προσθήκη" $\alpha - 1$ επιτυχιών και $\beta - 1$ αποτυχιών στο σύνολο δεδομένων.

### 1.2.2 Μη-γνήσιες εκ των προτέρων κατανομές και η κατανομή του Jeffreys

Σε πολλές περιπτώσεις υπάρχει η ανάγκη για επιλογή πρότερων κατανομών οι οποίες δεν επηρεάζουν σε μεγάλο βαθμό την εκ των υστέρων κατανομή, που εκφράζουν δηλαδή την άγνοια μας σχετικά με τις τιμές της $\boldsymbol{\theta}$. Τότε, κάνουμε λόγο για μη-πληροφοριακές πρότερες κατανομές, η οποίες συχνά αποκαλούνται και επίπεδες ή διάχυτες.

Η πρώτη πρόταση μη-πληροφοριακής κατανομής ήταν η ομοιόμορφη, η οποία αποδίδει ίση πιθανότητα σε όλες τις πιθανές τιμές της παραμέτρου $\boldsymbol{\theta}$. Ωστόσο, ένα πρόβλημα που συνοδεύει τη χρήση της ομοιόμορφης ως μη-πληροφοριακής πρότερης είναι ότι δεν είναι αναλλοίωτη ως προς τους διάφορους μετασχηματισμούς της $\boldsymbol{\theta}$. Για παράδειγμα, ας θεωρήσουμε ότι δεν διαθέτουμε πληροφορίες για την $\theta \in \mathbb{R}$ και ας επιλέξουμε για αυτήν πρότερη

$$\theta \sim U[0, 1].$$

Η άγνοια μας για την $\theta$ συνεπάγεται άγνοια και για οποιονδήποτε μετασχηματισμό της, π.χ για την $\theta^2$. Ωστόσο, δεν ισχύει ότι

$$\theta^2 \sim U[0, 1],$$

με αποτέλεσμα η χρήση της ομοιόμορφης πρότερης να μην είναι πλέον μη-πληροφοριακή.

Ακόμα, σε μη-συμπαγείς παραμετρικούς χώρους (π.χ $\theta \in \mathbb{R}$), δε μπορεί να οριστεί ομοιόμορφη κατανομή της μορφής

$$p(\theta) \propto c,$$

αφού δεν ολοκληρώνει στη μονάδα. Ωστόσο η χρήση τους είναι ακόμα επιτρεπτή, αν και μόνο αν

$$\int f(\boldsymbol{x}|\theta)d\theta = K < \infty,$$

αφού τότε η εκ των προτέρων κατανομή ορίζεται κανονικά:

$$p(\theta|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\theta)p(\theta)}{\int f(\boldsymbol{x}|\theta)p(\theta)d\theta} = \frac{f(\boldsymbol{x}|\theta)c}{\int f(\boldsymbol{x}|\theta)cd\theta} = \frac{f(\boldsymbol{x}|\theta)}{\int f(\boldsymbol{x}|\theta)d\theta}.$$

Τέτοιου είδους εκ των προτέρων κατανομές καλούνται μη-γνήσιες (*improper*) και ο καθορισμός τους γίνεται συνήθως με τη μέθοδο του Jeffreys (Jeffreys 1946): Μέσω ενός 1-1 μετασχηματισμού της παραμέτρου $\theta$,

$$\varphi = h(\theta),$$

η $p(\varphi)$ γράφεται σύμφωνα με τον τύπο αλλαγής μεταβλητών ως

$$p(\phi) = p(\theta)\left|\frac{d\theta}{d\phi}\right| = p(\theta)|h'(\theta)|^{-1}.$$

Σύμφωνα με τη μέθοδο του Jeffreys, η εκ των προτέρων κατανομή επιλέγεται ώστε

$$p_0(\theta) \propto [I(\theta)]^{\frac{1}{2}}, \tag{1.2.1}$$

όπου $I(\theta)$ η αναμενόμενη πληροφορία κατά Fisher που ορίζεται ως

$$I(\theta) = -\mathbb{E}_{\boldsymbol{X}|\theta}\left[\frac{\partial^2}{\partial\theta^2}\log f(\boldsymbol{x}|\theta)\right].$$

Η εκ των προτέρων κατανομή που ορίζεται σύμφωνα με την (1.2.1) είναι αναλλοίωτη ως προς 1-1 μετασχηματισμούς της $\theta$, και άρα αντανακλά τον ίδιο βαθμό "γνώσης" για την άγνωστη παράμετρο, όποια (επιτρεπτή) παραμετροποίηση και αν χρησιμοποιηθεί, ενώ γενικεύεται με φυσικό τρόπο στην περίπτωση πολυδιάστατης παραμέτρου $\boldsymbol{\theta}$.

## 1.3 Προβλεπτικές Κατανομές

### 1.3.1 Εκ των προτέρων προβλεπτική κατανομή

Προτού παρατηρηθούν οι τιμές του δείγματος, η κατανομή των άγνωστων παρατηρήσεων δίνεται από την

$$p(\boldsymbol{x}) = \int p(\boldsymbol{x},\boldsymbol{\theta})d\boldsymbol{\theta} = \int p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{1.3.1}$$

Η 1.3.1 είναι η περιθώρια κατανομή της $\boldsymbol{X}$, που καλείται εναλλακτικά *εκ των προτέρων προβλεπτική* κατανομή. Η ονομασία αυτή οφείλεται στο γεγονός ότι δεν είναι δεσμευμένη ως προς κάποια άλλη παρατήρηση (*εκ των προτέρων*) και ότι είναι η κατανομή μιας μετρήσιμης ποσότητας (*προβλεπτική*).

### 1.3.2 Εκ των υστέρων προβλεπτική κατανομή

Αφού παρατηρηθεί το δείγμα $\boldsymbol{X}$, μπορούμε να προβλέψουμε μία άγνωστη τιμή $\tilde{\boldsymbol{x}}$ ακολουθώντας την ίδια διαδικασία. Η εκ των υστέρων προβλεπτική κατανομή γράφεται ως

$$p(\tilde{\boldsymbol{x}}) = \int p(\tilde{\boldsymbol{x}},\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} \tag{1.3.2}$$

$$= \int p(\tilde{\boldsymbol{x}}|\boldsymbol{\theta},\boldsymbol{x})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} \tag{1.3.3}$$

$$\overset{\tilde{\boldsymbol{x}}\text{ ανεξάρτητο του }\boldsymbol{x}}{=} \int p(\tilde{\boldsymbol{x}}|\boldsymbol{\theta},\boldsymbol{x})p(\boldsymbol{\theta}|\boldsymbol{x})d\boldsymbol{\theta} \tag{1.3.4}$$

Από την 1.3.4 γίνεται φανερό ότι η ύστερη προβλεπτική κατανομή για το $\tilde{\boldsymbol{X}}$ είναι η μέση τιμή των δεσμευμένων ως προς το δείγμα $\boldsymbol{X}$, πάνω στον παραμετρικό χώρο των παραμέτρων $\boldsymbol{\theta}$.

Τα διαστήματα πρόβλεψης που προκύπτουν με βάση τον παραπάνω υπολογισμό είναι πλατύτερα από τα διαστήματα που υπολογίζονται κλασικά, αφού λαμβάνουν υπόψιν και την αβεβαιότητα των παραμέτρων $\boldsymbol{\theta}$.

## 1.4 Markov Chain Monte Carlo (MCMC)

Στην προηγούμενη ενότητα αναφερθήκαμε στην χρήση συζυγών κατανομών, ως μέθοδο απλούστευσης της διαδικασίας υπολογισμού της εκ των υστέρων κατανομής. Η τακτική αυτή αν και βοηθητική, δε μπορεί να εφαρμοστεί πάντα, ειδικά όταν εργαζόμαστε σε παραμετρικούς χώρους υψηλής διάστασης. Για αυτό, σε τέτοιες περιπτώσεις, είναι αναγκαία η επιστράτευση υπολογιστικών μεθόδων, όπως αυτών της οικογένειας Markov Chain Monte Carlo.

Από την (1.1.1), έχουμε

$$p(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{p(\boldsymbol{\theta})p(\boldsymbol{x}|\boldsymbol{\theta})}{p(\boldsymbol{x})},$$

όπου

$$p(\boldsymbol{x}) = \int_{\boldsymbol{\theta}\in\Theta} p(\boldsymbol{x}|\boldsymbol{\theta})p(\boldsymbol{\theta})d\boldsymbol{\theta}. \tag{1.4.1}$$

Το ολοκλήρωμα στην (1.4.1) δεν έχει κλειστό τύπο και για αυτό ενδιαφερόμαστε για μία αριθμητική προσέγγιση της εκ των υστέρων κατανομής.

### 1.4.1 Η μέθοδος Monte Carlo

Η κλασική μέθοδος Monte Carlo, που προτάθηκε στα μέσα του εικοστού αιώνα (Metropolis and Ulam 1949), στηρίζεται στον Νόμο των Μεγάλων Αριθμών και συναντάται σε πλήθος εφαρμογών, τόσο στα Μαθηματικά όσο και στις Φυσικές Επιστήμες γενικότερα.

**Θεώρημα 1.4.1** ((Ισχυρός) Νόμος των μεγάλων Αριθμών)**.** Έστω $\{X_n\}_{n\in\mathbb{N}}$ μια ακολουθία από ανεξάρτητες, ισόνομες τυχαίες μεταβλητές με κατανομή $\pi$ με μέση τιμή $\mathbb{E}[X] < \infty$. Τότε,

$$\mathbb{P}\left[\frac{1}{n}\sum_{i=1}^{n} f(X_i) \to \mathbb{E}^{\pi}[f(X)]\right] = 1.$$

Η πρακτική αξία της μεθόδου Monte Carlo είναι η εξής: Αν μπορούμε να "φανταστούμε" την προς εκτίμηση ποσότητα ως αναμενόμενη τιμή μιας (συνάρτησης) τυχαίας μεταβλητής, μπορούμε να την προσέγγισουμε με τον δειγματικό μέσο ενός μεγάλου αριθμού τυχαίων δειγμάτων της μεταβλητής αυτής.

Η εφαρμογή της κλασικής μεθόδου Monte Carlo έγκειται στην προσομοίωση ενός δείγματος από την κατανομή $\pi$, κάτι που αποδεικνύεται σε πολλές περιπτώσεις εξαιρετικά δύσκολο, είτε λόγω της πολυπλοκότητας του χώρου της τυχαίας μεταβλητής, είτε γιατί η κατανομή $\pi$ δεν είναι γνωστή. Το πρόβλημα αυτό μπορεί να επιλύσει η χρήση Μαρκοβιανών αλυσίδων κατά την κατασκευή των μεθόδων Monte Carlo, οδηγώντας σε μια νέα οικογένεια μεθόδων, των Markov Chain Monte Carlo (MCMC). Το πλεονέκτημα των μεθόδων αυτών απορρέει από την Μαρκοβιανή ιδιότητα: η δεσμευμένη κατανομή της $X_{n+1}$ εξαρτάται αποκλειστικά από το προηγούμενο βήμα της αλυσίδας, $X_n$. Αρκεί λοιπόν να κατασκευάσουμε τη Μαρκοβιανή αλυσίδα, να ορίσουμε δηλαδή τις πιθανότητες μετάβασης, με τέτοιο τρόπο ώστε να εξασφαλίζεται ότι η $\pi$ είναι η κατανομή ισορροπίας της αλυσίδας. Ένας τέτοιος τρόπος περιγράφεται από τον αλγόριθμο Metropolis-Hastings (Metropolis et al. 1953; Hastings 1970).

### 1.4.2 Ο αλγόριθμος Metropolis-Hastings

Ο αλγόριθμος Metropolis-Hastings ονομάστηκε εν μέρει προς τιμήν του Hastings και αποτελεί γενίκευση της προγενέστερης μεθόδου του Metropolis. Αρχικό βήμα του αλγορίθμου είναι η επιλογή μιας προτεινόμενης κατανομής (*κατανομή* εισήγησης), $g$, από την οποία προσομοιώνουμε στο βήμα $t+1$ την προτεινόμενη τιμή $\boldsymbol{\theta}_{t+1}$. Δεχόμαστε την τιμή αυτή ως την τρέχουσα κατάσταση της αλυσίδας, αν ο $u \sim U(0,1)$ που επιλέξουμε τυχαία είναι μικρότερος από την πιθανότητα αποδοχής

$$\alpha = \min\left\{1, \frac{p(\boldsymbol{\theta}^*|\boldsymbol{x})g(\boldsymbol{\theta}_t|\boldsymbol{\theta}^*, \boldsymbol{x})}{p(\boldsymbol{\theta}_t|\boldsymbol{x})g(\boldsymbol{\theta}^*|\boldsymbol{\theta}_t, \boldsymbol{x})}\right\}. \tag{1.4.2}$$

Αποδεικνύεται ότι η κατανομή ισορροπίας της αλυσίδας $\{\boldsymbol{\theta}\}_t$ είναι η εκ των υστέρων κατανομή, $p(\boldsymbol{\theta}|\boldsymbol{x})$.

**Αλγόριθμος 1** Αλγόριθμος Metropolis-Hastings

Επιλέγουμε $\boldsymbol{\theta}^{(0)}$ αυθαίρετα
**για** $t = 1, 2, \ldots, T$
  "Προσομοιώνουμε" $\theta^*$ από την προτεινόμενη κατανομή $g(\boldsymbol{\theta}|\boldsymbol{\theta}^{t-1}, \boldsymbol{x})$
  Υπολογίζουμε τον λόγο αποδοχής $\alpha$
  Παράγουμε τυχαίο $u \sim \text{Uniform}(0,1)$
  **αν** $u \leq \alpha$ **τότε**
    Δεχόμαστε την πρόταση: $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$
  **αλλιώς**
    Απορρίπτουμε την πρόταση: $\boldsymbol{\theta}^{(t)} = \boldsymbol{\theta}^*$
  **τέλος αν**
**τέλος για**

Την παραπάνω διαδικασία την επαναλαμβάνουμε μέχρι να διακρίνουμε ότι επιτυγχάνεται σύγκλιση. Είναι επίσης συνήθης τακτική να απορρίπτεται ένα πλήθος αρχικών δειγμάτων, ώστε να βεβαιωθούμε ότι το δείγμα είναι πράγματι αντιπροσωπευτικό της κατανομής ισορροπίας[1].

Από το δείγμα MCMC που θα παραχθεί με χρήση του αλγορίθμου Metropolis-Hastings, μπορούμε να υπολογίσουμε ενδιαφέροντα στατιστικά σχετικά με την ύστερη κατανομή της παραμέτρου $\boldsymbol{\theta}$, όπως μέση τιμή,

[1] Το διάστημα μέχρι να συμβεί αυτό καλείται burn-in period.

διάμεσο, τυπική απόκλιση κ.α. Στο παράδειγμα 1.4.1 χρησιμοποιούμε τον αλγόριθμο Metropolis-Hastings για να προσομοιώσουμε δείγματα από την εκ των υστέρων κατανομή των παραμέτρων ενός Μπεϋζιανού γραμμικού μοντέλου.

**Παράδειγμα 1.4.1.** *Μπεϋζιανή Γραμμική Παλινδρόμηση με τη βοήθεια του αλγορίθμου Metropolis-Hastings*

Το μοντέλο της γραμμικής παλινδρόμησης είναι από τα πιο ευρέως διαδεδομένα στατιστικά εργαλεία που χρησιμοποιείται για την περιγραφή του τρόπου που μια ποσότητα $Y$ (μεταβλητή απόκρισης) εξαρτάται από μία ή περισσότερες επεξηγηματικές μεταβλητές. Στην πιο απλή του μορφή, ισχύει

$$\boldsymbol{y}|\boldsymbol{X} \sim N_n(\boldsymbol{X}\boldsymbol{\beta}, \sigma^2\boldsymbol{I})$$

όπου $\boldsymbol{y} = (y_1, \ldots, y_n)$ το διάνυσμα των παρατηρήσεων, $\boldsymbol{\beta} = (\beta_0, \ldots, \beta_p)$ το διάνυσμα των παραμέτρων του μοντέλου, $\boldsymbol{X}$ ο πίνακας σχεδιασμού και $\sigma^2\boldsymbol{I}$ ο πίνακας διασποράς της τ.μ $\boldsymbol{Y}$.

Όταν προσεγγίζουμε το πρόβλημα της γραμμικής παλινδρόμησης από την Μπευζιανή σκοπιά, δεν μας ενδιαφέρει να βρούμε την "βέλτιστη τιμή" της $\boldsymbol{\beta}$ (π.χ. με εκτιμήτριες ελαχίστων τετραγώνων), αλλά να καθορίσουμε την εκ των υστέρων κατανομή της. Δηλαδή, θεωρούμε ότι όχι μόνο η μεταβλητή απόκρισης, αλλά και η παράμετρος προέρχεται από κάποια κατανομή την οποία και θέλουμε να προσδιορίσουμε. Η τυπική απόκλιση $\sigma$ θεωρείται επίσης ως παράμετρος του μοντέλου που επιθυμούμε να μοντελοποιήσουμε, δηλαδή $\boldsymbol{\theta} = (\beta_0, \beta_1, \ldots, \beta_p, \sigma)$. Πιο συγκεκριμένα, η εκ των υστέρων κατανομή των παραμέτρων του μοντέλου είναι η [2]

$$p(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X}) = \frac{p(\boldsymbol{y}|\boldsymbol{X}, \boldsymbol{\theta})p(\boldsymbol{\theta}|\boldsymbol{X})}{p(\boldsymbol{y}|\boldsymbol{X})}.$$

Ως εκ των προτέρων κατανομή των παραμέτρων συνηθίζεται να επιλέγεται για τις $\{\beta_i\}_{i=0,\ldots,p}$ η κανονική κατανομή ενώ για την *ακρίβεια* $\tau = \sigma^{-2}$, η Γάμμα. Για τις παραμέτρους των εκ των προτέρων κατανομών μπορούμε να πάρουμε την άποψη ενός ειδικού πάνω στο πρόβλημα που εξετάζουμε. Η χρήση της Μπεϋζιανής προσέγγισης στην γραμμική παλινδρόμηση, λοιπόν, προσφέρει την δυνατότητα ενσωμάτωσης πρότερης γνώσης για τις τιμές των ζητούμενων παραμέτρων, κάτι που δεν είναι εφικτό με τις κλασικές μεθόδους εκτιμητικής.

Ας προχωρήσουμε τώρα σε ένα αριθμητικό παράδειγμα. Προσομοιώνουμε συνθετικά δεδομένα από ένα απλό γραμμικό μοντέλο $y = \alpha x + \beta$, $x \sim N(0, 1)$ με $\alpha = 3$, $\beta = 2$ στα οποία προσθέτουμε ένα "θόρυβο" $\varepsilon \sim N(0, \sigma^2)$, $\sigma = 1$. Ως εκ των προτέρων κατανομές θα επιλέξουμε τις εξής:

$$\alpha \sim N(0, 10)$$
$$\beta \sim N(0, 10)$$
$$\sigma^2 \sim InvGamma(2, 2).$$

Ως προτεινόμενη κατανομή για τον αλγόριθμο Metropolis-Hastings, επιλέγουμε την κανονική, και για τις τρεις παραμέτρους: Σε κάθε βήμα, "μετακινούμε" την αλυσίδα για κάθε παράμετρο κατά μία τυχαία μεταβλητή κανονικά κατανεμημένη, με μέση τιμή 0 και τυπική απόκλιση 0.1. Η κατανομή αυτή είναι συμμετρική, δηλαδή $g(x|y) = g(y|x)$ επομένως οι σχετικοί όροι στην 1.4.2 απλοποιούνται. Η πιθανότητα αποδοχής λοιπόν, υπολογίζεται στην προκειμένη περίπτωση σε κάθε βήμα ως

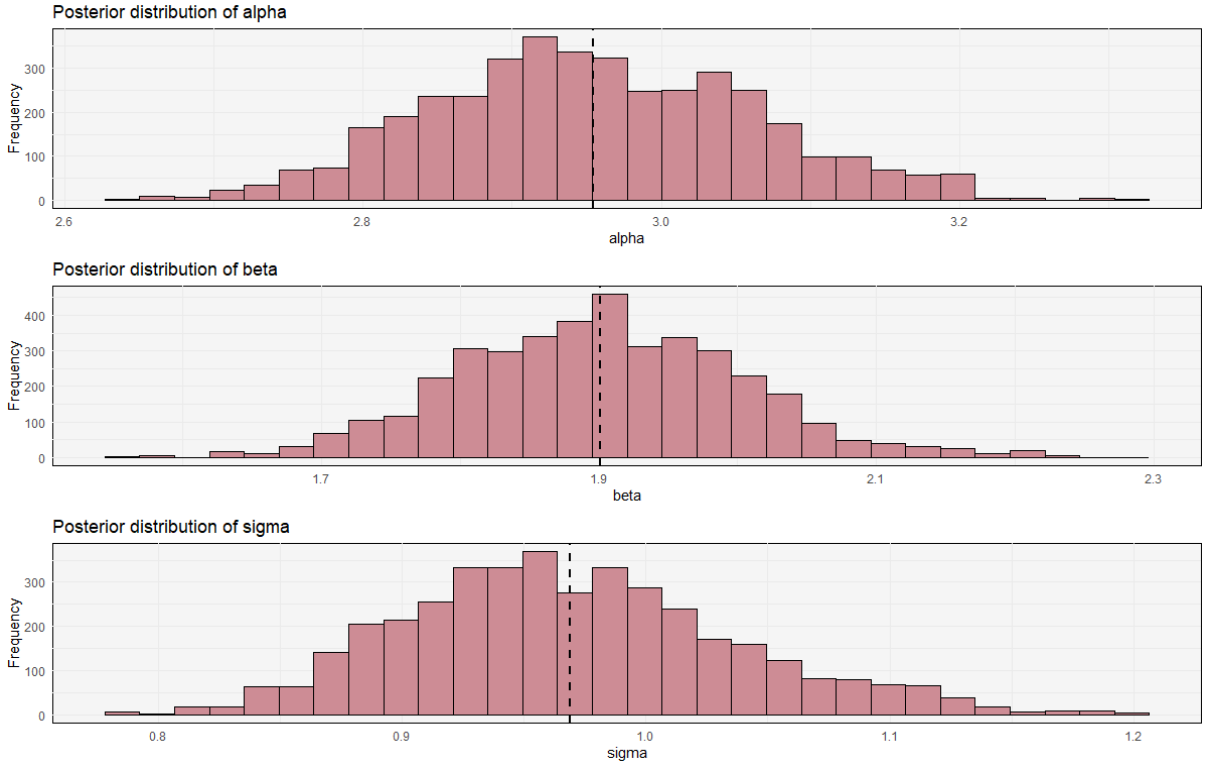
$$a = \min\left(1, \frac{p(\alpha^i, \beta^i, \sigma^i|\boldsymbol{x})}{p(a^{i-1}, \beta^{i-1}, \sigma^{i-1}|\boldsymbol{x})}\right)$$
$$\Rightarrow a = \min\left(1, \frac{p(\boldsymbol{x}|\alpha^i, \beta^i, \sigma^i)p(\alpha^i, \beta^i, \sigma^i)}{p(\boldsymbol{x}|a^{i-1}, \beta^{i-1}, \sigma^{i-1})p(a^{i-1}, \beta^{i-1}, \sigma^{i-1})}.\right)$$

Στο Διάγραμμα 1.1 παρουσιάζονται τα ιστογράμματα των δειγμάτων που ελήφθησαν με τη μέθοδο Metropolis-Hastings για τις τρεις παραμέτρους. Με μάυρη διακεκομένη γραμμή αναπαρίσταται ο δειγματικός μέσος.

### 1.4.3 Δειγματοληψία κατά Gibbs

Στο Παράδειγμα 1.4.1, σε κάθε βήμα προσομοιώνουμε διανύσματα $\boldsymbol{\theta} = (\alpha, \beta, \sigma)$ από την ίδια κατανομή. Μπορούμε ωστόσο να εργαστούμε και διαφορετικά, προσομοιώνοντας για κάθε συνιστώσα της $\boldsymbol{\theta}$ τιμές από την πλήρους δέσμευσης εκ των υστέρων κατανομή της. Η μέθοδος αυτή ονομάζεται Δειγματοληψία Gibbs και

[2] Αποδεικνύεται ότι η κατανομή του $\boldsymbol{X}$ δεν επηρεάζει την εκ των υστέρων κατανομή του $(\boldsymbol{\theta}|\boldsymbol{y}, \boldsymbol{X})$




Διάγραμμα 1.1: Ιστογράμματα των MCMC δειγμάτων για τις τρεις παραμέτρους $\alpha, \beta, \sigma$

αποτελεί ειδική περίπτωση του αλγορίθμου Metropolis-Hastings. Στην Δειγματοληψία Gibbs, αν $\boldsymbol{\theta} \in \mathbf{R}^d$, η κατανομή εισήγησης για την συνιστώσα $\theta_i$ στον χρόνο $t$ είναι η

$$g_i(\theta_i^*|\theta_{t,i}, \boldsymbol{\theta}_{t,-i}, \boldsymbol{x}) = p(\theta_i^*|\boldsymbol{\theta}_{t,-i}, \boldsymbol{x}),$$

όπου $\boldsymbol{\theta}_{t,-i} = (\theta_1, \theta_2, \ldots, \theta_{i-1}, \theta_{i+1}, \ldots, \theta_d)$.

Σε πολλά προβλήματα, είναι εύκολη η προσομοίωση τιμών από τις δεσμευμένες κατανομές των συνιστωσών της $\boldsymbol{\theta}$. Για παράδειγμα, γνωρίζουμε ότι αν η $\boldsymbol{\theta} = (\theta_1, \theta_2)^T|\boldsymbol{x}$ ακολουθεί δισδιάστατη κανονική κατανομή, τότε οι $\theta_1|\theta_2, \boldsymbol{x}$ και $\theta_2|\theta_1, \boldsymbol{x}$ ακολουθούν κανονική κατανομή. Η δειγματοληψία Gibbs σε αυτή την περίπτωση διευκολύνει τους υπολογισμούς, προσομοιωνοντας τιμές από δύο κανονικές κατανομές:

Πράγματι, ας θεωρήσουμε μία παρατήρηση $(y_1, y_2)$ από έναν δισδιάστατο κανονικά κατανεμημένο πληθυσμό με άγνωστη μέση τιμή $\boldsymbol{\theta} = (\theta_1, \theta_2)$ και γνωστό πίνακα διασπορών-συνδιασπορών,

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}.$$

Με μια ομοιόμορφη πρότερη στην $\boldsymbol{\theta}$, η εκ των υστέρων κατανομή είναι

$$\begin{pmatrix} \theta_1 \\ \theta_2 \end{pmatrix} \Bigg| y \sim N\left( \begin{pmatrix} y_1 \\ y_2 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} \right).$$

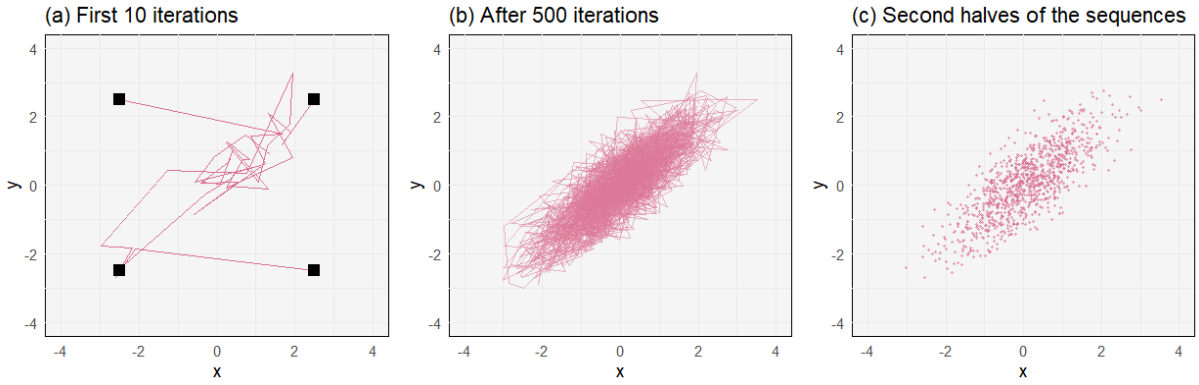
Αν και είναι απλό να προσομοιώσουμε άμεσα από την κοινή εκ των υστέρων κατανομή των $(\theta_1, \theta_2)$, για τους σκοπούς της εργασίας θα παρουσιάσουμε το πώς εφρμόζεται εδώ η Δειγματοληψία Gibbs.

Χρειαζόμαστε τις εκφράσεις για τις δεσμευμένες εκ των υστέρων κατανομές, οι οποίες, από τις ιδιότητες της πολυμεταβλητής κανονικής κατανομής[3], είναι

$$\theta_1|\theta_2, y \sim N(y_1 + \rho(\theta_2 - y_2), 1 - \rho^2)$$

$$\theta_2|\theta_1, y \sim N(y_2 + \rho(\theta_1 - y_1), 1 - \rho^2).$$

[3]Οι σχετικές ιδιότητες αναφέρονται εν συντομία και στο Κεφάλαιο 5




Διάγραμμα 1.2: Τέσσερα ανεξάρτητα μονοπάτια τις δειγματοληψίας Gibbs, για μία δισδιάστατη κανονική κατανομή με $\rho = 0.8$. Με μαύρα σημεία συμβολίζονται τα αρχικά σημεία. Βλ. Παράρτημα Β για τον αντίστοιχο κώδικα.

Για παράδειγμα, ας θεωρήσουμε την περίπτωση $\rho = 0.8$, τα δεδομένα $(y_1, y_2) = (0, 0)$, και τέσσερις ανεξάρτητες ακολουθίες προσομοιωμένων σημείων, με αρχικά σημεία τα $(\pm 2.5, \pm 2.5)$. Τα αποτελέσματα της προσομοίωσης αυτής παρουσιάζονται στο Διάγραμμα 1.2.

# Κεφάλαιο 2

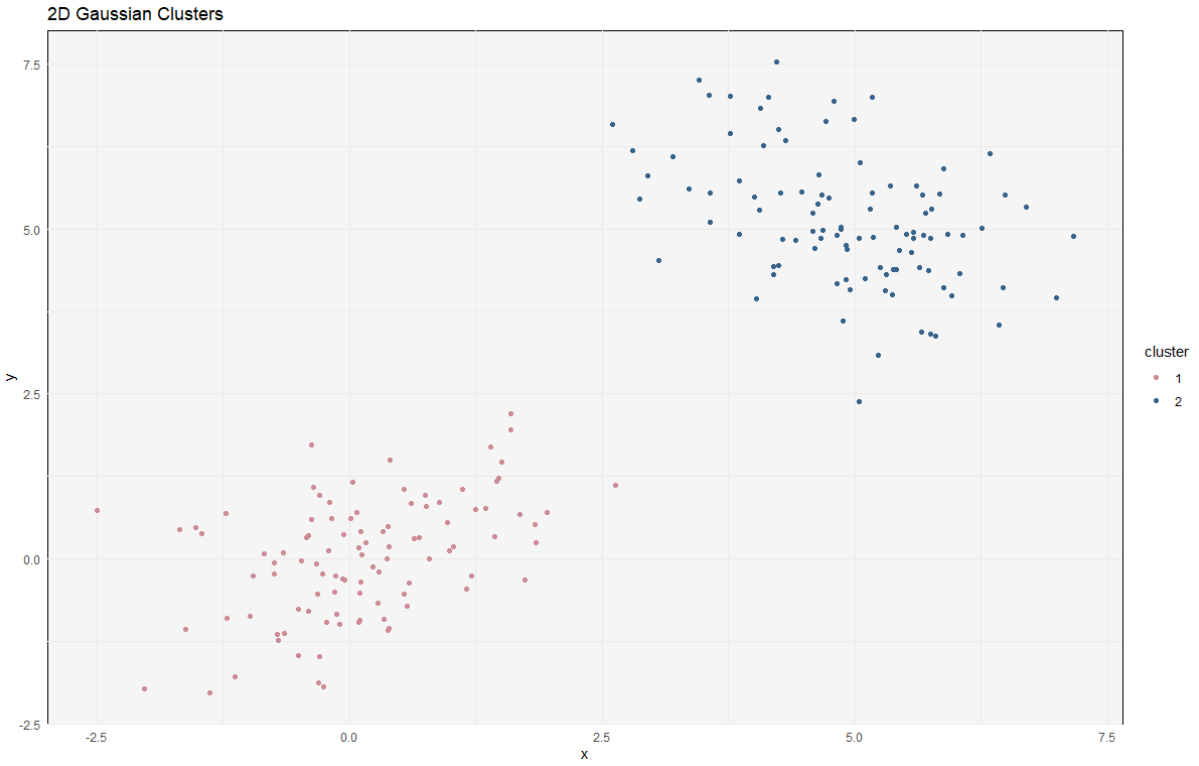
# Η Διαδικασία Dirichlet

Στο προηγούμενο κεφάλαιο, εισαχθήκαμε στον Μπεϋζιανό τρόπο σκέψης, εξετάζοντας την ενσωμάτωση της προηγούμενης γνώσης μέσω της χρήσης των εκ των προτέρων κατανομών και την ενημέρωση των πεποιθήσεων μας μετά την συλλογή νέων δεδομένων. Με την εφαρμογή του Νόμου του Bayes, διερευνήσαμε πώς τα στατιστικά μοντέλα μπορούν να αναπτυχθούν και να επεκταθούν ώστε να συμπεριλάβουν την αβεβαιότητα και την πιθανότητα στις προβλέψεις τους. Η παραμετρική προσέγγιση που ακολουθήθηκε, αν και παρέχει μια σταθερή βάση για την εκτίμηση και την συμπερασματολογία, αποδεικνύεται ανεπαρκής σε προβλήματα όχι τόσο αυστηρώς ορισμένα. Ένα τέτοιο πρόβλημα είναι αυτό της συσταδοποίησης (*clustering*). Σε αυτά τα προβλήματα αποδεικνύεται πολύτιμη η χρήση της διαδικασίας Dirichlet.

## 2.1 Το πρόβλημα της συσταδοποίησης

### 2.1.1 Κανονικά μοντέλα μίξης, ο αλγόριθμος EM και η ανάγκη για μη παραμετρική προσέγγιση

Το πρόβλημα της συσταδοποίησης έγκειται στην διαμέριση ενός δοσμένου συνόλου $\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_n\}$ σε $K$ συστάδες έτσι ώστε να ελαχιστοποιείται η διασπορά εντός των συστάδων αλλά να μεγιστοποιείται η διασπορά μεταξύ των συστάδων.




Διάγραμμα 2.1: Παράδειγμα δεδομένων προς συσταδοποίηση

Ας υποθέσουμε ότι τα δεδομένα μας είναι αυτά του Διάγραμματος 2.1. Θέλουμε να τα ομαδοποιήσουμε, να προσδιορίσουμε δηλαδή δύο ή περισσότερες κατανομές από τις οποίες προέρχεται κάθε παρατήρηση. Στο Διάγραμμα 2.1 οι συστάδες είναι διακεκριμένες, 2 στο πλήθος και εύκολα διαχωρίσιμες ακόμα και οπτικά. Συγκεκριμένα, τα δεδομένα παράχθηκαν από δύο δισδιάστατες κανονικές κατανομές:

$$N_1 = N\left(\begin{bmatrix}0\\0\end{bmatrix}, \begin{bmatrix}1 & 0.5\\0.5 & 1\end{bmatrix}\right)$$
$$N_2 = N\left(\begin{bmatrix}5\\5\end{bmatrix}, \begin{bmatrix}1 & -0.5\\-0.5 & 1\end{bmatrix}\right),$$

ας υποθέσουμε όμως πως αυτό δεν το γνωρίζουμε. Πώς μπορούμε λοιπόν να περιγράψουμε τα δεδομένα αυτά; Για την μοντελοποίηση δεδομένων αυτής της μορφής ενδείκνυται η χρήση (Πεπερασμένων) Μοντέλων *Μίξης* (Finite Mixture Models): ένα μοντέλο μίξης είναι ενα στατιστικό μοντέλο για την αναπαράσταση υποπληθυσμών μέσα σε έναν πληθυσμό, καθένας εκ των οποίων προέρχεται από μια διαφορετική κατανομή. Υποθέτουμε λοιπόν ότι τα δεδομένα προέρχονται από την κατανομή

$$f(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{K} \rho_i p(\boldsymbol{x}; \boldsymbol{\theta}_i) \tag{2.1.1}$$

και θέλουμε να εκτιμήσουμε τις παραμέτρους $\{\rho_i\}_{i=1,\dots,K}$, δηλαδή την πιθανότητα η παρατήρηση $x$ να ανήκει στην $i$ συστάδα και $\{\boldsymbol{\theta}_i\}_{i=1,\dots,K}$, δηλαδή τις παραμέτρους της κατανομής κάθε επι μέρους συστάδας.

Ο κλασικός (παραμετρικός) τρόπος αντιμετώπισης του προβλήματος είναι η εφαρμογή του αλγορίθμου Expectation-Maximization (EM), για την εκτίμηση των παραμέτρων $\{\rho_i\}_{i=1,\dots,K}$ και $\{\boldsymbol{\theta}_i\}_{i=1,\dots,K}$, και προϋποθέτει τον εκ των προτέρων προσδιορισμό του πλήθους συστάδων, $K$. Ο αλγόριθμος EM είναι μια αριθμητική μέθοδος βελτιστοποίησης, μέσω της οποίας υπολογίζουμε θέσεις (τοπικών) μέγιστων της συνάρτησης πιθανοφάνειας και συνίσταται από δύο βήματα, τα οποία στο πρόβλημα μας περιγράφονται ως εξής:

- Θεωρούμε την λανθάνουσα (*latent*) μεταβλητή $\boldsymbol{Z} \sim \text{Categorical}(K, \boldsymbol{\rho} = [\rho_1, \rho_2, \cdots, \rho_K])$ που κωδικοποιεί την συνιστώσα της μίξης από την οποία προέρχεται η κάθε παρατήρηση.
- **Expectation:** Εκτιμούμε την "κατανομή" των παρατηρήσεων στις συνιστώσες, δεδομένων των παραμέτρων στην τρέχουσα επανάληψη.
- **Maximization:** Ανανεώνουμε τις τιμές των αγνώστων παραμέτρων ώστε να μεγιστοποιείται η αναμενόμενη τιμή της λογαριθμοποιημένης πιθανοφάνειας των αγνώστων παραμέτρων, δεδομένου του $\mathbf{Z}$.

Στην περίπτωση που οι συνιστώσες της μίξης είναι κανονικές κατανομές, κάνουμε λόγο για Gaussian Mixture Model (GMM) και η προς μεγιστοποίηση ποσότητα στο βήμα Maximization, είναι η

$$\ell(\boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{i=1}^{K} \log\Big(p(\boldsymbol{x}^{(i)}; \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\Big) = \sum_{i=1}^{K} \log\left(\sum_{z^{(i)}=1}^{k} p(\boldsymbol{x}^{(i)}|z^{(i)}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(z^{(i)}; \boldsymbol{\rho})\right).$$

Στο βήμα *Expectation* του αλγορίθμου καλούμαστε σε κάθε επανάληψη $r$ να υπολογίσουμε την δεσμευμένη πιθανότητα η παρατήρηση $i$ να ανήκει στη $j$ συνιστώσα της μίξης, δεδομένων των τρεχουσών τιμών των αγνώστων παραμέτρων. Για την $j$ συνιστώσα, $j = 1, \dots, K$, έχουμε

$$r_j^{(i)} := p(z^{(i)} = j|\boldsymbol{x}^{(i)}; \boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma}).$$

Εφαρμόζοντας τον κανόνα του Bayes έχουμε

$$r_j^{(i)} = \frac{p(\boldsymbol{x}^{(i)}|z^{(i)} = j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(z^{(i)} = j; \boldsymbol{\rho})}{\sum_{l=1}^{k} p(\boldsymbol{x}^{(i)}|z^{(i)} = l; \boldsymbol{\mu}, \boldsymbol{\Sigma}) p(z^{(i)} = l; \boldsymbol{\rho})},$$

όπου

$$p(\boldsymbol{x}^{(i)}|z^{(i)} = j; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{n/2}|\boldsymbol{\Sigma}_j|^{1/2}} \exp\left(-\frac{1}{2}(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)^T \boldsymbol{\Sigma}_j^{-1} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)\right),$$
$$p(z^{(i)} = j; \boldsymbol{\rho}) = \boldsymbol{\rho}_j.$$

Αποδεικνύεται (Rudin 2020) ότι η $\mathbb{E}_{\boldsymbol{Z}}\left(\ell(\boldsymbol{\rho}, \boldsymbol{\mu}, \boldsymbol{\Sigma})\right)$ σε κάθε βήμα μεγιστοποιείται από τις παρακάτω εκφράσεις των αγνώστων παραμέτρων, όπου $j = 1, \ldots, K$.

$$\begin{aligned}
\boldsymbol{\rho}_j &:= \frac{1}{m}\sum_{i=1}^{m} r_j^{(i)}, \\
\boldsymbol{\mu}_j &:= \frac{\sum_{i=1}^{m} r_j^{(i)} \boldsymbol{x}^{(i)}}{\sum_{i=1}^{m} r_j^{(i)}}, \\
\boldsymbol{\Sigma}_j &:= \frac{\sum_{i=1}^{m} r_j^{(i)} (\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)(\boldsymbol{x}^{(i)} - \boldsymbol{\mu}_j)^T}{\sum_{i=1}^{m} r_j^{(i)}}.
\end{aligned}$$

Ας υποθέσουμε ότι μας δίνονται τα δεδομένα του Διάγραμματος 2.1 και θέλουμε να εκτιμήσουμε με την χρήση του αλγορίθμου ΕΜ τις παραμέτρους $\boldsymbol{\mu}_1, \boldsymbol{\mu}_2, \boldsymbol{\Sigma}_1, \boldsymbol{\Sigma}_2$ των κανονικών κατανομών από τις οποίες προέρχονται. Με χρήση του κώδικα που παρατίθεται στο Παράρτημα C, λαμβάνουμε τις εκτιμήσεις

$$\begin{aligned}
\rho_1 = 0.499, \rho_2 = 0.501 \\
\boldsymbol{\mu}_1 = \begin{bmatrix} 4.914 \\ 5.122 \end{bmatrix}, \boldsymbol{\mu}_2 = \begin{bmatrix} 0.132 \\ 0.025 \end{bmatrix} \\
\boldsymbol{\Sigma}_1 = \begin{bmatrix} 0.899 & -0.403 \\ -0.403 & 0.975 \end{bmatrix}, \boldsymbol{\Sigma}_2 = \begin{bmatrix} 0.887 & 0.387 \\ 0.387 & 0.813 \end{bmatrix}.
\end{aligned}$$

οι οποίες κρίνονται αρκετά ικανοποιητικές αν λάβουμε υπόψιν το σχετικά μικρό μέγεθος του δείγματος.

Τι συμβαίνει όμως στην περίπτωση που δεν γνωρίζουμε εκ των προτέρων το πλήθος των συστάδων; Στην περίπτωση αυτή, πρέπει να εκτιμηθεί από τα δεδομένα του προβλήματος. Μία τέτοια προσέγγιση επιτρέπει ακόμα την "ανακατασκευή" του μοντέλου, όταν εφαρμοστεί σε νέα δεδομένα, ώστε να μπορεί να "εμφανίσει" συστάδες που δεν υπήρχαν αρχικά. Το μαθηματικό εργαλείο που διευκολύνει αυτή την μοντελοποίηση είναι η Διαδικασία Dirichlet, που χρησιμοποιείται ως εκ των προτέρων κατανομή για την τυχαία μεταβλητή που συμβολίζει το πλήθος των συστάδων.

## 2.2 Η πεπερασμένης διάστασης κατανομή Dirichlet

Η διαδικασία Dirichlet είναι η φυσική απειροδιάστατη γενίκευση της πεπερασμένης διάστασης κατανομής Dirichlet, η οποία με την σειρά της βασίζεται στη μονοδιάστατη κατανομή Βήτα. Στην ενότητα αυτή θα παρουσιάσουμε ορισμένες ιδιότητες της κατανομής Dirichlet.

Ας περιοριστούμε αρχικά στην μονοδιάστατη περίπτωση, και στις πιο γνώριμες κατανομές Βήτα και Γάμμα.

**Ορισμός 2.2.1** (Κατανομή Γάμμα)**.** Έστω τυχαία μεταβλητή $X \in \mathbb{R}^+$. Θα λέμε ότι η $X$ ακολουθεί κατανομή *Γάμμα* με παραμέτρους $\alpha$ και $\beta$, αν η σύνάρτηση πυκνότητας πιθανότητας της δίνεται από την σχέση

$$f(x) = \begin{cases} \frac{\beta^\alpha}{\Gamma(\alpha)} x^{\alpha-1} e^{-\beta x}, \text{ αν } 0 < x < \infty \\ 0, \text{ αλλιώς} \end{cases},$$

όπου $\alpha, \beta > 0$ και $\Gamma(\alpha) = \int_0^\infty t^{\alpha-1} e^{-t} dt$ η συνάρτηση Γάμμα. Θα σύμβολίζουμε παρακάτω την κατανομή Γάμμα ως $\Gamma(\alpha, \beta)$ όπου οι παράμετροι $\alpha, \beta$ θα καλούνται παράμετροι σχήματος και κλίμακας, αντίστοιχα.

**Ορισμός 2.2.2** (Κατανομή Βήτα)**.** Έστω τυχαία μεταβλητή $X \in [0, 1]$. Θα λέμε ότι η $X$ ακολουθεί κατανομή *Βήτα* με παραμέτρους $\alpha$ και $\beta$, αν η σύνάρτηση πυκνότητας πιθανότητας της δίνεται από την σχέση

$$f(x) = \begin{cases} \frac{\Gamma(\alpha+\beta)}{\Gamma(\alpha)\Gamma(\beta)} x^{\alpha-1}(1-x)^{\beta-1}, \text{ αν } 0 \leq x \leq 1 \\ 0, \text{ αλλιώς} \end{cases},$$

όπου $\alpha, \beta > 0$. Θα σύμβολίζουμε παρακάτω την κατανομή Βήτα ως $\text{Beta}(\alpha, \beta)$.

**Ορισμός 2.2.3** (Κατανομή Dirichlet)**.** Η κατανομή Dirichlet αποτελεί την πολυδιάστατη γενίκευση της κατανομής Βήτα. Έστω $\boldsymbol{X} = (X_1, X_2, \ldots, X_K) \in S_K \subset \mathbb{R}^K$, όπου $S_K = \{\boldsymbol{x} \in \mathbb{R}^K : 0 \leq x_k \leq$

$1, \sum_{k=1}^{K} x_k = 1\}$. Θα λέμε ότι η $\boldsymbol{X}$ ακολουθεί κατανομή *Dirichlet* με παράμετρο $\boldsymbol{\alpha} = (\alpha_1, \alpha_2, \dots \alpha_K) \in \mathbb{R}^K, \alpha_k > 0, k = 1, \dots, K$, αν η από κοινού συνάρτηση πυκνότητας πιθανότητας της δίνεται από την σχέση

$$f(\boldsymbol{x}) = \begin{cases} \frac{1}{B(\boldsymbol{a})} \prod_{k=1}^{K} x_k^{a_k - 1}, & \text{αν } \boldsymbol{x} \in S_k \\ 0, & \text{αλλιώς,} \end{cases}$$

όπου $B(\alpha_1, \alpha_2, \dots \alpha_K)$ η φυσική γενίκευση της συνάρτησης Βήτα σε $K$ διαστάσεις, δηλαδή

$$B(\alpha_1, \alpha_2, \dots, \alpha_K) = \frac{\prod_{i=1}^{K} \Gamma(\alpha_\kappa)}{\Gamma\left(\sum_{k=1}^{K} \alpha_k\right)}.$$

Θα συμβολίζουμε την κατανομή Dirichlet ως $Dir(\boldsymbol{a})$. Για τη μέση τιμή, διάμεσο και διασπορά ισχύουν, αντίστοιχα,

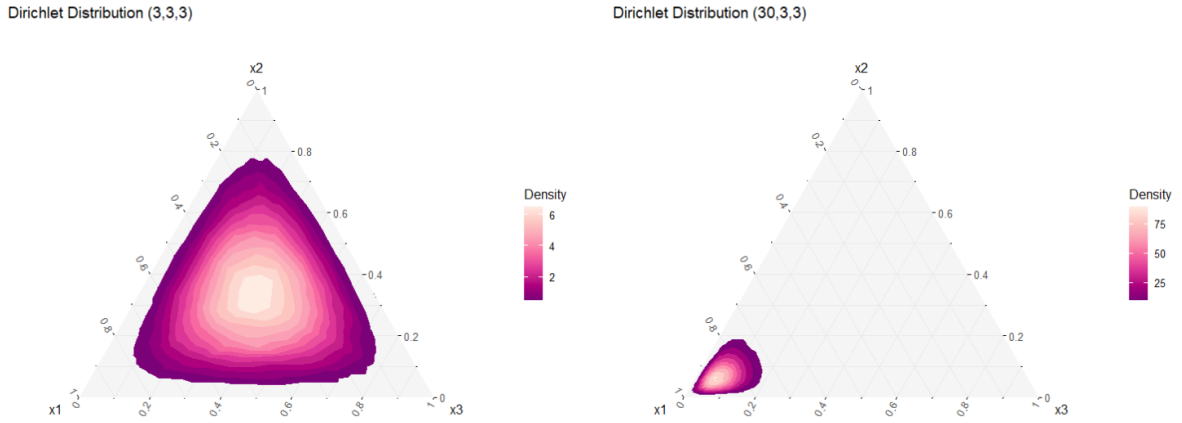
$$\mathbb{E}(x_k) = \frac{\alpha_k}{\sum_{k=1}^{K} \alpha_k}, \ mode(x_k) = \frac{\alpha_k - 1}{\sum_{k=1}^{K} \alpha_k - K}, Var(x_k) = \frac{\alpha_k(\sum_{k=1}^{K} \alpha_k - \alpha_k)}{\left(\sum_{k=1}^{K} \alpha_k\right)^2 (\sum_{k=1}^{K} \alpha_k + 1)}. \tag{2.2.1}$$

Συχνά χρησιμοποιείται η *συμμετρική Dirichlet* κατανομή, όπου $\alpha_k = \frac{\alpha}{K}, k = 1, \dots, K$, για κάποια παράμετρο $\alpha > 0$. Τότε,

$$\mathbb{E}(x_k) = \frac{1}{K}, \ Var(x_k) = \frac{K - 1}{K^2(\alpha + 1)}. \tag{2.2.2}$$

Γενικά, η παράμετρος $\boldsymbol{a}$ και συγκεκριμένα το άθροισμα των συνιστωσών της, ρυθμίζει το πόσο "πλατιά" είναι η κατανομή, ενώ για κάθε $k = 1, \dots, K$ η συνιστώσα $\alpha_k$ ρυθμίζει το *πού* επιτυγχάνεται η κορυφή. Για παράδειγμα, η $Dir(1, 1, 1)$ είναι ομοιόμορφη στο $S_3$, η $Dir(3, 3, 3)$ είναι μια "πλατιά" κατανομή γύρω από το $(\frac{1}{3}, \frac{1}{3}, \frac{1}{3})$ ενώ η $Dir(30, 3, 3)$ είναι μια κατανομή με την πυκνότητα της συσσωρευμένη γύρω από το $(1, 0, 0)$.

Στην περίπτωση της τρισδιάστατης κατανομής Dirichlet, είναι εύκολο-και χρήσιμο-να οπτικοποιήσουμε την πυκνότητα της για διαφορετικές τιμές των παραμέτρων $\alpha_1, \alpha_2, \alpha_3$ με την βοήθεια ενός τριγωνικού διαγράμματος. Στο Διάγραμμα 2.2 απεικονίζονται $10^6$ δείγματα από την $Dir(3, 3, 3)$ και την $Dir(30, 3, 3)$.




Διάγραμμα 2.2: Τριγωνικά διαγράμματα $10^6$ τυχαίων δειγμάτων από την τρισδιάστατη κατανομή $Dir(\boldsymbol{a})$ για διαφορετικές παραμέτρους $\boldsymbol{a}$. Αριστερά: $\boldsymbol{a} = (3, 3, 3)$, δεξιά: $\boldsymbol{a} = (30, 3, 3)$.

## 2.3 Το Μπεϋζιανό ιστόγραμμα

Ας θεωρήσουμε ένα απλό σενάριο στο οποίο $y_i \overset{\text{iid}}{\sim} f, i = 1, \dots, n$. Στόχος είναι να εκτιμήσουμε την πυκνότητα $f$, δουλεύοντας πάντα στο Μπεϋζιανό πλαίσιο αναφοράς. Γνωρίζουμε ότι το ιστόγραμμα, χρησιμοποιείται συχνά ως μία απλή μορφή εκτίμησης της πυκνότητας. Η παραμετρική μορφή του ιστογράμματος που θα παρουσιαστεί ευθύς αμέσως, θα λειτουργήσει και ως έναυσμα για την μετάβαση στη πλήρως μη παραμετρική εκτίμηση πυκνότητας που θα εισάγουμε παρακάτω.

Έστω μία διαμέριση του στηρίγματος, $\{\xi_1, \dots \xi_k\}$, τέτοια ώστε $\xi_0 < \xi_1 < \cdots < \xi_k$ και $y_i \in [\xi_0, \xi_k]$. Η εκτίμηση μέσω ιστογράμματος δίνεται από την σχέση

$$f(y) = \sum_{h=1}^{k} \mathbb{1}_{\xi_{h-1} < y \leq \xi_h} \frac{\pi_h}{\xi_h - \xi_{h-1}}$$

όπου $\boldsymbol{\pi} = (\pi_1, \dots, \pi_k)$ το άγνωστο διάνυσμα μάζας πιθανότητας. Ολοκληρώνουμε τον προσδιορισμό του μοντέλου θεωρώντας ως πρότερη κατανομή για το $\boldsymbol{\pi}$ την $Dir(\alpha_1, \dots, \alpha_k)$,

$$p(\boldsymbol{\pi}|\boldsymbol{a}) = \frac{\Gamma(\sum_{h=1}^{k} \alpha_h)}{\prod_{h=1}^{k} \Gamma(\alpha_h)} \prod_{h=1}^{k} \pi_h^{\alpha_h - 1}.$$

Αν τώρα εκφράσουμε το διάνυσμα των υπερπαραμέτρων ως $\boldsymbol{\alpha} = \alpha \boldsymbol{\pi}_0$, όπου $\alpha$ η κλίμακα που ερμηνευούμε συνήθως ως εκ των προτέρων μέγεθος δείγματος (Gelman et al. 2013) και

$$\boldsymbol{\pi}_0 = \mathbb{E}(\boldsymbol{\pi}|\boldsymbol{\alpha}) = \left( \frac{\alpha_1}{\sum_h \alpha_h}, \dots, \frac{\alpha_k}{\sum_h \alpha_h} \right),$$

ο εκ των προτέρων μέσος, η εκ των υστέρων κατανομή του $\boldsymbol{\pi}$ υπολογίζεται ως

$$p(\boldsymbol{\pi}|y) \propto \prod_{h=1}^{k} \pi_h^{\alpha_h - 1} \prod_{i: y_i \in (\xi_{h-1}, \xi_h]} \frac{\pi_h}{\xi_h - \xi_{h-1}} \tag{2.3.1}$$

$$\propto \pi_h^{a_h + n_h - 1} \overset{D}{=} Dir(\alpha_1 + n_1, \dots, \alpha_k + n_k), \tag{2.3.2}$$

όπου $n_h = \sum_i \mathbb{1}_{\xi_{h-1} < y_i \leq \xi_h}$ το πλήθος των παρατηρήσεων στο $h-$στό κελί του ιστογράμματος.

Η εκ των υστέρων κατανομή είναι επίσης μία κατανομή Dirichlet, δηλαδή η Dirichlet είναι συζυγής κατανομή στο μοντέλο. Διαισθητικά, επιλέγοντας την κατανομή Dirichlet ως εκ των προτέρων κατανομή για το $\boldsymbol{\pi}$, μπορούμε να ερμηνεύσουμε τις υπερπαραμέτρους $\alpha_i$, ως το πλήθος των παρατηρήσεων που "πιστεύουμε" εκ των προτέρων ότι βρίσκονται σε κάθε διαδοχικό υποδιάστημα της διαμέρισης (κελί).

**Παράδειγμα 2.3.1.** *Μπεϋζιανό ιστόγραμμα για εκτίμηση πυκνότητας δεδομένων που προέρχονται από μοντέλο μίξης δύο Βήτα κατανομών.*

Έστω δείγμα $n = 100$ παρατηρήσεων από τη μίξη

$$f(y) = 0.75 \cdot \text{Beta}(y|1, 5) + 0.25 \cdot \text{Beta}(y|20, 2).$$

Εφαρμόζοντας την προσέγγιση του Μπεϋζιανού ιστογράμματος με 10 ισαπέχοντα σημεία να σχηματίζουν τα ορθογώνια στο $[0, 1]$, λαμβάνουμε την εκ των υστέρων κατανομή όπως στην 2.3.2. Στο Διάγραμμα 2.3 αντιπαρατίθεται η γραφική παράσταση της πυκνότητας πιθανότητας του μοντέλου μίξης καθώς και το ιστόγραμμα των 100 δειγμάτων από την ύστερη κατανομή, όπως αυτή προκύπτει με τη μέθοδο του Μπεϋζιανού ιστογράμματος.

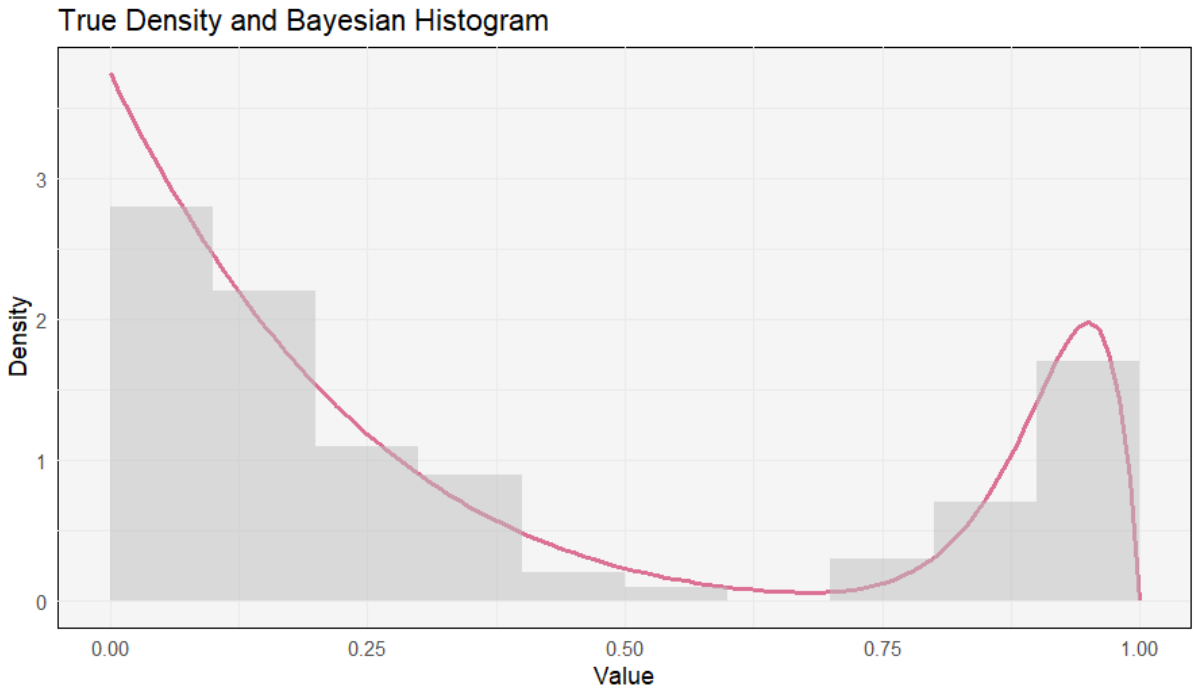
Παρατηρούμε ότι η μέθοδος προσεγγίζει επαρκώς την πραγματική πυκνότητα. Εντούτοις, πάσχει από ευαισθησία στην τοποθέτηση και το πλάτος των κελιών (Gelman et al. 2013), αλλά προσφέρει το πλεονέκτημα της συζυγίας και της ευκολίας στην ερμηνεία των υπερπαραμέτρων. Για να ξεπεράσουμε την ανάγκη προσδιορισμού των κελιών, διατηρώντας όμως την απλότητα της προσέγγισης του Μπεϋζιανού ιστογράμματος, εισάγουμε την έννοια της Διαδικασίας Dirichlet.

## 2.4 Η διαδικασία Dirichlet

Η διαδικασία Dirichlet προτάθηκε πρώτη φορά από τον Ferguson (Ferguson 1973), ως μία κατάλληλη πρότερη κατανομή για μη-παραμετρικά προβλήματα και αποτελεί απειροδιάστατη γενίκευση της κατανομής Dirichlet.

Μια πρότερη διαδικασία Dirichlet είναι ένα μέτρο πιθανότητας (κατανομή) πάνω στον χώρο των κατανομών $G : \Theta \to [0, 1]$ με $G(\theta) \geq 0, \int_\Theta G(\boldsymbol{\theta}) d\boldsymbol{\theta} = 1$, τέτοιό ώστε για κάθε διαμέριση $B_1, \dots, B_K$ του στηρίγματος $\Theta$, το διάνυσμα $(G(B_1), \dots, G(B_K))$ να ακολουθεί κατανομή Dirichlet και συγκεκριμένα

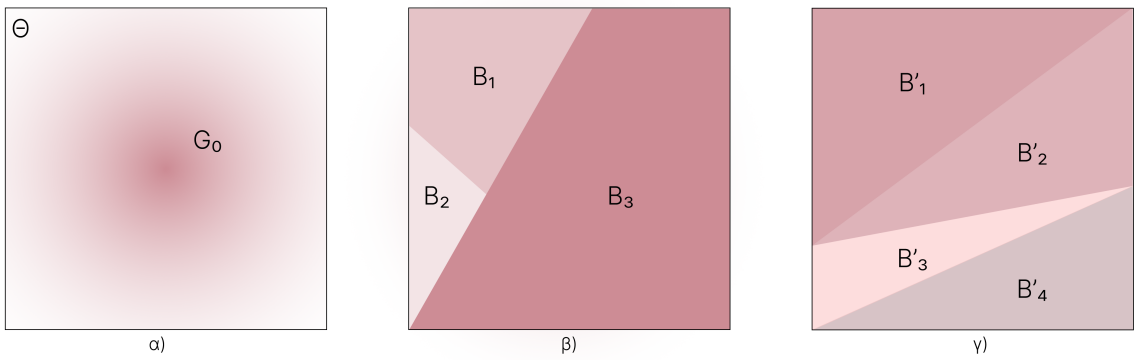
$$(G(B_1), \dots, G(B_K)) \sim Dir(\alpha G_o(B_1), \dots, \alpha G_0(B_K)) \tag{2.4.1}$$




Διάγραμμα 2.3: Πραγματική πυκνότητα και ιστόγραμμα 100 δειγμάτων από τη κατανομή που προκύπτει με τη μέθοδο του Μπεϋζιανού ιστογράμματος

όπου η παράμετρος $\alpha$ ονομάζεται *παράμετρος συγκέντρωσης* ενώ η κατανομή $G_0$ ονομάζεται *μέτρο βάσης*. Τότε, συμβολίζουμε $G \sim DP(\alpha, G_0)$.

Στο Διάγραμμα 2.4 παρουσιάζεται στο υπο-Διάγραμμα α) ως παράδειγμα ένα μέτρο βάσης $G_0$ (δισδιάστατη κανονική κατανομή) στον δισδιάστατο χώρο Θ, καθώς και δύο πιθανές διαμερίσεις. Ο χρωματισμός κάθε χωρίου $B_k$ της διαμέρισης είναι ανάλογος της $\mathbb{E}[G(B_k)] = G_0(B_k)$.




Διάγραμμα 2.4: α) Ένα μέτρο βάσης $G_0$ στον δισδιάστατο χώρο Θ. β) Μία πιθανή διαμέριση του Θ σε $K = 3$ χωρία. γ) Μια εκλεπτυσμένη διαμέριση του Θ σε $K = 4$ χωρία.

Από τον ορισμό της διαδικασίας Dirichlet προκύπτει ότι, για κάθε $B \in \Theta$ ισχύει ότι η περιθώρια κατανομή στο χωρίο $B$ είναι Βήτα, δηλαδή

$$G(B) \sim Beta(\alpha G_0(B), \alpha(1 - G_0(B)).$$

Για τον πρότερο μέσο, ισχύει ότι

$$\mathbb{E}(G(B)) = G_0(B),$$

δηλαδή η πρότερη $G$ είναι κεντραρισμένη στην $G_0$. Για για την πρότερη διασπορά, ισχύει

$$Var(G(B)) = \frac{G_0(B)(1 - G_0(B))}{1 - \alpha}.$$

Μια ακόμα ελκυστική ιδιότητα της διαδικασίας Dirichlet είναι η συζυγία: Έστω $\boldsymbol{\theta}_i \overset{iid}{\sim} G, i = 1, \ldots, n$ και $G \sim DP(\alpha, G_0)$. Τότε, με χρήση της 2.4.1 και λόγω της συζυγίας της πεπερασμένης κατανομής Dirichlet, για κάθε διαμέριση $B_1, \ldots, B_K$, θα ισχύει

$$G(B_1), G(B_2), \ldots, G(B_K)|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim Dir\left(\alpha G_0(B_1) + \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{\theta}_i \in B_1}, \ldots, \alpha G_0(B_K) + \sum_{i=1}^{n} \mathbb{1}_{\boldsymbol{\theta}_i \in B_K}\right)$$

Ισοδύναμα, με βάση τον ορισμό της διαδικασίας Dirichlet,

$$G|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \sim DP\left(\alpha + n, \frac{1}{\alpha + n}\alpha G_0 + \sum_{i=1}^{n} n_i\right).$$

Επομένως, η διαδικασία Dirichlet είναι ελκυστική αφού προκύπτει από ένα μοντέλο παρόμοιο με το Μπεϋζιανό Ιστόγραμμα, χωρίς όμως να υπάρχει η εξάρτηση από τον καθορισμό των κελιών. Ταυτόχροα, η απλή διατύπωση του μέσου και της διασποράς της έχει διαισθητικά οφέλη: μπορούμε να φανταστούμε ότι η πρότερη είναι κεντραρισμένη σε μία κατανομή $G_0$, ενώ η παράμετρος $\alpha$ ρυθμίζει την αβεβαιότητα της επιλογής αυτής. Ακόμα, η ιδιότητα της συζυγίας επιτρέπει υπολογιστικά ευκολότερες "ανανεώσεις" της ύστερης κατανομής.

Συμπληρωματικά, όσον αφορά την προβλεπτική κατανομή του $\boldsymbol{\theta}_{n+1}|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n$ όπου $\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n \overset{iid}{\sim} G$ και $G \sim DP(a, G_0)$, για κάθε μετρήσιμο $A \in \Theta$ έχουμε

$$p(\boldsymbol{\theta}_{n+1} \in A|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n) = \mathbb{E}[G(A)|\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_n] \tag{2.4.2}$$

$$= \frac{1}{a + n}\left(aG_0(A) + \sum_{h=1}^{n} \delta_{\boldsymbol{\theta}_h}(A)\right), \tag{2.4.3}$$

ενώ μάλιστα αποδεικνύεται (Blackwell and MacQueen 1973) ότι για κάθε $i = 1, \ldots, n$, η δεσμευμένη εκ των προτέρων κατανομή της $\boldsymbol{\theta}_i$ δεδομένων των $\boldsymbol{\theta}_{-i} = (\boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_{i-1}, \boldsymbol{\theta}_{i+1}, \ldots, \boldsymbol{\theta}_n)$ δίνεται από την

$$\boldsymbol{\theta}_i|\boldsymbol{\theta}_{-i} \sim \frac{1}{\alpha + n - 1}\left(G_0(\boldsymbol{\theta}_i) + \sum_{h=1}^{k^{(-i)}} \frac{n_h^{(-i)}}{\alpha + n - 1}\delta_{\boldsymbol{\theta}_h^{*(-i)}}\right), \tag{2.4.4}$$

όπου $\theta_h^*, h = 1, \ldots, k^{(-i)}$ οι μοναδικές τιμές των $\boldsymbol{\theta}^{-i}$ και $n_h^{(-i)} = \sum_{j \neq i} \mathbb{1}_{\boldsymbol{\theta}_j}(\boldsymbol{\theta}_h^*)$.

### 2.4.1 Κατασκευή: Stick Breaking

Η παραπάνω συζήτηση ενδεχομένως να φαίνεται αρκετά αφηρημένη. Ας δώσουμε σε αυτό το σημείο έναν κατασκευαστικό ορισμό για την διαδικασία Dirichlet, γνωστό ως διαδικασία Stick-Breaking (Sethuraman 1994). Η αναπαράσταση αυτή επιτρέπει μια πιο διαισθητική ερμηνεία του πώς "μοιάζουν" οι πραγματοποιήσεις της $G \sim DP(\alpha, G_0)$. Συγκεκριμένα, ας θεωρήσουμε

$$G(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}}(\cdot) \tag{2.4.5}$$

$$\pi_h = v_h \prod_{l<h}(1 - v_l) \tag{2.4.6}$$

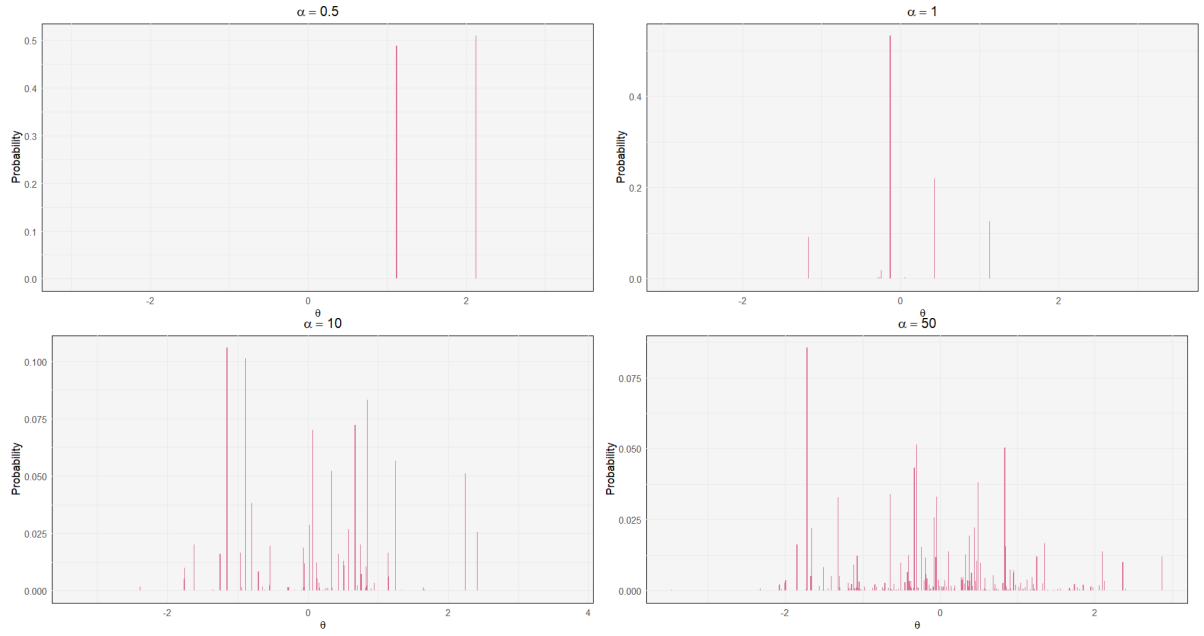
$$v_h \sim Beta(1, \alpha) \tag{2.4.7}$$

$$\boldsymbol{\theta}_h \sim G_0. \tag{2.4.8}$$

Αποδεικνύεται τότε ότι

$$G \sim DP(\alpha, G_0),$$

ενώ συμβολίζουμε συνήθως $\{\pi\}_{h=1}^{\infty} = \boldsymbol{\pi} \sim GEM(\alpha)$, προς τιμήν των Griffiths, Engen, McCloskey, ενώ η ορολογία αυτή αποδόθηκε από τον Ewens (Ewens 1990). Ισοδύναμα, το φυσικό ανάλογο των παραπάνω εκφράσεων συνοψίζεται ως εξής: Θεωρούμε ένα ξυλάκι μήκους 1, που συμβολίζει την συνολική πιθανότητα που πρέπει να ανακατανεμηθεί σε όλα τα κομμάτια στο οποίο θα το σπάσουμε διαδοχικά. Το μήκος κάθε κομματιού αντιπροσωπεύει την πιθανότητα που αναθέτουμε στο κάθε κομμάτι. Αρχικά σπάμε ένα τυχαίο κομμάτι μήκους $v_1 \sim Beta(1, a)$. Το μήκος του κομματιού αυτού μας δίνει και το πρώτο "βάρος" $\pi_1$. Έπειτα, σπάμε ένα τυχαίο κομμάτι $v_2$ από το εναπομείναν ξυλάκι. Και πάλι, το $\pi_2$ συμβολίζει το μήκος του




Διάγραμμα 2.5: Πραγματοποιήσεις της διαδικασίας Stick Breaking για διαφορετικές τιμές της παραμέτρου συγκέντρωσης α. Ως $G_0$ χρησιμοποιήθηκε η τυπική κανονική κατανομή.

δεύτερου κομματιού. Όσο το $h \to \infty$, παρατηρούμε ότι τα μήκη των κομματιών γίνονται όλο και μικρότερα. Η παράμετρος συγκέντρωσης $\alpha$ καθορίζει την κατανομή των μηκών των κομματιών: μικρότερες τιμές του $\alpha$ δίνουν μεγαλύτερα μήκη για τα πρώτα κομμάτια, με τα υπόλοιπα να έχουν μικρότερα μήκη. Αυτό αντιστρέφεται για μεγαλύτερες τιμές του $\alpha$. Αυτό διαπιστώνεται εύκολα αν παρατηρήσουμε ότι

$$\mathbb{E}[v_h] = \frac{1}{1+a}.$$

Στο Διάγραμμα 2.5 αναπαρίστανται τέσσερις διαφορετικές πραγματοποιήσεις της Διαδικασίας Dirichlet μέσω της κατασκευής αυτής για διαφορετικές τιμές της παραμέτρου συγκέντρωσης α. Παρατηρούμε ότι για μικρότερες τιμές της $\alpha$ η κατανομές που προκύπτουν είναι λιγότερο "διακεχυμένες".

### 2.4.2 Η διαδικασία κινέζικου εστιατορίου

Μία ακόμα ερμηνεία της διαδικασίας Dirichlet δίνεται μέσω της λεγόμενης Διαδικασίας Κινέζικου Εστιατορίου (Blackwell and MacQueen 1973), μίας στοχαστικής διαδικασίας διακριτού χρόνου που περιγράφεται ως εξής: μπορεί κανείς να φανταστεί ένα κινέζικο εστιατόριο με άπειρα τραπέζια άπειρης χωρητικότητας. Ο πρώτος πελάτης κάθεται στο πρώτο τραπέζι. Ο κάθε πελάτης που μπαίνει στο εστιατόριο στην συνέχεια, επιλέγει πού θα κάτσει με πιθανότητα ανάλογη του αριθμού των ήδη καθήμενων σε κάθε τραπέζι. Πιο συγκεκριμένα, δεδομένου ότι $n$ πελάτες έχουν ήδη καθίσει στο εστιατόριο, ο $n+1$-οστός πελάτης επιλέγει πού θα καθίσει με πιθανότητες

$$\mathbb{P}(\text{κάθεται στο τραπέζι } k) = \frac{n_k}{n+\alpha},$$

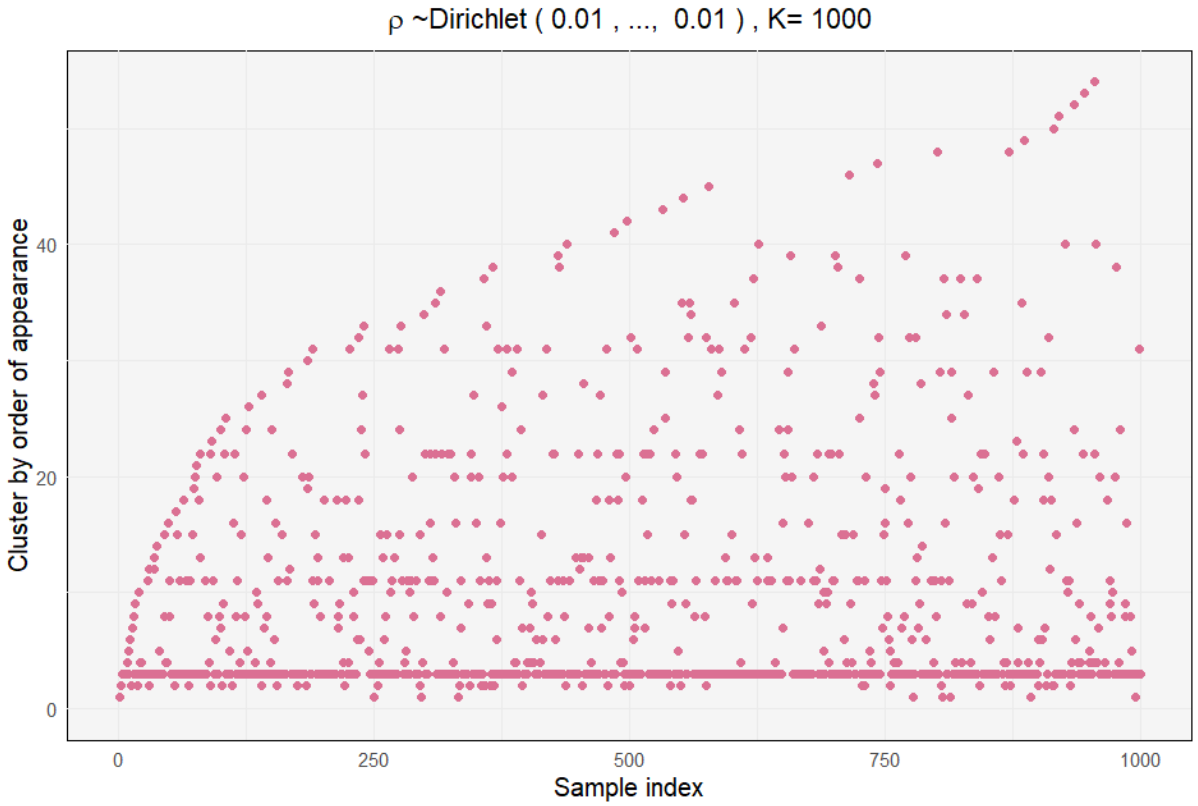
ενώ η πιθανότητα να κάτσει σε νέο (άδειο) τραπέζι είναι

$$\mathbb{P}(\text{κάθεται σε νέο τραπέζι}) = \frac{\alpha}{n+\alpha},$$

όπου $n_k$ το πλήθος των ήδη καθήμενων πελατών στο τραπέζι $k$ και $\alpha$ η παράμετρος συγκέντρωσης.

Αν επαναδιατυπώσουμε το πρόβλημα αυτό ορίζοντας $z_{n+1}$ την δείκτρια μεταβλητή που συμβολίζει το τραπέζι στο οποίο κάθεται ο $(n+1)$-οστός πελάτης, έχουμε

$$p(z_{n+1} = z | z_1, z_2, \ldots, z_n; a) = \frac{1}{a+n}\left(a\mathbb{1}_z(k^*) + \sum_{k=1}^{K}\mathbb{1}_z(k)\right), \tag{2.4.9}$$




Διάγραμμα 2.6: Ανάθεση σε συστάδες $N_{max} = 1000$ παρατηρήσεων από την Categorical$(\rho)$ όπου $\rho \sim Dir(a)$.

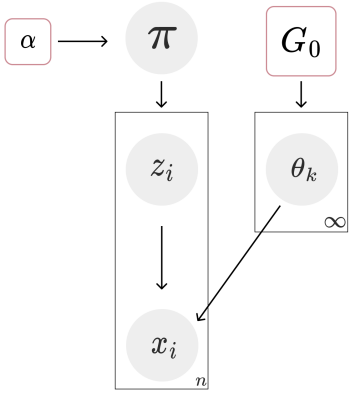
από όπου είναι άμεσα φανερή η αναλογία με την 2.4.2, αν φανταστούμε τα τραπέζια της Διαδικασίας Κινέζικου Εστιατορίου ως τις διάφορες συστάδες και τους πελάτες ως διαφορετικές παρατηρήσεις. Το αποτέλεσμα είναι μια κατανομή πάνω σε διαμερίσεις των ακεραίων, ή αλλιώς, κατανομή των πελατών σε τραπέζια. Το αναμενόμενο πλήθος των νέων συστάδων που εμφανίζονται στο δείγμα αποδεικνύεται ότι τείνει σχεδόν βεβαίως στο $a \log(N)$ όσο $N \to \infty$, δηλαδή η πολυπλοκότητα του μοντέλου αυξάνεται λογαριθμικά σε σχέση με το μέγεθος του δείγματος. Αυτό καθίσταται φανερό και από το Διάγραμμα 2.6.

## 2.5 Μοντέλα Μίξης Διαδικασίας Dirichlet (DPMM)

Ας δούμε τώρα πως η Διαδικασία Dirichlet μπορεί να αποδειχθεί χρήσιμη σε ένα πρόβλημα συσταδοποίησης, όπως αυτό που περιγράφτηκε στη πρώτη ενότητα του κεφαλαίου. Οι πραγματοποιήσεις της Διαδικασίας Dirichlet είναι διακριτές κατανομές, και ως τέτοιες, δε μπορούν να χρησιμοποιηθούν ως πρότερες κατανομές για συνεχή δεδομένα, Ωστόσο, ενδύκνεινται ως πρότερη για τις παραμέτρους του μηχανισμού που παράγει τα εν λόγω δεδομένα, όπως ενός μοντέλου μίξης. Συγκεκριμένα, κατ'αναλογία με την 2.1.1, ας θεωρήσουμε το *άπειρο μοντέλο μίξης*

$$g(\boldsymbol{x}, \boldsymbol{\theta}) = \sum_{i=1}^{\infty} \pi_i f(\boldsymbol{x}; \boldsymbol{\theta}_i). \tag{2.5.1}$$

Το σύμβολο του απείρου στην άθροιση δεν σημαίνει ότι έχουμε άπειρες κατειλημμένες συστάδες, παρά μόνο ότι το μοντέλο είναι αρκετά ευέλικτο για να προσθέσει νέες συστάδες όσο εισάγονται περισσότερα δεδομένα. Ισοδύναμα, μπορούμε να προσδιορίσουμε το μοντέλο ιεραρχικά. Αν θεωρήσουμε ότι η πρότερη κατανομή των παραμέτρων $\theta_i$ είναι η $G_0$, μπορούμε να λάβουμε δείγματα από το άπειρο αυτό μοντέλο μίξης $g$, ως




Διάγραμμα 2.7: Σχηματική αναπαράσταση της δειγματοληψίας από ένα DPMM.

εξής:

$$\boldsymbol{\pi} \sim GEM(\alpha) \tag{2.5.2}$$
$$z_i \sim \boldsymbol{\pi} \tag{2.5.3}$$
$$\boldsymbol{\theta}_k \sim G_0(\lambda) \tag{2.5.4}$$
$$x_i \sim g(\boldsymbol{\theta}_{z_i}). \tag{2.5.5}$$

Στο Διάγραμμα 2.7 παρατίθεται μια σχηματική αναπαράσταση της δειγματοληψίας από ένα Μοντέλο Μίξης Διαδικασίας Dirichlet.

### 2.5.1 Δειγματοληψία κατά Gibbs για τα DPMM

Η χρήση των DPMM έχει καταστεί υπολογιστικά επιτεύξιμη, χάρη στην ανάπτυξη μεθόδων Markov Chain Monte Carlo (MCMC) για την δειγματοληψία από την ύστερη κατανομή των παραμέτρων των επιμέρους τμημάτων του μοντέλου μίξης (Neal 2000). Οι μέθοδοι αυτές υλοποιούνται πιο εύκολα για μοντέλα βασισμένα σε συζυγείς πρότερες κατανομές, ενώ απαιτούν πιο δύσκολες αριθμητικές ολοκληρώσεις στην περίπτωση μη-συζυγών πρότερων.

Όπως αναφέρθηκε εν συντομία στο πρώτο κεφάλαιο, η μέθοδος της Δειγματοληψίας Gibbs έγκειται στην δειγματοληψία από τις πλήρους δέσμευσης ύστερες κατανομές κάθε παραμέτρου, με επαναληπτικό τρόπο. Παρακάτω θα μελετήσουμε την εφαρμογή της μεθόδου στα DPMM, αλλά πρώτα, ας επικεντρωθούμε στην περίπτωση ενός πεπερασμένου μοντέλου μίξης (FMM) της ακόλουθης μορφής:

$$\begin{aligned}
\boldsymbol{\pi} &\sim \text{Dir}\left(\frac{\alpha}{K}, \ldots, \frac{\alpha}{K}\right) \\
z_i &\sim \boldsymbol{\pi} \\
\boldsymbol{\theta}_k &\sim G_0(\lambda) \\
x_i &\sim g(\boldsymbol{\theta}_{z_i}).
\end{aligned}$$

Το DPMM αποτελεί απειροδιάστατη επέκταση του μοντέλου αυτού, αφού αν επιτρέψουμε $K \to \infty$, η Dirichlet πρότερη κατανομή της $\boldsymbol{\pi}$ αντικαθίσταται από την $GEM(\alpha)$, με βάση την Stick-Breaking κατασκευή.

Συμβολίζουμε $\boldsymbol{z}_{-i} = \{z_1, z_2, \ldots, z_{i-1}, z_{i+1}, \ldots z_k\}$ και $\boldsymbol{x} = \{x_i\}_{i=1}^n$ το διάνυσμα των παρατηρήσεων. Τότε, για κάθε δείκτρια μεταβλητή $z_i$ έχουμε

$$\begin{aligned}
p(z_i = k | \boldsymbol{z}_{-i}, \boldsymbol{x}, \boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \alpha, \lambda) &= p(z_i = k | x_i, \boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K) \\
&\propto p(z_i = k | \boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K) p(x_i | z_i = k, \boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K) \\
&= p(z_i = k | \boldsymbol{\pi}) p(x_i | \boldsymbol{\theta}_k) \\
&= \pi_k g(x_i | \boldsymbol{\theta}_k).
\end{aligned}$$

Όσον αφορά τις παραμέτρους μίξης $\boldsymbol{\pi}$, αν ορίσουμε $n_k = \sum_{i=1}^{n} \mathbb{1}_{z_i}(k)$, έχουμε

$$\begin{aligned} p(\boldsymbol{\pi}|\boldsymbol{z}, \boldsymbol{x}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \alpha, \lambda) &= p(\boldsymbol{\pi}|\boldsymbol{z}, \alpha) \\ &= \text{Dir}\left(n_1 + \frac{\alpha}{K}, \ldots, n_K + \frac{\alpha}{K}\right), \end{aligned}$$

λόγω της ιδιότητας συζυγίας που έχει η κατανομή Dirichlet.

Τέλος, όσον αφορά τις παραμέτρους της κατανομής κάθε συστάδας, $\boldsymbol{\theta}_i$, συμβολίζοντας $\boldsymbol{x}_k$ τις παρατηρήσεις που ανήκουν στην συστάδα αυτή και λόγω ανεξαρτησίας[1], λαμβάνουμε

$$\begin{aligned} p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k}, \boldsymbol{\pi}, \boldsymbol{z}, \boldsymbol{x}, \alpha, \lambda) &= p(\boldsymbol{\theta}_k|\boldsymbol{\theta}_{-k}, z, \boldsymbol{x}, \lambda) \\ &= p(\boldsymbol{\theta}_k|\boldsymbol{x}_k, \lambda) \\ &\propto G_0(\boldsymbol{\theta}_k|\lambda) L(\boldsymbol{x}_k|\boldsymbol{\theta}_k). \end{aligned}$$

Προφανώς, αν η $G_0$ είναι συζυγής πρότερη για τα $\boldsymbol{\theta}_k$, η εκ των υστέρων κατανομή της θα είναι της ίδιας μορφής.

Συνοψίζοντας, έχουμε την εξής περιγραφή για την διαδικασία Δειγματοληψίας Gibbs σε ένα FMM.

**Αλγόριθμος 2** Δειγματοληψία Gibbs για FMM

1: Δεδομένων των $\boldsymbol{\pi}^{(t-1)}$, $\{\boldsymbol{\theta}_k^{(t-1)}\}_{k=1}^K$ από την προηγούμενη επανάληψη, λαμβάνουμε νέο δείγμα των $\boldsymbol{\pi}^{(t)}$ και $\{\boldsymbol{\theta}_k^{(t-1)}\}_{k=1}^K$ ως εξής:
2: **για** $i = 1, \ldots, n$
3: Προσομοιώνουμε $z_i$ από την κατανομή:
4: $p(z_i^{(t)} = k) \propto \boldsymbol{\pi}_k^{(t-1)} f(z_i|\boldsymbol{\theta}_k^{(t-1)})$
5: **τέλος για**
6: Προσομοιώνουμε νέο βάρος μίξης $\boldsymbol{\pi}^{(t)}$ από την κατανομή
7: $\boldsymbol{\pi}^{(t)} \sim \text{Dir}\left(n_1^{(t)} + \frac{\alpha}{K}, \ldots, n_K^{(t)} + \frac{\alpha}{K}\right)$
8: όπου $n_k^{(t)} = \sum_{i=1}^{n} \mathbb{1}_{z_i^{(t)}}(k)$
9: **για** $k = 1, \ldots, K$
10: Προσομοίωσε τις παραμέτρους κάθε συστάδας, $\boldsymbol{\theta}_k$, από την κατανομή:
11: $\boldsymbol{\theta}_k^{(t)} \propto G_0(\boldsymbol{\theta}_k|\lambda) L(\boldsymbol{\theta}_k^{(t)}|\boldsymbol{\theta}_k^{(t-1)})$
12: **τέλος για**

Όπως είναι φανερό, για την ανανέωση των παραμέτρων μίξης $\pi_i$, απαιτείται σε κάθε επανάληψη η προσομοίωση τιμών από την κατανομή Dirichlet. Ωστόσο, όταν επιτρέπουμε $K \to \infty$, αυτό δεν είναι απλό εγχείρημα. Για να ξεπεράσουμε το εμπόδιο αυτό, μπορούμε να υπολογίσουμε την πλήρους δέσμευσης ύστερη κατανομή της $z_i$ ως εξής:

$$\begin{aligned} p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \alpha, \lambda) &= p(z_i = k|\boldsymbol{z}_{-i}, x_i, \boldsymbol{\theta}_k, \alpha) && (2.5.6) \\ &= p(z_i = k|\boldsymbol{z}_{-i}, \alpha, \boldsymbol{\theta}_k) p(x_i|z_i = k, \boldsymbol{z}_{-i}, \boldsymbol{\theta}_k, \alpha) && (2.5.7) \\ &= p(z_i = k|\boldsymbol{z}_{-i}, \alpha) p(x_i|\boldsymbol{\theta}_k) && (2.5.8) \\ &= \frac{n_{k,-i} + \frac{\alpha}{K}}{n + \alpha - 1} g(x_i|\boldsymbol{\theta}_k), && (2.5.9) \end{aligned}$$

όπου για την 2.5.6 χρησιμοποιήθηκε η μαρκοβιανή ιδιότητα καθώς και το αποτέλεσμα της υποσημείωσης[1], για την 2.5.7 χρησιμοποιήθηκε ο κανόνας του Bayes, για την 2.5.8 και πάλι η μαρκοβιανή ιδιότητα ενώ

[1] Αποδεικνύεται (Frey 2003) ότι δεδομένης της δείκτριας μεταβλητής $\boldsymbol{z}$, οι παράμετροι μίξης $\boldsymbol{\pi}$ και οι παράμετροι κάθε κατανομής, $\boldsymbol{\theta}_k$ είναι ανεξάρτητες:

$$p\left(\boldsymbol{\pi}, \{\boldsymbol{\theta}_k\}_{k=1}^K \mid \mathbf{z}, \mathbf{x}, \alpha, \lambda\right) = p\left(\boldsymbol{\pi} \mid \mathbf{z}, \alpha\right) \prod_{k=1}^{K} p\left(\boldsymbol{\theta}_k \mid \mathbf{x}_k, \lambda\right).$$

Επομένως η δεσμευμένη ύστερη κατανομή της $\boldsymbol{\theta}_k$ εξαρτάται μόνο από τις παρατηρήσεις $\boldsymbol{x}_k$ που ανήκουν στην εν λόγω συστάδα.

για την 2.5.9 αξιοποιήθηκε το εξής αποτέλεσμα:

$$\begin{aligned} p(z_i = k|\boldsymbol{z}_{-i}, \alpha) &= \frac{p(\boldsymbol{z}|\alpha)}{p(\boldsymbol{z}_{-i}|\alpha)} \\ &= \frac{\Gamma(n+\alpha)}{\Gamma(n+\alpha-1)} \times \frac{\Gamma(n_k + \frac{\alpha}{K})}{\Gamma(n_k - i + \frac{\alpha}{k})} \\ &= \frac{1}{n+\alpha-1} \times \frac{n_{k,-i} + \frac{\alpha}{K}}{1} \\ &= \frac{n_{k,-i} + \frac{\alpha}{K}}{n+\alpha-1}. \end{aligned}$$

Γενικεύοντας για $K \to \infty$, έχουμε

- Για την συστάδα $k$ με $n_{k,-i} > 0$ (δηλαδή, για τις ήδη υπάρχουσες συστάδες):

$$P(z_i = k|z_{-i}, \alpha) = \frac{n_{k,-i}}{n+\alpha-1}$$

- Για τις υπόλοιπες συστάδες:

$$P(z_i \neq z_j \text{ για κάθε } j \neq i|z_{-i}, \alpha) = \frac{\alpha}{n+\alpha-1},$$

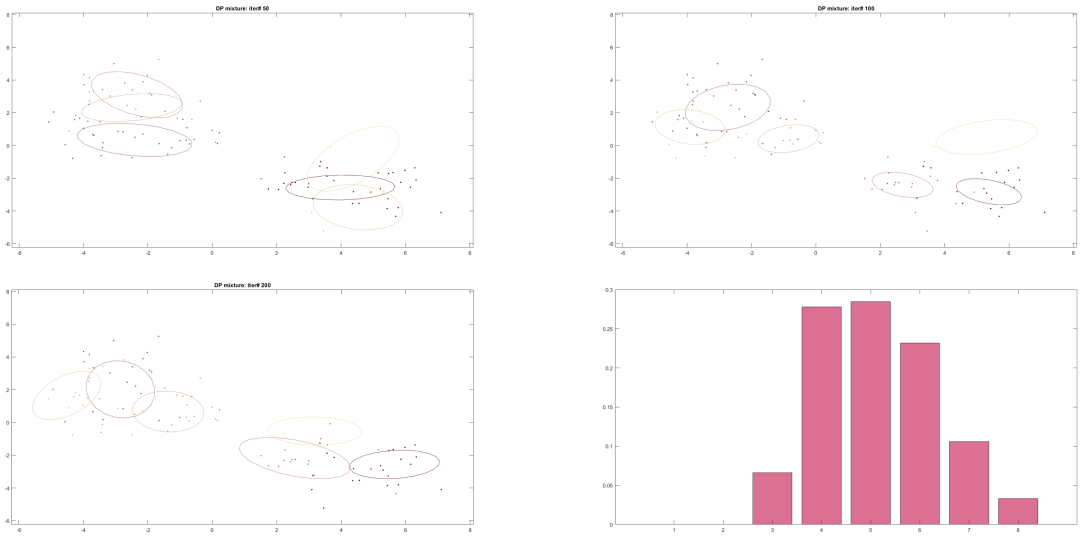
κατ' αναλογία με την κατασκευή της διαδικασίας Κινέζικου Εστιατορίου. Επομένως, η εκ των υστέρων πιθανότητα η παρατήρηση $i$ να τοποθετηθεί σε μία από τις ήδη υπάρχουσες $K$ συστάδες, είναι, αντικαθιστώντας τον πολλαπλασιαστικό παράγοντα στην 2.5.9 με το ακριβώς προηγούμενο αποτέλεσμα,

$$p(z_i = k|\boldsymbol{z}_{-i}, \boldsymbol{x}, \{\boldsymbol{\theta}_k\}_{k=1}^K, \alpha, \lambda) = \frac{n_{k,-i}}{n+\alpha-1} g(x_i|\boldsymbol{\theta}_k).$$

Αντίστοιχα, η πιθανότητα να τοποθετηθεί σε νέα συστάδα, την οποία αριθμούμε χωρίς βλάβη της γενικότητας ως $K+1$, είναι

$$\begin{aligned} p(z_i = K+1|\boldsymbol{z}_{-i}, \boldsymbol{x}, \alpha, \lambda) &= p(z_i = K+1|\boldsymbol{z}_{-i}, x_i, \alpha, \lambda) \\ &= p(z_i = K+1|\boldsymbol{z}_{-i}, \alpha, \lambda) p(x_i|z_i = K+1, \boldsymbol{z}_{-i}, \alpha, \lambda) \\ &= p(z_i = K+1|\boldsymbol{z}_{-i}, \alpha) p(x_i|\lambda) \\ &= \frac{\alpha}{n+\alpha-1} \int g(x_i|\boldsymbol{\theta}) G_0(\boldsymbol{\theta}|\lambda) d\boldsymbol{\theta}. \end{aligned}$$

Γενικά, τα DPMM είναι ανθεκτικά όσον αφορά την παράμετρο συγκέντρωσης $\alpha$, όμως το πλήθος συστάδων $K$ δεν είναι (Escobar and West 1995). Συγκεκριμένα, μεγαλύτερες τιμές της παραμέτρου συγκέντρωσης αναμένεται να οδηγούν σε μεγαλύτερο πλήθος συστάδων στο τελικό μοντέλο. Για αυτό, είναι συχνά χρήσιμο να αναθέτουμε στην $\alpha$ μία μη-πληροφοριακή πρότερη κατανομή, έστω $\Gamma(\alpha_0, \beta_0)$. Η διαδικασία της Δειγματοληψίας Gibbs για ένα DPMM συνοψίζεται στον Αλγόριθμο 3.



Διάγραμμα 2.8: Συσταδοποίηση 200 σημείων του επιπέδου που έχουν προσομοιωθεί από 5 κανονικές κατανομές για $s_0 = 3, s_1 = 1$, με χρήση ενός DPMM. Δείγματα από την ύστερη κατανομή μετά από 50,100 και 200 επαναλήψεις της δειγματοληψίας Gibbs. Στο τελευταίο Διάγραμμα παρατίθεται η ύστερη κατανομή του $K$ αγνοώντας τα 50 πρώτα δείγματα ως burn-in.

**Αλγόριθμος 3** Δειγματοληψία Gibbs για DPMM

Δεδομένων των $\boldsymbol{\iota}^{(t-1)}$, $\{\boldsymbol{\theta}_k^{(t-1)}\}_{k=1}^K$ και $\{z_i^{(t-1)}\}_{i=1}^n$ από την προηγούμενη επανάληψη, λαμβάνουμε νέο δείγμα των $\{\boldsymbol{\theta}_k^{(t)}\}_{k=1}^K$ και $\{z_i^{(t)}\}_{i=1}^n$ ως εξής:
Θέτουμε $z = z^{(t-1)}$, $\alpha = \alpha^{(t-1)}$
**για** $i = 1, \dots, n$
    Αφαιρούμε την παρατήρηση $x_i$ από την συστάδα $i$, καθώς θα προσοσμοιώσουμε νέο $z_i$.
    **αν** $x_i$ είναι η μόνη παρατήρηση της τρέχουσας συστάδας **τότε**
        Η συστάδα αυτή καθίσταται άδεια. Την αφαιρούμε, μαζί με τις παραμέτρους της, $\boldsymbol{\theta}_i$ και ελαττώνουμε το $K$ κατά 1.
    **τέλος αν**
    Προσομοιώνουμε το νέο $z_i$ με βάση τις εξής πιθανότητες:
    $p(z_i = k, k \leq K) \propto \frac{n_{k,-i}-1}{n+\alpha-1} g(x_i|\boldsymbol{\theta}_k^{(t-1)})$
    $p(z_i = K+1) \propto \frac{\alpha}{n+\alpha-1} \int g(x_i|\boldsymbol{\theta}) G_0(\boldsymbol{\theta}|\lambda) d\boldsymbol{\theta}$
    **αν** $z_i = K+1$ **τότε**
        Παίρνουμε νέα συστάδα, την $K+1$. Για αυτήν, προσομοιωνουμε από την $H$ παραμέτρο $\boldsymbol{\theta_{K+1}}$.
    **τέλος αν**
**τέλος για**
**για** $k = 1, \dots, K$
    Προσομοιώνουμε τις παραμέτρους κάθε συστάδας, $\boldsymbol{\theta}_k$ από την ακόλουθη κατανομή
    $\boldsymbol{\theta}_k^{(t)} \propto G_0(\boldsymbol{\theta}_k|\lambda) L(x_k^{(t)}|\boldsymbol{\theta}_k^{(t-1)})$
**τέλος για**
Θέτουμε $z^{(t)} = z$
**αν** $\alpha \sim \text{Gamma}(a_0, b_0)$ **τότε**
    Προσομοιώνουμε $\alpha^{(t)} \sim p(\alpha|K, n, a_0, b_0)$.
**τέλος αν**

### 2.5.2 Εφαρμογή: κανονικό μοντέλο μίξης

Προσομοιώνουμε $n = 100$ σημεία στο επίπεδο από $K = 5$ δισδιάστατες κανονικές κατανομές με κέντρα $\mu \sim s_0 N(0, 1)$ στα οποία προσθέτουμε επίσης κανονικά κατανεμημένο θόρυβο $\varepsilon \sim s_1 N(0, 1)$[2]. Προσαρμόζουμε στη συνέχεια ένα DPMM με χρήση $n = 200$ επαναλήψεων της Δειγματοληψίας Gibbs. Θέτουμε την

[2]Χρησιμοποιούμε $s_0 = 3, s_1 = 1$. Ο αντίστοιχος κώδικας παρατίθεται στο Παράρτημα 3.

παράμετρο συγκέντρωσης $\alpha = 1$ και αρχικοποιούμε το πλήθος των συστάδεων ως $K = 1$. Στο Διάγραμμα 2.8 παρουσιάζεται η εξέλιξη της συσταδοποίησης καθώς και το ιστόγραμμα της εκ των υστέρων κατανομής του πλήθους συστάδων $K$.

Είναι φανερό πως η μέθοδος φαίνεται να ανακαλύπτει σχετικά σύντομα μία "λογική" συσταδοποίηση των σημείων, την στιγμή που η προσαρμογή ενός πεπερασμένου μοντέλου μίξης ενδέχεται να παγιδευόταν σε τοπικά ελάχιστα. Το DPMM ξεπερνάει το εμπόδιο αυτό, σχηματίζοντας από τις πρώτες επαναλήψεις περιττές συστάδες, στις οποίες "καταφεύγει" για να αποφύγει να παγιδευτεί σε τοπικά ελάχιστα (Murphy 2022).

# Κεφάλαιο 3

# Η Ιεραρχική Διαδικασία Dirichlet (HDP)

Η έννοια των ιεραρχικών μοντέλων είναι θεμελιώδης στην Μπεϋζιανή Στατιστική. Ένα Μπεϋζιανό μοντέλο ορίζεται από μία κατανομή των οποίων οι παράμετροι ακολουθούν με την σειρά τους μια άλλη κατανομή, κάτι που μπορεί δυνητικά να συνεχιστεί επ'άπειρον. Για παράδειγμα, στο μη-παραμετρικό πλαίσιο στο οποίο εργαζόμαστε, στη παράμετρο συγκέντρωσης $\alpha$ της $DP(\alpha, G_0)$ συχνά αποδίδεται μία πρότερη (παραμετρική) κατανομή, όπως και στις υπερπαραμέτρους της κατανομής βάσης, $G_0$. Η ιδέα αυτή μπορεί να επεκταθεί, αν διαχειριστούμε την $G_0$ μη-παραμετρικά. Παρακάτω παρουσιάζεται η Ιεραρχική Διαδικασία Dirichlet (Hierarchical Dirichlet Process - HDP), όπως αυτή προτάθηκε από τους Teh, Jordan, Beal και Blei (Teh et al. 2006).

## 3.1 Ορισμός

Έστω $J$ ομάδες δεδομένων, καθεμία εκ των οποίων αποτελείται από $n_j$ παρατηρήσεις $(x_{j1}, \ldots, x_{jn_j})$. Σκοπος μας είναι να μοντελοποιήσουμε τα δεδομένα κάθε ομάδας με ένα μοντέλο μίξης. Κάθε επιμέρους μοντέλο θα έχει τις δικές του παραμέτρους μίξης. Αυτές θα προέρχονται από Διαδικασίες Dirichlet με το ίδιο μέτρο βάσης, το οποίο με τη σειρά του είναι επίσης μια Διαδικασία Dirichlet. Έτσι καθορίζεται το μοντέλο της Ιεραρχικής Διαδικασίας Dirichlet.

Η HDP παραμετροποιείται από την κατανομή βάσης $H$ που καθορίζει την εκ των προτέρων κατανομή πάνω στα δεδομένα, και από ένα πλήθος υπερπαραμέτρων συγκέντρωσης που καθορίζουν το πλήθος των συστάδων και την "σύνδεση" μεταξύ των διαφορετικών ομάδων. Συγκεκριμένα, η $j$-οστή ομάδα συσχετίζεται με το μέτρο $G_j$, κατανεμημένο σύμφωνα με την Διαδικασία Dirichlet

$$G_j \sim DP(\alpha, G_0),\ j = 1, \ldots, J,$$

ενώ το κοινό μέτρο βάσης $G_0$ είναι επίσης κατανεμημένο σύμφωνα με μία Διαδικασία Dirichlet,

$$G_0 \sim DP(\gamma, H).$$

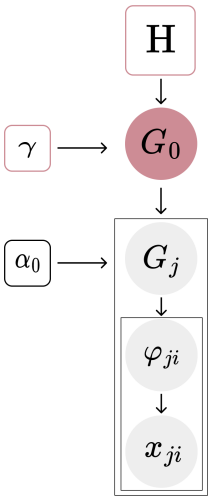
Η Ιεραρχική Διαδικασία Dirichlet μπορεί να χρησιμοποιηθεί ως πρότερη κατανομή των παραγόντων $\varphi_{ji}$, σε ομαδοποιημένα δεδομένα. Συγκεκριμένα, αν $\varphi_{ji} \overset{\text{iid}}{\sim} G_j$, όπου κάθε $\varphi_{ji}$ αντιστοιχεί στο επίπεδο του παράγοντα $\varphi$ της παρατήρησης $x_{ji}$. Τότε,

$$\begin{aligned} \varphi_{ji} &\sim G_j \\ x_{ji}|\varphi_{ji} &\sim f(\varphi_{ji}). \end{aligned}$$

Στο Διάγραμμα 3.1 παρατίθεται μια σχηματική αναπαράσταση του εν λόγω ιεραρχικού μοντέλου.

## 3.2 Κατασκευή: Stick Breaking

Όπως και στην περίπτωση της Διαδικασίας Dirichlet, έτσι και η Ιεραρχική Διαδικασία Dirichlet μπορεί να γίνει πιο εύκολα αντιληπτή μέσω της κατασκευής Stick Breaking.




Διάγραμμα 3.1: Ένα μοντέλο Ιεραρχικής Διαδικασίας Dirichlet.

Πιο συγκεκριμένα, έχουμε αφού η $G_0$ είναι κατανεμημένη σύμφωνα με την διαδικασία Dirichlet, έχουμε ότι

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\boldsymbol{\theta}_k^{**}}.$$

όπου

$$V_k \sim Beta(1, \gamma), k = 1, \ldots, \infty$$
$$\beta_k = V_k \prod_{l=1}^{k-1} (1 - V_l)$$
$$\boldsymbol{\theta}_k^{**} \sim H.$$

Αντίστοιχα, για τις $G_j, j = 1, \ldots, J$ έχουμε

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\boldsymbol{\theta}_k^{**}}, \tag{3.2.1}$$

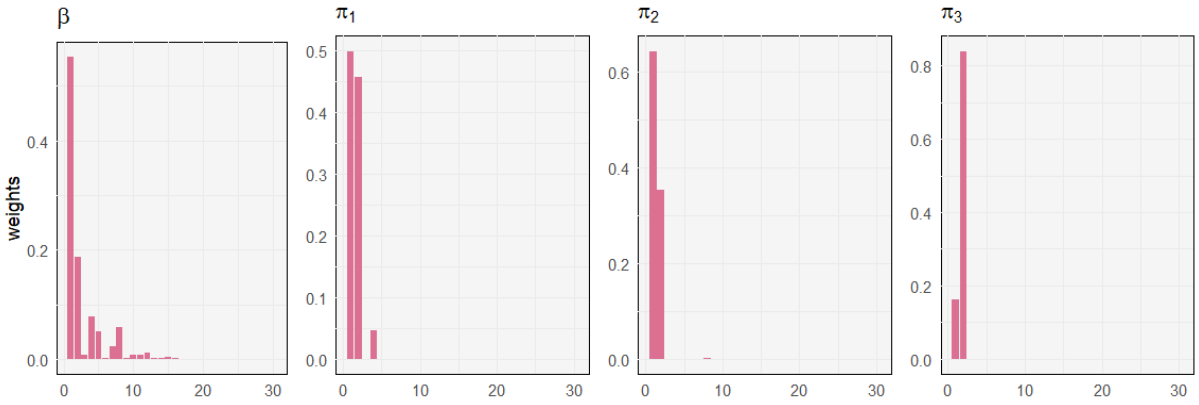
όπου για την σχέση μεταξύ των βαρών $\boldsymbol{\beta} = (\beta_1, \beta_2, \ldots)$ και $\boldsymbol{\pi} = (\pi_1, \pi_2, \ldots)$ αποδεικνύεται (Teh et al. 2006) ότι

$$V_{jk} \sim Beta\left(a\beta_k, a\left(1 - \sum_{l=1}^{k} \beta_l\right)\right)$$
$$\pi_{jk} = V_{jk} \prod_{l=1}^{k-1} (1 - V_{jl}).$$

Παρατηρούμε λοιπόν ότι κάθε κάθε πραγματοποίηση της $G_j$ "κληρονομεί", μέσω των παραμέτρων $\pi_{jh}$ χαρακτηριστικά από την κατανομή βάσης $G_0$. Αυτή η μορφή ιεραρχείας αποδεικνύεται ιδιαίτερα χρήσιμη στη μοντελοποίηση δεδομένων των οποίων οι συστάδες μοιράζονται κοινά χαρακτηριστικά. Στο Διάγραμμα 3.2 αναπαρίστανται οι τρεις κατανομές $\pi_{jk}, j = 1, 2, 3$ όπως αυτές κατασκευάστηκαν με τη μέθοδο Stick Breaking που περιγράφηκε παραπάνω.

## 3.3 Η διαδικασία κινέζικου franchise

Όπως η διαδικασία Dirichlet γίνεται πιο εύκολα αντιληπτή με την χρήση της μεταφοράς του Κινέζικου Εστιατορίου, έτσι και για την Ιεραρχική Διαδικασία Dirichlet, προτάθηκε η μεταφορά του Κινέζικου Franchise




Διάγραμμα 3.2: Δείγματα από την κατασκευή Stick Breaking για ένα HDP. Αριστερά: $\boldsymbol{\beta}$ για $\gamma = 2$ και στα επόμενα Διάγραμματα, $\boldsymbol{\pi}_1, \boldsymbol{\pi}_2, \boldsymbol{\pi}_3$ για $\alpha = 1$.

(Teh et al. 2006) που επεκτείνει την Διαδικασία Κινέζικου Εστιατορίου, επιτρέποντας πολλαπλά εστιατόρια, τα οποία όμως μοιράζονται τον ίδιο κατάλογο.

Για την παρακάτω ανάλυση, θεωρούμε ότι $\varphi_{ji} \sim G_j$, $\theta_1, \ldots, \theta_K \overset{\text{iid}}{\sim} H$ και για κάθε $j = 1, \ldots, J$, θεωρούμε $T_j$ τυχαίες μεταβλητές $\psi_{j1}, \ldots, \psi_{jT_j} \overset{\text{iid}}{\sim} G_0$. Στο πλαίσιο της μεταφοράς της Διαδικασίας Κινέζικου Franchise, φανταζόμαστε ένα Franchise $J$ εστιατορίων με κοινό κατάλογο. Σε κάθε τραπέζι, επιλέγεται ένα πιάτο του καταλόγου από τον πρώτο πελάτη που κάθεται εκεί, το οποίο και μοιράζεται από όλους τους πελάτες που κάθονται στο τραπέζι αυτό. Τα εστιατόρια αντιστοιχούν σε ομάδες του πληθυσμού, οι πελάτες αντιστοιχούν στις τυχαίες μεταβλητές $\varphi_{ji}$, τα τραπέζια στις μεταβλητές $\psi_{jt}$ ενώ τα πιάτα στις $\theta_k$.

Κάθε $\varphi_{ji}$ σχετίζεται με μία $\psi_{jt}$, ενώ κάθε $\psi_{jt}$ σχετίζεται με την σειρά της με μία $\theta_k$. Αν $t_{ji}$ ο δείκτης της μεταβλητής $\psi_{jt}$ που αντιστοιχεί στην $\varphi_{ji}$ (δηλαδή, το νούμερο του τραπεζιού που αντιστοιχεί στον εν λόγω πελάτη), και $k_{jt}$ ο δείκτης της $\theta_k$ που σχετίζεται με το $\psi_{jt}$ (δηλαδή, το πιάτο που αντιστοιχεί στο τραπέζι αυτό), ορίζουμε $m_k = \sum_j m_{jk}$ το πλήθος των $\psi_{jt}$ που σχετίζονται με το $\theta_K$ για όλα τα $j$. Επειδή $\varphi_{ji} \sim DP(\alpha, G_0)$, από την 2.4.4 έχουμε ότι

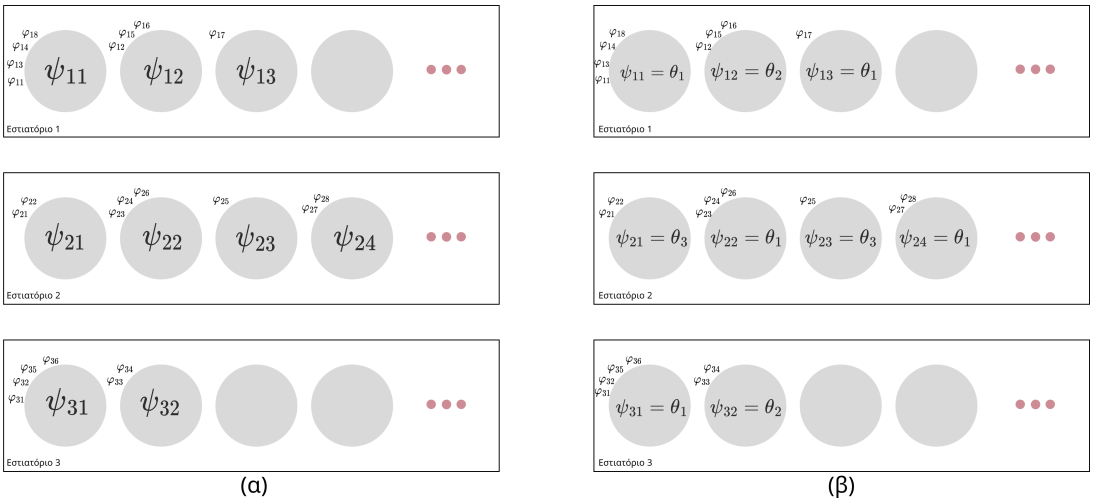
$$\phi_{ji} \mid \phi_{j1}, \ldots, \phi_{ji-1}, \alpha, G_0 \sim \sum_{t=1}^{T_j} \frac{n_{jt}}{i - 1 + \alpha} \delta_{\psi_{jt}} + \frac{\alpha}{i - 1 + \alpha} G_0. \tag{3.3.1}$$

Για να ληφθεί δείγμα από την κατανομή αυτή, την οποία μπορούμε να θεωρήσουμε ως μοντέλο μίξης, ακολουθούμε την εξής διαδικασία: με πιθανότητες που δίνονται από τα ανάλογα βάρη, διαλέγουμε έναν όρο από το δεξί μέλος της 3.3.1 . Αν επιλεχθεί ο όρος $t$ από το άθροισμα, θέτουμε $\varphi_{ji} = \psi_{jt}$ και $t_{ji} = t$. Αν επιλεγεί ο δεύτερος όρος, αυξάνουμε το $T_j$ κατά ένα, θέτουμε $\varphi_{ji} = \varphi_{jT_j} \sim G_0$ και $t_{ji} = T_j$ Όμως, επειδή $G_0 \sim DP(\gamma, H)$, χρησιμοποιώντας και πάλι την 2.4.4 μπορούμε να την απαλείψουμε από την 3.3.1, αφού

$$\psi_{jt} \mid \psi_{j1}, \psi_{j2}, \ldots, \psi_{21}, \ldots, \psi_{j,t-1}, \gamma, H \sim \sum_{k=1}^{K} \frac{m_k}{\sum_k m_k + \gamma} \delta_{\theta_k} + \frac{\gamma}{\sum_k m_k + \gamma} H. \tag{3.3.2}$$

Επομένως, ακολουθωντας όμοια διαδικασία με προηγουμένως, αν επιλέξουμε τον όρο $k$ του αθροίσματος στο δεξί μέλος της 3.3.2, θέτουμε $\psi_{jt} = \theta_k$ και $k_{jt} = k$. Αν επιλεχθεί ο δεύτερος όρος, αυξάνουμε το $K$ κατά ένα και θέτουμε $\psi_{jt} = \theta_K \sim H$ και $k_{jt} = K$.

Επιστρέφοντας στην αναλογία του Κινέζικου Franchise, ένας πελάτης που μπαίνει στο $j-$οστό εστιατόριο, κάθεται σε ένα από τα κατειλημμένα τραπέζια με ορισμένη πιθανότητα, ή σε νέο τραπέζι με πιθανότητα την υπολειπόμμενη. Αυτό είναι η γνωστή Διαδικασία Κινέζικου Εστιατορίου και αντιστοιχεί στην 3.3.1. Αν καθίσει σε ήδη κατειλημμένο τραπέζι, τρώει το πιάτο που έχει ήδη επιλεγεί σε αυτό. Αν καθίσει σε νέο τραπέζι, επιλέγει το πιάτο που θα παραγγείλει. Η επιλογή αυτή εξαρτάται από την δημοφιλία κάθε πιάτου σε όλο το Franchise, αλλά με μη μηδενική πιθανότητα δύναται να παραγγείλει και νέο πιάτο. Η διαδικασία αυτή περιγράφεται στην 3.3.2. Στο Διάγραμμα 3.3 απωτυπώνεται μια σχηματική αναπαράσταση της διαδικασίας, με και χωρίς την απαλοιφή της $G_0$.




(α) (β)

Διάγραμμα 3.3: (α): Κάθε ορθογώνο αντιστοιχεί σε ένα εστιατόριο (ομάδα). Κάθε τραπέζι σχετίζεται με μια παράμετρο $\psi_{jt} \sim G_0$ και κάθε $\varphi_{ji}$ κάθεται στο τραπέζι που του ανατίθεται στην 3.3.1. (β) απαλοίφοντας την $G_0$, ανατίθεται σε κάθε $\psi_{jt}$ ένα πιάτο.

## 3.4 Κρυφά Μαρκοβιανά Μοντέλα HDP

### 3.4.1 Μαρκοβιανές αλυσίδες

Ας ανακαλέσουμε από την θεωρία τον στοχαστικών διαδικασιών την κεντρική ιδέα πίσω από τις Μαρκοβιανές αλυσίδες: θεωρούμε ότι κάθε χρονική στιγμή $t$, η τυχαία μεταβλητή $X_t$ εμπεριέχει όλη την πληροφορία που είναι απαραίτητη για την πρόβλεψη της εξέλιξης της.

**Ορισμός 3.4.1** (Μαρκοβιανή Αλυσίδα)**.** Μια στοχαστική διαδικασία $\{X_n\}_{n\in\mathbb{N}}$ με τιμές στον $\mathbb{X}$ καλείται Μαρκοβιανή αλυσίδα, αν για κάθε $n \in \mathbb{N}$ και για κάθε $v_0, \ldots, v_{n-1}, x, y \in \mathbb{X}$ ισχύει

$$\mathbb{P}\left[X_{n+1} = y \mid X_0 = v_0, \ldots, X_{n-1} = v_{n-1}, X_n = x\right] = \mathbb{P}\left[X_{n+1} = y \mid X_n = x\right].$$

Μάλιστα, αν ακόμα η δεσμευμένη πιθανότητα $\mathbb{P}\left[X_{n+1}|X_n\right]$ είναι ανεξάρτητη του $n$, η Μαρκοβιανή αλυσίδα καλείται χρονικά ομοιογενής και μπορεί να οριστεί για αυτήν ο πίνακας μετάβασης αποτελούμενος από τις πιθανότητες μετάβασης

$$p : \mathbb{X} \times \mathbb{X} \to [0,1], p(x,y) = \mathbb{P}\left[X_{n+1} = y \mid X_n = x\right],$$

Αν πάλι ο χώρος καταστάσεων $\mathbb{X}$ είναι διακριτός, $\mathbb{X} = \{1, 2, \ldots, K\}$, ο πίνακας μετάβασης είναι ο στοχαστικός πίνακας $\mathbf{A}$ όπου $A_{ij} = \mathbb{P}\left[X_{n+1} = j | X_n = i\right]$ είναι η πιθανότητα μετάβασης από την κατάσταση $i$ στην κατάσταση $j$.

Σε ορισμένες περιπτώσεις, η υπόθεση ότι η κατάσταση $X_{n-1}$ της αλυσίδας ενσωματώνει όλη την πληροφορία της "ιστορίας" της, $\mathbf{X}_{1:n-2}$, είναι υπερβολικά ισχυρή. Για αυτό, μπορούμε να προσθέσουμε άλλο ένα επίπεδο εξάρτησης, $X_{n-2}$ έτσι ώστε η από κοινού συνάρτηση πυκνότητας πιθανότητας να έχει την ακόλουθη μορφή

$$p(\mathbf{x}_{1:T}) = p(x_1, x_2)p(x_3 \mid x_1, x_2)p(x_4 \mid x_2, x_3)\ldots = p(x_1, x_2)\prod_{n=3}^{T} p(x_t \mid x_{n-1}, x_{n-2}).$$

Τότε, η αλυσίδα ονομάζεται Μαρκοβιανή αλυσίδα *δεύτερης τάξης*. Εντούτοις, σε περιπτώσεις που υπάρχουν συσχετίσεις μεταξύ των παρατηρήσεων που θέλουμε να μελετήσουμε που εκτείνονται πολύ στο "παρελθόν" της αλυσίδας, ακόμα και οι Μαρκοβιανές αλυσίδες δεύτερης τάξης δεν αρκούν για τη μοντελοποίηση τους. Με παρόμοιο τρόπο μπορούμε να ορίσουμε Μαρκοβιανές αλυσίδες ανώτερης τάξης, όμως κάτι τέτοιο αυξάνει τις παραμέτρους του μοντέλου. Μια εναλλακτική προσέγγιση δίνεται από τα Κρυφά Μαρκοβιανά Μοντέλα (Hidden Markov Models - HMM).

### 3.4.2 Κρυφά μαρκοβιανά μοντέλα (HMM)

Ένα Κρυφό Μαρκοβιανό Μοντέλο (HMM) αποτελείται από μία Μαρκοβιανή αλυσίδα διακριτού χρόνου με πεπερασμένο χώρο "κρυφών" καταστάσεων $\Theta = \{\theta_1, \theta_2, \ldots, \theta_T\}$ και ένα μοντέλο $f_{\theta_t}(x_t) = p(\mathbf{x}_t|\theta_t)$ από το οποίο προέρχονται οι παρατηρήσεις. Η αντίστοιχη από κοινού συνάρτηση πυκνότητας πιθανότητας έχει τη μορφή

$$p(\boldsymbol{\theta}_{1:T}, \mathbf{x}_{1:T}) = p(\boldsymbol{\theta}_{1:T}) p(\mathbf{x}_{1:T} \mid \boldsymbol{\theta}_{1:T}) = \left[ p(\theta_1) \prod_{t=2}^{T} p(\theta_t \mid \theta_{t-1}) \right] \left[ \prod_{t=1}^{T} p(\mathbf{x}_t \mid \theta_t) \right].$$

Οι παρατηρήσεις $\{x_t\}_{t=1,\ldots,T}$ μπορεί να προέρχονται είτε από διακριτές είτε από συνεχείς τυχαίες μεταβλητές. Στην πρώτη περίπτωση, συνηθίζεται το μοντέλο παρατηρήσεων να δίνεται από έναν πίνακα $\mathbf{B}$,

$$p(\mathbf{x}_t = l \mid \theta_t = k, \boldsymbol{\lambda}) = \mathbf{B}(k, l).$$

Στην συνεχή περίπτωση, θεωρούμε συχνά ότι η παρατηρήσεις προέρχονται από μια κανονική κατανομή

$$p(\mathbf{x}_t \mid \theta_t = k) = \mathcal{N}(\mathbf{x}_t \mid \boldsymbol{\mu}_k, \boldsymbol{\Sigma}_k).$$

Στο Διάγραμμα 3.4 θεωρούμε ότι έχουμε μια Μαρκοβιανή αλυσίδα τριών καταστάσεων, κάθε μία εκ των οποίων αντιστοιχεί σε μία δισδιάστατη κανονική κατανομή.


Διάγραμμα 3.4: (α): 100 σημεία του επιπέδου προσομοιωμένα από ένα HMM τριών καταστάσεων. Κάθε κατάσταση αντιστοιχεί σε μία κανονική κατανομή (οι παράμετροι των κανονικών κατανομών καθώς και ο σχετικός κώδικας στην R βρίσκεται στο Παράρτημα D). (β): Αλληλουχία καταστάσεων της Μαρκοβιανής αλυσίδας.

Φυσικά, στα πραγματικά HMM μόνο οι παρατηρήσεις $\mathbf{x}$ είναι ορατές. Οι καταστάσεις της υποβόσκουσας Μαρκοβιανής αλυσίδας δεν είναι απευθείας ορατές, όπως και η δεσμευμένη συνάρτηση πυκνότητας πιθανότητας των παρατηρήσεων δεδομένης της κατάστασης. Ακόμα, δεν είναι γνωστό το πλήθος των κρυφών καταστάσεων ή οι πιθανότητες μετάβασης μεταξύ τους. Αυτές είναι και οι παράμετροι που περιγράφουν πλήρως τα HMM, που καλούμαστε να εκτιμήσουμε. Για τον σκοπό αυτό έχει προταθεί πλήθος αλγορίθμων: Ο προς τα μπρος αλγόριθμος για εκτίμηση των δεσμευμένων σ.π.π, ο αλγόριθμος Viterbi (Viterbi 1967) για τον προσδιορισμό της κρυφής ακολουθίας των καταστάσεων που οδήγησε στα παρατηρούμενα δεδομένα, και ο πιο διαδεδομένος αλγόριθμος Baum-Welch (Baum et al. 1970) για την εκτίμηση του πίνακα μετάβασης της κρυφής Μαρκοβιανής αλυσίδας αλλά και των δεσμευμένων σ.π.π. Ωστόσο, η αναλυτική περιγραφή των εν λόγω αλγορίθμων εκτείνεται πέραν του πλαισίου της παρούσας εργασίας.

### 3.4.3 Άπειρα κρυφά Μαρκοβιανά μοντέλα

Μπορούμε να επεκτείνουμε την ιδέα των HMM ώστε να επιτρέπουν έναν αριθμήσιμο άπειρο χώρο καταστάσεων, με τρόπο παρόμοιο με τη μετάβαση από τα πεπερασμένα στα άπειρα μοντέλα μίξης (Beal et al. 2001; Teh et al. 2006). Ωστόσο, το εργαλείο που θα χρησιμοποιηθεί τώρα δεν είναι η Διαδικασία Dirichlet, αλλά η Ιεραρχική Διαδικασία Dirichlet.

Για να γίνει πιο εμφανής η ανάγκη για χρήση της HDP, ας σκεφτούμε το HMM ως ένα σύνολο μοντέλων μίξης: για κάθε τιμή της τωρινής κατάστασης $\theta_t$ της αλυσίδας, η παρατήρηση $x_{t+1}$ επιλέγεται αφού πρώτα επιλεχθεί η κατάσταση $\theta_{t+1}$, από την κατανομή $f_{\theta_{t+1}}$. Επομένως, οι πιθανότητες μετάβασης $p(\theta_t, \theta_{t+1})$ αντιστοιχούν στις παραμέτρους μίξης του μοντέλου, και οι δεσμευμένες κατανομές $f_{x_t}$ αντιστοιχούν στα επιμέρους τμήματα του. Αν χρησιμοποιούσαμε την Διαδικασία Dirichlet για να γενικεύσουμε τα μοντέλα αυτά, θα λαμβάναμε ένα σύνολο DPMM, ένα για κάθε κατάσταση $\theta_t$, τα οποία όμως είναι ανεξάρτητα μεταξύ τους: Το σύνολο των καταστάσεων προσβάσιμων από την τιμή της τωρινής κατάστασης είναι ξένο από το σύνολο των καταστάσεων προσβάσιμων από μια άλλη τιμή της τωρινής κατάστασης. Η δομή που θα δημιουργούταν έτσι, θα θύμιζε περισσότερο δέντρο παρά αλυσίδα. Το πρόβλημα αυτό επιλύεται χρησιμοποιώντας την Ιεραρχική Διαδικασία Dirichlet.

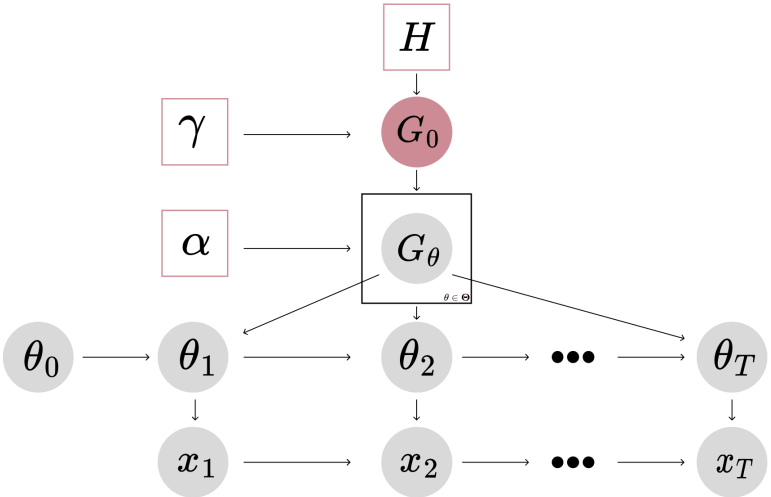
Πιο συγκεκριμένα, έστω $\{G_\theta : \theta \in \Theta\}$, από την HDP

$$G_0|\gamma, H \sim DP(\gamma, H)$$
$$G_\theta|\alpha, G_0 \sim DP(\alpha, G_0),\ \theta \in \Theta.$$

Όπως θα δούμε στην πορεία, το μέτρο βάσης $G_0$ επιτρέπει στις μεταβάσεις από κάθε κατάσταση να μοιράζονται προσβάσιμες καταστάσεις. Έστω επίσης $\theta_0 = \theta_0^{**} \in \Theta$ η προκαθορισμένη αρχική κατάσταση. Τότε, οι δεσμευμένες κατανομές της ακολουθίας των λανθανουσών μεταβλητών $\theta_1, \ldots, \theta_T$ και των παρατηρούμενων μεταβλητών $x_1, \ldots, x_T$ δίνονται από τις

$$\theta_t \mid \theta_{t-1}, G_{\theta_{t-1}} \sim G_{\theta_{t-1}},\ t = 1, \ldots, T$$
$$x_t \mid \theta_t \sim f_{\theta_t}.$$

Το μοντέλο αυτό ονομάζεται Hierarchical Dirichlet Process Hidden Markov Model (HDP HMM) και περιγράφεται σχηματικά στο Διάγραμμα 3.5.




Διάγραμμα 3.5: Ένα μοντέλο HDP HMM.

Στην πραγματικότητα, τα μονοπάτια του HDP HMM μπορούν να εμφανίσουν πεπερασμένο μόνο πλήθος καταστάσεων. Αυτό, καθώς και η σχέση μεταξύ του παραμετρικού HMM και του HDP HMM, καθίσταται φανερό αν λάβουμε υπόψιν την Stick Breaking κατασκευή του τελευταίου. Πράγματι, σύμφωνα με τις 2.4.5 και 3.2.1, θα ισχύει

$$G_0 = \sum_{k=1}^{\infty} \beta_k \delta_{\theta_k^{**}}$$
$$G_{\theta_l^{**}} = \sum_{k=1}^{\infty} \boldsymbol{\pi}_{\theta_{l,k}^{**}} \delta_{\theta_k^{**}} \quad l = 0, 1, \ldots, \infty,$$

όπου

$$\phi_k^{**} \mid H \sim H \quad k = 1, \ldots, \infty$$
$$\boldsymbol{\beta} \mid \gamma \sim \text{GEM}(\gamma)$$
$$\pi_{\theta_k^{**}} \mid \alpha, \boldsymbol{\beta} \sim \text{DP}(\alpha, \boldsymbol{\beta}).$$

Συγκρίνοντας με το HMM, η πιθανότητα μετάβασης από την κατάσταση $\theta_l^{**}$ στην κατάσταση $\theta_k^{**}$ δίνεται από την $\pi_{\theta_{l,k}^{**}}$, ενώ η κατανομή των παρατηρήεων δίνεται από την $f_{\theta_k^{**}}$. Αν μάλιστα ταυτίσουμε την κατάσταση $\theta_k^{**}$ με τον ακέραιο $k$, για $k = 0, 1, \ldots, \infty$, και ορίσουμε τις κατηγορικές μεταβλητές $z_t$ που αντιστοιχούν στην κατάσταση που βρίσκεται η αλυσίδα την στιγμή $t$, η σχέση με το HMM γίνεται πιο προφανής. Συγκεκριμένα, αν $\theta_t = \theta_k^{**}$ η κατάσταση την στιγμή $t$, τότε $z_t = k$ και $\boldsymbol{\pi}_k = \boldsymbol{\pi}_{\theta_k^{**}}$, μπορούμε να εκφράσουμε το HDP HMM ως

$$z_t \mid z_{t-1}, \boldsymbol{\pi}_{z_{t-1}} \sim \boldsymbol{\pi}_{z_{t-1}}$$
$$x_t \mid z_t, \theta_{z_t}^{**} \sim f_{\theta_{z_t}^{**}}.$$

Η αναπαράσταση αυτή επιτρέπει άμεσα να ερμηνεύσουμε το HDP HMM ως ένα HMM με αριθμήσιμο άπειρο χώρο καταστάσεων.

## 3.5 Εφαρμογή: Θεματικά Μοντέλα (Topic Models)

Η Μοντελοποίηση Θεμάτων (Topic Modelling) αποτελεί ένα είδος στατιστικής μοντελοποίησης που σκοπό έχει την αναγνώριση κοινών θεμάτων μέσα σε ένα σύνολο κειμένων και έγκειται στην ομαδοποίηση (συσταδοποίηση) λέξεων που τείνουν να συνυπάρχουν συχνά. Στην παρούσα ενότητα της εργασίας, θα χρησιμοποιήσουμε τη μέθοδο της Ιεραρχικής Διαδικασίας Dirichlet για να "ανακαλύψουμε" τα υποκείμενα θέματα σε ένα σύνολο άρθρων από την ιστοσελίδα της εφημερίδας *Καθημερινή*. Το μοντέλο που θα προσαρμόσουμε ονομάζεται θεματικό μοντέλο HDP (HDP topic model).

Ένα θεματικό μοντέλο HDP αποτελείται από ένα σύνολο κειμένων, που καλείται συνήθως σώμα (corpus). Το σώμα με την σειρά του απαρτίζεται από τα επιμέρους κείμενα $d = 1, \ldots, D$ (documents), τα οποία απαρτίζονται από τις λέξεις $\boldsymbol{w}_d$. Για κάθε κείμενο $d$ έστω $\boldsymbol{\theta}_d$ το διάνυσμα των βαρών μίξης θεμάτων. Ακόμα, για κάθε θέμα $k$, έστω $\boldsymbol{\varphi}_k$ το διάνυσμα των πιθανοτήτων για κάθε λέξη στο θέμα. Τότε, οι λέξεις σε κάθε κείμενο επιλέγονται ως εξής (Teh et al. 2007):

- Με πιθανότητα $\theta_{dk}$ επιλέγεται το θέμα $k$.
- Με πιθανότητα $\varphi_{kw}$ επιλέγεται η λέξη $w$.

Έστω $x_{id}$ η ετικέτα της $i$-οστής λέξης στο κείμενο d και $z_{id}$ το επιλεχθέν θέμα της. Τότε,

$$z_{id}|\boldsymbol{\theta}_d \sim \text{Categorical}(\boldsymbol{\theta}_d)$$
$$x_{id}|z_{id}, \boldsymbol{\phi}_{z_{id}} \sim \text{Categorical}(\boldsymbol{\phi}_{z_{id}}).$$

Για τις παραμέτρους $\boldsymbol{\theta}_d$ και $\boldsymbol{\phi}_k$ χρησιμοποιούμε πρότερες κατανομής Dirichlet,

$$\boldsymbol{\theta}_d|\boldsymbol{\pi} \sim Dir(\alpha\boldsymbol{\pi})$$
$$\boldsymbol{\phi}_k|\boldsymbol{\tau} \sim Dir(\beta\boldsymbol{\tau}),$$

όπου $\boldsymbol{\pi}$ η κατανομή των θεμάτων σε όλο το σώμα κειμένων και $\boldsymbol{\tau}$ η κατανομή των λέξεων σε όλο το σώμα κειμένων, ενώ $\alpha, \beta$ είναι παράμετροι συγκέντρωσης.

Αν το πλήθος θεμάτων $K$ είναι γνωστό και πεπερασμένο, η παραπάνω προσέγγιση ονομάζεται LDA (Latent Dirichlet Allocation). Ωστόσο, συνήθως αυτό δεν είναι ρεαλιστικό και θα θέλαμε ιδανικά ένα μοντέλο που καθορίζει αυτόματα το πλήθος $K$ των θεμάτων. Στην περίπτωση αυτή, έχουμε εν δυνάμη αριθμήσιμο άπειρο πλήθος θεμάτων και τα $\boldsymbol{\theta}_d, \boldsymbol{\pi}$ είναι απειροδιάστατα διανύσματα. Χρησιμοποιώντας την συνήθη stick-breaking κατασκευή για το $\boldsymbol{\pi}$, παίρνουμε

$$\pi_k = V_k \prod_{l=1}^{k-1} (1 - V_l)$$
$$V_k \sim Beta(1, \gamma), k = 1, \ldots, \infty.$$

Ισοδύναμα, $G_d \sim DP(\alpha, G_0)$ και $G_0 \sim DP(\gamma, Dir(\beta\boldsymbol{\tau}))$, όπου $G_d = \sum_{k=1}^{\infty} \theta_{dk}\delta_{\varphi_k}$ και $G_0 = \sum_{k=1}^{\infty} \pi_k\delta_{\varphi_k}$.

### 3.5.1 Συλλογή κειμένων

Για την συλλογή των κειμένων στο πείραμα μας χρησιμοποιήθηκε η Βιβλιοθήκη "BeautifulSoup" της Python, επιτρέπει την εξαγωγή δεδομένων από αρχεία HTML, αναγνωρίζοντας τις κατάλληλες ετικέτες HTML που περιέχουν τα κείμενα των άρθρων και αποθηκεύοντας τα σε λίστα για περαιτέρω ανάλυση. Συλλέχθηκαν συνολικά 90 άρθρα πολιτικού περιεχομένου που δημοσιεύτηκαν στο διάστημα 31 Αυγούστου-9 Σεπτεμβρίου 2024.

### 3.5.2 Προεπεξεργασία των δεδομένων

Μετά την συλλογή των κειμένων, ακολούθησε η προεπεξεργασία τους. Η προεπεξεργασία περιλαμβάνει τον καθαρισμό των κειμένων από ανεπιθύμητους χαρακτήρες, την μετατροπή των λέξεων σε πεζά γράμματα και την αφαίρεση των κοινών λέξεων (stopwords) που δεν προσδίδουν σημαντική πληροφορία στο μοντέλο. Τέτοιες λέξεις είναι, για παράδειγμα, άρθρα, προθέσεις, ρήματα όπως "είναι, βρίσκεται, δήλωσε", σύνδεσμοι κ.α.

### 3.5.3 Εκπαίδευση μοντέλου HDP

Στη συνέχεια, δημιουργήθηκε ένα λεξικό που περιέχει όλους τους μοναδικούς όρους στα κείμενα και ένα σώμα κειμένων που αποτελείται από τα κείμενα σε μορφή bag-of-words, όπου κάθε κείμενο αναπαριστάται από τις συχνότητες εμφάνισης των όρων του. Για την εξαγωγή των θεμάτων από τα κείμενα χρησιμοποιήθηκε το μοντέλο Hierarchical Dirichlet Process (HDP) της βιβλιοθήκης *gensim*. Το μοντέλο εκπαιδεύτηκε με το σώμα κειμένων και το λεξικό που δημιουργήθηκαν προηγουμένως. Η ανανέωση των παραμέτρων του μοντέλου σε κάθε βήμα, γίνεται με τη μέθοδο Online Variational Inference (Wang et al. 2011). Ακόμα, για την επιλογή των υπερπαραμέτρων $\gamma, \alpha$, χρησιμοποιήθηκε το πλέγμα $\alpha : \{0.1, 1, 10\}, \gamma : \{0.1, 1, 10\}$. Το μοντέλο εκπαιδεύτηκε με όλα τα πιθανά ζεύγη και ως τελικό, επιλέχθηκε αυτό με τη μεγαλύτερη τιμή της μετρικής *coherence* (συνοχή), η οποία αξιολογεί τη συνοχή των θεμάτων εξετάζοντας πόσο συχνά οι λέξεις που απαρτίζουν ένα θέμα συνυπάρχουν στα έγγραφα. Πιο συγκεκριμένα, η μετρική coherence χρησιμοποιείται συνήθως στην ανάλυση της σχέσης μεταξύ δύο συνόλων δεδομένων, ή για τον προσδιορισμό της ομοιότητας τους. Στη Μοντελοποίηση Θεμάτων, η συνοχή συγκρίνει την σημειολογική ομοιότητα μεταξύ επαναλαμβανόμεων λέξεων μέσα στα θέματα. Υπολογίζεται συχνά ως (Mimno et al. 2011)

$$c_v = \frac{2}{M(M-1)} \sum_{m=2}^{M} \sum_{l=1}^{m-1} \log \frac{p(w_m, w_l) + \varepsilon}{p(w_l)},$$

όπου $M$ το πλήθος των πιο συχνών λέξεων μέσα σε κάθε κείμενο, $p(w_l)$ η συχνότητα της λέξης $w_l$ και $p(w_m, w_l)$ η πιθανότητα οι λέξεις $w_m, w_l$ να συνυπάρχουν μέσα σε ένα κείμενο. Παίρνει τιμές στο $[0, 1]$ με τις καλύτερες τιμές (υψηλή ομοιότητα) να είναι αυτές κοντά στο 1. Βέλτιστες υπερπαράμετροι από το πλέγμα αυτό ήταν οι $\alpha = 1, \gamma = 0.1$ με τιμή coherence $c_v = 0.734$

### 3.5.4 Αποτελέσματα

Η παραπάνω διαδικασία εξάγει τα πιο σημαντικά θέματα από το σύνολο των κειμένων και δημιουργεί σύννεφα λέξεων για κάθε θέμα, όπου το μέγεθος κάθε λέξης είναι ανάλογο της σημαντικότητάς της στο θέμα. Τα αποτελέσματα αυτά μπορούν να χρησιμοποιηθούν για την ανάλυση του περιεχομένου και την εξαγωγή συμπερασμάτων για τα κυρίαρχα θέματα σε ένα σύνολο άρθρων. Παρατηρούμε στο Διάγραμμα 3.6 ότι η διαδικασία αυτή κατάφερε πράγματι να ανακαλύψει τα κύρια θέματα μέσα στο σύνολο των 90 άρθρων.


Διάγραμμα 3.6: Τα 6 πιο σημαντικά θέματα, όπως αυτά ανακαλύφθηκαν μέσω της HDP. Κάθε θέμα αναπαρίσταται ως ένα σύννεφο λέξεων, όπου το μέγεθος κάθε λέξης είναι ανάλογο της συχνότητας εμφάνισης της στο εκάστοτε θέμα.

# Κεφάλαιο 4

# Η Εξαρτημένη Διαδικασία Dirichlet

Μια υπόθεση που κάπως "βουβά" κάναμε στην ανάλυση που αφορούσε στην Διαδικασία Dirichlet, είναι ότι οι παρατηρήσεις ενδιαφέροντος είναι ανταλλάξιμες (exchangable). Δηλαδή, όλα τα υποσύνολά τους, με οποιαδήποτε διάταξη μοιράζονται την ίδια από κοινού κατανομή. Ωστόσο, σε πλήθος εφαρμογών ενδιαφερόμαστε για τη μοντελοποίηση δεδομένων που εξελίσσονται με τον χρόνο. Τότε, φυσικά η υπόθεση της ανταλλαξιμότητας δε μπορεί να ισχύει. Για αυτό, προτάθηκε (MacEachern 1999) η Εξάρτημένη Διαδικασία Dirichlet (Dependent Dirichlet Process - DDP) ως μία μη-παραμετρική πρότερη σε χρονικά εξελισσόμενα μοντέλα μίξης.

## 4.1 Κατασκευή

Μέχρι τώρα έχουμε ασχοληθεί κυρίως με προβλήματα όπου αναθέτουμε μία μη-παραμετρική πρότερη σε μία μόνο κατανομή. Ωστόσο, πολλές φορές απαιτείται η μοντελοποίηση μίας συλλογής κατανομών, $\mathcal{G} = \{G_{\boldsymbol{s}} : \boldsymbol{s} \in S\}$, όπου για κάθε $\boldsymbol{s} \in S$ η $G_{\boldsymbol{s}}$ είναι μία κατανομή πιθανότητας. Για παράδειγμα, το $S$ μπορεί να είναι χρονικό διάστημα ή μία περιοχή του χώρου. Αν θεωρήσουμε ότι οι $G_{\boldsymbol{s}}$ ταυτίζονται, δηλαδή αν $G \equiv DP(\alpha, G_0)$ για κάθε $s$, αυτό είναι περιοριστικό. Αν πάλι θεωρήσουμε ότι $\{G_{\boldsymbol{s}}\}_{s\in S} \overset{\text{i.i.d}}{\sim} DP(\alpha, G_0)$ έχουμε πιο κοστοβόρους υπολογισμούς. Η DDP επιχειρεί να βρει μία "μέση λύση", εισάγοντας εξάρτηση μεταξύ των $G_s$ μέσω της τροποποίησης των επιμέρους τμημάτων της κάθε DP, δηλαδή των βαρών $\pi_h$ και των ατόμων $\delta_{\boldsymbol{\theta}_h}$ (Kottas and Krnjajić 2009).

Ας ανακαλέσουμε την κατασκευή Stick Breaking της Διαδικασίας Dirichlet, σύμφωνα με την οποία,

$$G \sim DP(\alpha, G_0)$$

αν και μόνο αν μπορούμε να γράψουμε

$$G(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h}(\cdot)$$

όπου

$$\pi_h = V_h \prod_{l<h}(1 - V_l),\ V_h \sim Beta(1, \alpha),\ \boldsymbol{\theta}_h \sim G_0.$$

Ας θεωρήσουμε τώρα μία συλλογή διαδικασιών, $\mathcal{G} = \{G_{\boldsymbol{s}} : \boldsymbol{s} \in S\}$, ορίζουμε

$$G_{\boldsymbol{s}}(\cdot) = \sum_{h=1}^{\infty} \pi_h(\boldsymbol{s}) \delta_{\boldsymbol{\theta}_h(\boldsymbol{s})}(\cdot), \tag{4.1.1}$$

όπου $\boldsymbol{\theta}_1(\boldsymbol{s}), \boldsymbol{\theta}_2(\boldsymbol{s}), \ldots$ είναι μονοπάτια της στοχαστικής διαδικασίας $G_{0,\boldsymbol{s}}$ και $V_1(\boldsymbol{s}), V_2(\boldsymbol{s}), \ldots$ είναι μονοπάτια μίας στοχαστικής διαδικασίας στο $S$ τέτοιας ώστε $V_h(\boldsymbol{s}) \sim Beta(1, \alpha(\boldsymbol{s}))$.

Κύριο αποτέλεσμα της παραπάνω κατασκευής, είναι ότι οι περιθώριες κατανομές των παραπάνω διαδικασιών, είναι κατανεμημένες σύμφωνα με την Διαδικασία Dirichlet. Δηλαδή, για κάθε $\boldsymbol{s} \in S$, η $G_{\boldsymbol{s}}$ ακολουθεί την Διαδικασία Dirichlet.

### 4.1.1 DDP Κοινών Βαρών

Ο MacEachern ασχολήθηκε με μία είδική περίπτωση της DDP, αυτήν των Κοινών Βαρών (Single-p DDP), που κατ'αναλογία με την 4.1.1 ορίζεται ως

$$G_{\boldsymbol{s}}(\cdot) = \sum_{h=1}^{\infty} \pi_h \delta_{\boldsymbol{\theta}_h(s)}(\cdot), \tag{4.1.2}$$

όπου τα βάρη $\pi_h = V_h \prod_{l<h}(1 - V_l)$ είναι κοινά για κάθε $\boldsymbol{s}$, με $V_h \sim Beta(1, \alpha)$. Η εκδοχή αυτή της DDP παραμένει η πιο συνήθης επιλογή, καθώς η εκ των υστέρων προσομοίωση γίνεται με τον ίδιο τρόπο με την Διαδικασία Dirichlet. Η εξάρτηση στα Single-p DDP επάγεται από την εξάρτηση των $\delta_{\boldsymbol{\theta}_h(\boldsymbol{s})}$. Ωστόσο, ένα μειονέκτημα των μοντέλων αυτών είναι η αποτυχία μοντελοποίησης της τοπικής εξάρτησης που παρουσιάζεται συχνά σε χρονικά ή χωρικά δεδομένα. Το μοντέλο αυτό βρίσκει εφαρμογές (ενδεικτικά) στην Βιοστατιστική, σε αναλύσεις κλινικών μελετών όπου η μεταβλητή απόκρισης μετράται σε διαφορετικές χρονικές περιόδους ή ανάμεσα σε διάφορες ομάδες (De Iorio et al. 2004), αλλά και σε ανάλυση περιβαλλοντολογικών δεδομένων με χρονικούς και χωρικούς παράγοντες (Kottas et al. 2012).

### 4.1.2 DDP Κοινών Ατόμων

Εναλλακτικά, μπορούμε να σταθεροποιήσουμε τα *άτομα* (atoms) $\delta_{\boldsymbol{\theta}_h}$, οδηγούμενοι έτσι στην DDP Κοινών Ατόμων (Common-Atoms DDP) που ορίζεται ως

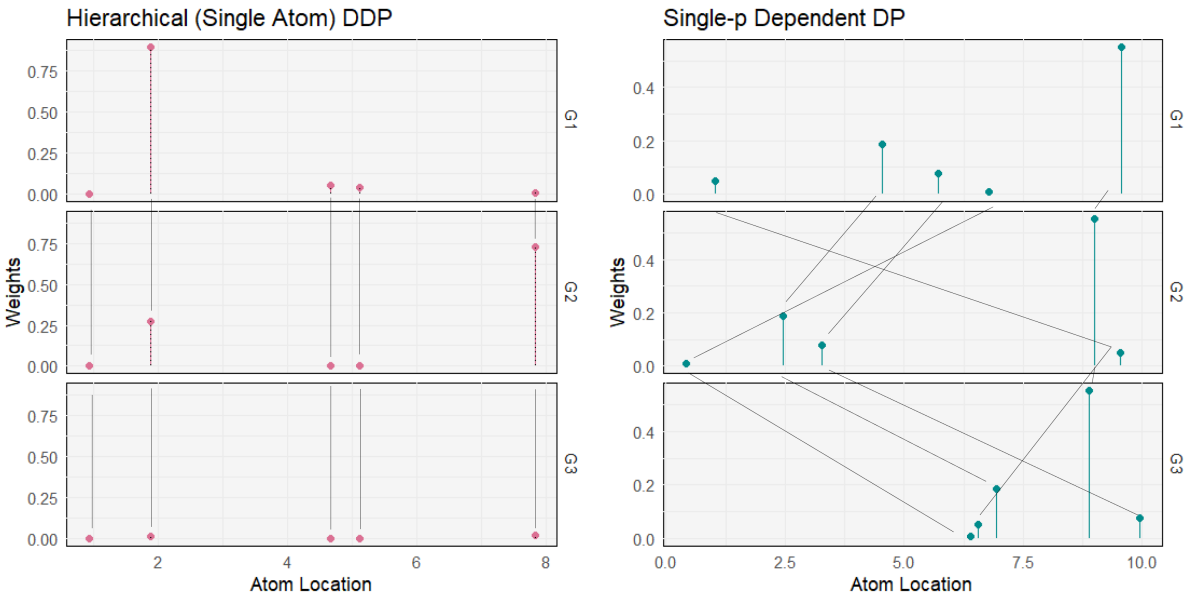
$$G_{\boldsymbol{s}}(\cdot) = \sum_{h=1}^{\infty} \pi_h(\boldsymbol{s}) \delta_{\boldsymbol{\theta}_h}(\cdot), \tag{4.1.3}$$

όπου κατά τα γνωστά,

$$\boldsymbol{\theta}_h \sim G_0.$$

Η εκδοχή αυτή της DDP έχει εφαρμοστεί σε δεδομένα εγκληματολογίας, για την πρόβλεψη γεωγραφικών σημείων με αυξημένο ποσοστό εγκληματικότητας (Taddy 2010) αλλά και σε κοινωνιολογικά δεδομένα όπου η μεταβλητή απόκρισης είναι κατηγορική μεταβλητή διάταξης (όπως οι απαντήσεις σε ερωτηματολόγια) (Kottas et al. 2005). Μάλιστα, η HDP με την οποία ασχοληθήκαμε στο προηγούμενο κεφάλαιο, μπορεί να θεωρηθεί ως ειδική περίπτωση της Single-Atom DDP.

Στο Διάγραμμα 4.1 σκιαγραφείται η διαφορά μεταξύ των δύο μοντέλων. Στην πρώτη περίπτωση (Single-Atom) παρατηρούμε ότι τα άτομα βρίσκονται στις ίδιες θέσεις, σε όλα τα δείγματα που παίρνουμε. Αντίθετα, στην περίπτωση Single-p, οι θέσεις των ατόμων διαφοροποιούνται. Ωστόσο, τα βάρη είναι κοινά για κάθε ομάδα $j = 1, 2, 3$.




Διάγραμμα 4.1: Αριστερά: Single-atom DDP. Δεξιά: Single-p DDP. Βλ. Παράρτημα G για τον αντίστοιχο κώδικα.

# Κεφάλαιο 5

# Γκαουσιανές Διαδικασίες

Οι Γκαουσιανές Διαδικασίες (Rasmussen and Williams 2006; O'Hagan and Kingman 1978) αποτελούν ένα ευέλικτο, μη-παραμετρικό εργαλείο για την μοντελοποίηση συναρτήσεων από δεδομένα. Όπως όλες οι Μπεϋζιανές μέθοδοι, δεν παρέχουν απλά μία μεμονωμένη εκτίμηση καλύτερης προσαρμογής, αλλά μια ύστερη κατανομή πάνω στις πιθανές συναρτήσεις. Όπως υποδεικνύει και το όνομα τους, οι Γκαουσιανές Διαδικασίες επεκτείνουν την κανονική κατανομή στον χώρο των συναρτήσεων. Γενικά, μπορούμε να φανταστούμε την Γκαουσιανή Διαδικασία ως μια απειροδιάστατη κατανομή πάνω σε συναρτήσεις (κατά αντιστοιχία με την Διαδικασία Dirichlet, την οποία σκεφτόμαστε ως μια κατανομή πάνω σε κατανομές).

## 5.1 Η Γκαουσιανή διαδικασία

Στα προηγούμενα κεφάλαια ασχοληθήκαμε κατά κύριο λόγο με την Διαδικασία Dirichlet, στην οποία για κάθε πεπερασμένη διαμέριση $B_1, \ldots, B_n$ ισχύει ότι αν

$$G \sim DP(\alpha, G_0),$$

τότε

$$(G(B_1), \ldots, G(B_K)) \sim Dir(\alpha G_0(B_1), \ldots, \alpha G_0(B_K)).$$

Μια ακόμα στοχαστική διαδικασία που χρησιμοποιείται σε πολλές εφαρμογές ως πρότερη, είναι η Γκαουσιανή Διαδικασία, της οποίας οι κατανομές πεπερασμένης διάστασης είναι κανονικές. Πιο συγκεκριμένα, συμβολίζουμε

$$f \sim GP(m, k),$$

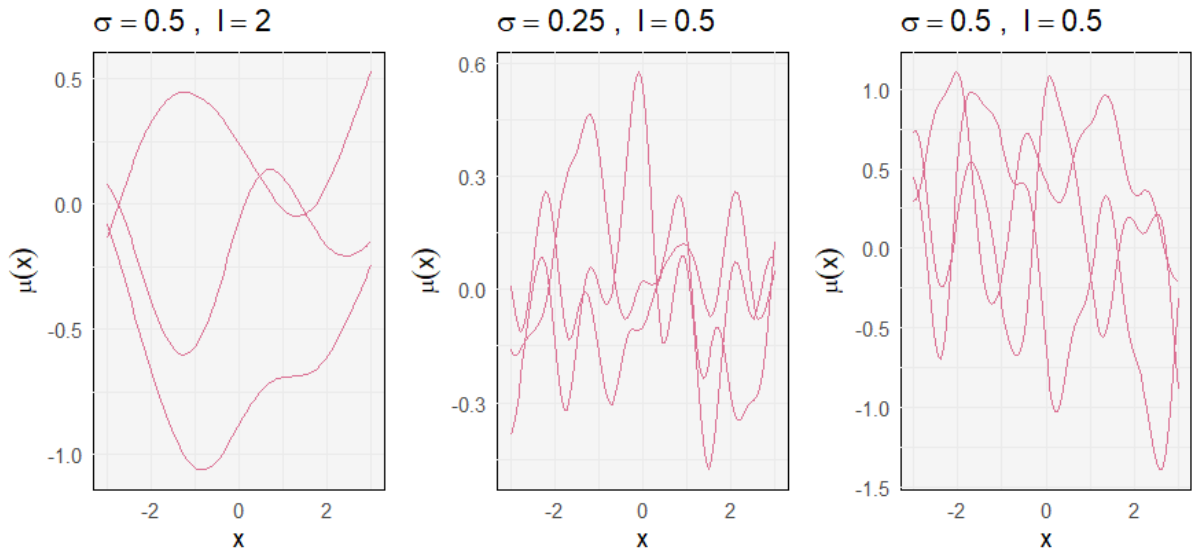
αν και μόνο αν για κάθε συλλογή $\{x_i\}_{i=1,\ldots,n}$ ισχύει

$$f(x_1), \ldots, f(x_n) \sim \mathcal{N}\left((m(x_1), \ldots, m(x_n)), K(x_1, \ldots, x_n)\right). \tag{5.1.1}$$

Όπως μια κανονική (Γκαουσιανή) κατανομή χαρακτηρίζεται πλήρως από τη μέση τιμή της και τον πίνακα διασπορών-συνδιασπορών, έτσι και μια Γκαουσιανή Διαδικασία ορίζεται πλήρως από τη μέση τιμή της $\boldsymbol{m}$ και τον πίνακα με στοιχεία $K_{ij} = k(x_i, x_j)$. Η συνάρτηση συνδιασποράς (πυρήνας) $k$ ελέγχει την ομαλότητα των πραγματοποιήσεων της Γκαουσιανής Διαδικασίας αλλά και το κατά πόσο αυτές αποκλίνουν από το μέσο. Μία συνήθης επιλογή είναι η τετραγωνική εκθετική συνάρτηση συνδιασποράς,

$$k(x_i, x_j) = \sigma^2 exp\left(-\frac{|x_i - x_j|^2}{l^2}\right),$$

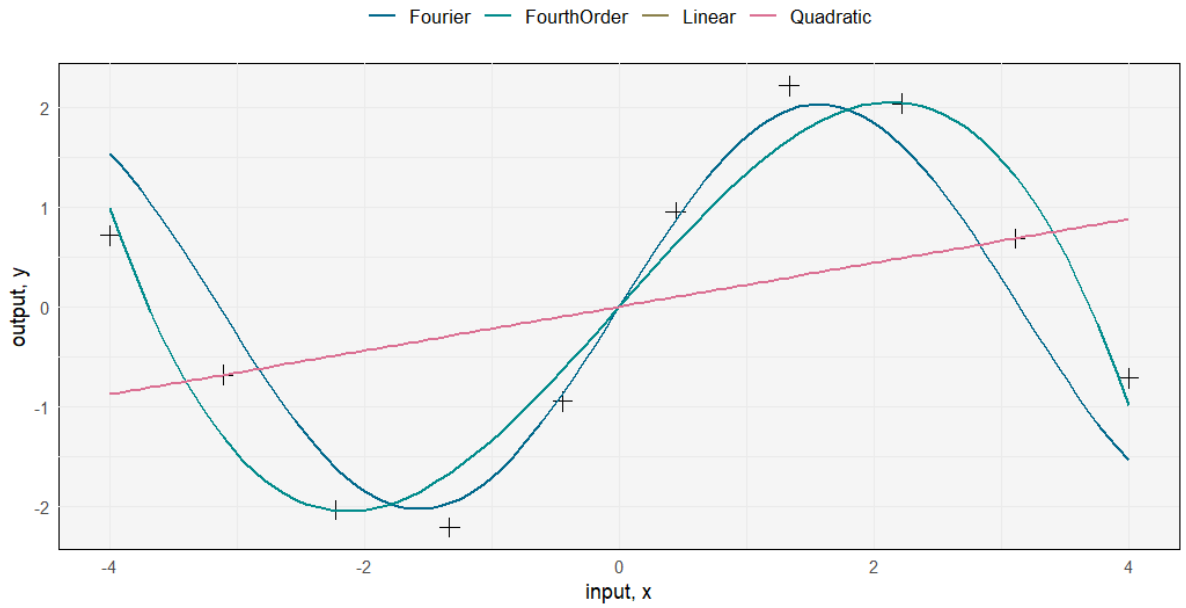
όπου $\sigma, l$ είναι παράμετροι προς εκτίμηση. Η παράμετρος $\sigma$ αντικατοπτρίζει την απόκλιση από το μέσο ενώ η παράμετρος $l$ την ομαλότητα της συνάρτησης. Στο Διάγραμμα 5.1 παρουσιάζονται τρείς διαφορετικές πραγματοποιήσεις της Γκαουσιανής Διαδικασίας με τετραγωνική κανονική συνάρτηση συνδιασποράς, για διαφορετικές τιμές των παραμέτρων $l, \sigma$.




Διάγραμμα 5.1: Τυχαία δείγματα από μία Γκαουσιανή Διαδικασία με τετραγωνική εκθετική συνάρτηση διασποράς, για διάφορες τιμές των παραμέτρων $l, \sigma$. Παρατηρούμε ότι μικρότερες τιμές της παραμέτρου $l$ οδηγούν σε λιγότερο ομαλές συναρτήσεις, ενώ μικρότερες τιμές της παραμέτρου $\sigma$ ωθούν την συνάρτηση προς το μέσο. (βλ. Παράρτημα 4)

## 5.2 Μη-γραμμική παλινδρόμηση με Γκαουσιανές διαδικασίες

Ένα σύνηθες πρόβλημα που εμφανίζεται σε εφαρμογές μοντελοποίησης δεδομένων, είναι ο προσδιορισμός μιας συνάρτησης $f(\boldsymbol{x})$, δεδομένων τιμών $\boldsymbol{y}$, της μεταβλητής απόκρισης για διάφορες τιμές των επεξηγηματικών τιμών $\boldsymbol{x}$. Φυσικά, οι παρατηρήσεις αυτές ενδέχεται να συνοδεύονται από κάποιον θόρυβο (σφάλμα). Μια απλή προσέγγιση στο πρόβλημα αυτό δίνει η προσαρμογή μιας συνάρτησης από κάποια προεπιλεγμένη κλάση συναρτήσεων (π.χ, γραμμικές ή πολυωνυμικές) με τη μέθοδο ελαχιστοποίησης του μέσου τετραγωνικού σφάλματος μεταξύ των προβλεπόμενων τιμών του εκάστοτε μοντέλου και των πραγματικών, παρατηρούμενων τιμών. Για παράδειγμα, στο Διάγραμμα 5.2 παρουσιάζονται τέσσερις διαφορετικές καμπύλες που προσαρμόζουμε σε ένα σύνολο δεδομένων προερχόμενων από μία "κρυφή" διαδικασία. Εδώ, τα εν λόγω δεδομένα είναι σημεία της $y = 2\sin(x) + 0.2x + \mathcal{N}(0, 0.3)$.




Διάγραμμα 5.2: Τέσσερις ενδεικτικές προσαρμοσμένες συναρτήσεις για ένα σύνολο παρατηρούμενων δεδομένων (μαύροι σταυροί).

Το πρόβλημα της απλής αυτής προσέγγισης έγκειται πρώτον, στην επιλογή της κατάλληλης κλάσης συναρτήσεων για τα δεδομένα με τα οποία εργαζόμαστε, ωστέ να αποφύγουμε το πρόβλημα της υπερπροσαρμογής αλλά και να περιγράψουμε επαρκώς την έμφυτη μεταβλητότητα των δεδομένων. Εμφανίζεται δηλαδή το σύνηθες πρόβλημα του Bias-Variance Tradeoff. Ωστόσο, ακόμα και αν επιχειρήσουμε να ξεπεράσουμε το πρόβλημα αυτό, επιστρατεύοντας για παράδειγμα τεχνικές cross-validation για να επιλέξουμε την κλάση μοντέλων με την βέλτιστη ικανότητα γενίκευσης, υπάρχει ένα ακόμα πρόβλημα εγγενές της μεθόδου αυτής: επιστρέφει μία και μόνο εκτίμηση, χωρίς να αναπαριστά την αβεβαιότητα του τελικού μοντέλου. Αν χρησιμοποιήσουμε το τελικό μοντέλο για προβλέψεις, αναμένουμε κακή επίδοση χωρίς κάποιο μέτρο που να ποσοτικοποιεί την "πίστη" μας στις προβλέψεις αυτές.

Ξεπερνάμε και τα δύο αυτά προβλήματα υιοθετώντας μια Μπεϋζιανή προσέγγιση. Κατά τα γνωστά, αναθέτουμε μία πρότερη κατανομή στην ζητούμενη συνάρτηση $f$, έστω $p(f)$ και έχουμε για την ύστερη κατανομή,

$$p(f|\boldsymbol{x},\boldsymbol{y}) = \frac{p(\boldsymbol{y}|f,\boldsymbol{x})p(f|\boldsymbol{x})}{p(\boldsymbol{y}|\boldsymbol{x})}.$$

Όσον αφορά το πρόβλημα της επιλογής κατάλληλης κλάσης συναρτήσεων, στρεφόμαστε στις Γκαουσιανές Διαδικασίες.

Έστω λοιπόν ένα σύνολο $N$ παρατηρήσεων $\{y_i\}_{i=1,\dots,N}$ που αντιστοιχούν σε ένα σύνολο $D-$διάστατων $\{\boldsymbol{x}_i\}_{i=1,\dots,N}$. Συμβολίζουμε τις τιμές των επεξηγηματικών μεταβλητών με το $N \times D$ πίνακα $X$, ενώ τις παρατηρήσεις της μεταβλητής απόκρισης με το $N \times 1$ διάνυσμα $\boldsymbol{y}$. Στόχος μας, είναι να προσδιορίσουμε την ύστερη κατανομή $p(\boldsymbol{f}|X,\boldsymbol{y})$, χρησιμοποιώντας για τις τιμές της $\boldsymbol{f}$ μια πρότερη Γκαουσιανής Διαδικασίας (GP). Δηλαδή, σύμφωνα με την 5.1.1,

$$p(\boldsymbol{f}|X) = \mathcal{N}(m(X), k(X,X)). \tag{5.2.1}$$

Η υπόθεση που θα κάνουμε στη συνέχεια της εργασίας θα είναι ότι τα δεδομένα έχουν προσθετικό, κανονικά κατανεμημένο θόρυβο, δηλαδή

$$y = f(\boldsymbol{x}) + \varepsilon \tag{5.2.2}$$

$$\varepsilon \sim \mathcal{N}(0, \sigma_\varepsilon^2). \tag{5.2.3}$$

Μπορούμε λοιπόν να γράψουμε την συνάρτηση πιθανοφάνειας ως

$$p(\boldsymbol{y}|\boldsymbol{f},X) = \prod_{i=1}^{N} \mathcal{N}(y_i : f_i, \sigma_\varepsilon^2) = \mathcal{N}(\boldsymbol{y}; \boldsymbol{f}, \sigma_\varepsilon^2 I_N).$$

Συνδυάζοντας με την 5.2.1 παίρνουμε[1]

$$\begin{bmatrix} \boldsymbol{f} \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X) \\ m(X) \end{bmatrix}, \begin{bmatrix} K & K \\ K & K + \sigma_\epsilon^2 I_N \end{bmatrix} \right),$$

ενώ για την ύστερη σ.π.π της $\boldsymbol{f}$ θα ισχύει[2]

$$p(\boldsymbol{f} \mid \boldsymbol{y}, X) = \mathcal{N}\left( m(X) + K\left[K + \sigma_\epsilon^2 I_N\right]^{-1}(\boldsymbol{y} - m(X)), \quad K - K\left[K + \sigma_\epsilon^2 I_N\right]^{-1} K \right). \tag{5.2.4}$$

Λόγω του ορισμού της Γκαουσιανής Διαδικασίας, για ένα νέο σύνολο παρατηρούμενων τιμών $X^*$, η από κοινού σ.π.π των προβλέψεων $\boldsymbol{f}^*$ με τις $\boldsymbol{y}$ είναι επίσης κανονική:

$$\begin{bmatrix} \boldsymbol{f}^* \\ \boldsymbol{y} \end{bmatrix} \sim \mathcal{N}\left( \begin{bmatrix} m(X^*) \\ m(X) \end{bmatrix}, \begin{bmatrix} K(X^*, X^*) & K(X^*, X) \\ K(X, X^*) & K + \sigma_\epsilon^2 I_N \end{bmatrix} \right).$$

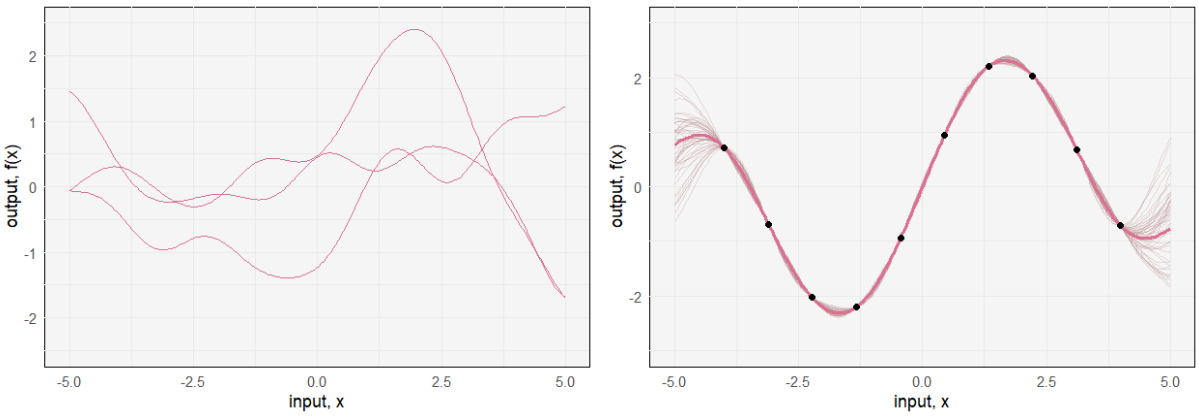
Εύκολα λοιπόν λαμβάνουμε την έκφραση για την προβλεπτική κατανομή της GP, κάνοντας χρήση της ίδιας ιδιότητας που χρησιμοποιήσαμε στην 5.2.4:

$$p(\boldsymbol{f}^* | X^*, X, \boldsymbol{y}) = \mathcal{N}(\boldsymbol{m}^*, \boldsymbol{s}^*), \tag{5.2.5}$$

όπου για τη μέση τιμή και τον πίνακα διασπορών-συνδιασπορών ισχύει

$$\boldsymbol{m}^* = m(X^*) + k(X^*, X)\left[K(X, X) + \sigma_\varepsilon^2 I_N\right](\boldsymbol{y} - m(X))$$
$$\boldsymbol{s}^* = k(X^*, X^*) - K(X^*, X)\left[K(X, X) + \sigma_\varepsilon^2 I_N\right]^{-1} K(X, X^*).$$

Επανερχόμαστε στο σύνολο δεδομένων του Διάγραμματος 5.2 και αυτή τη φορά τα μοντελοποιούμε με χρήση της GP. Αν θεωρήσουμε ότι οι παρατηρήσεις μας είναι ακριβείς, χωρίς θόρυβο, μπορούμε να μηδενίσουμε τον όρο $\sigma_\varepsilon^2$ στις παραπάνω εκφράσεις. Στο Διάγραμμα 5.3 παρουσιάζονται τα αποτελέσματα των υπολογισμών που περιγράψαμε, για αυτό το σύνολο δεδομένων. Παρατηρούμε ότι όλες οι τυχαίες συναρτήσεις-δείγματα από την ύστερη κατανομή διέρχονται από τα σημεία εκπαίδευσης, όπως φυσικά και η ύστερη μέση τιμή της.

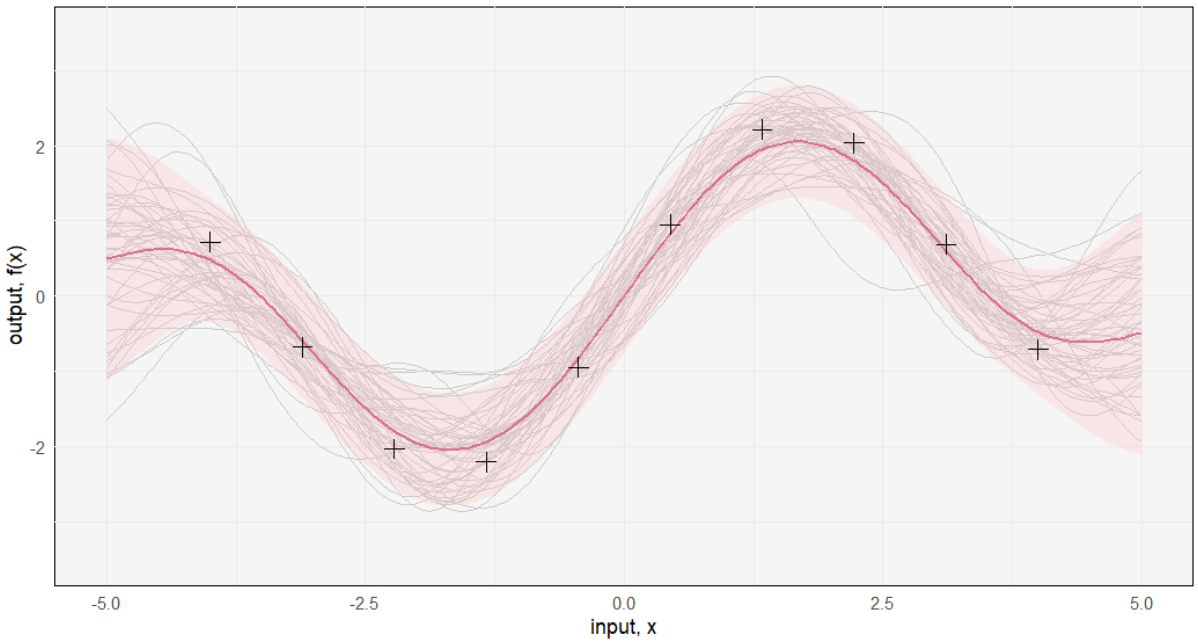



Διάγραμμα 5.3: Αριστερά: Δείγματα από την GP πρότερη κατανομή, με μέση τιμή τη μηδενική συνάρτηση και τετραγωνική εκθετική συνάρτηση διασποράς με υπερπαραμέτρους $l = 1, \sigma = 1$. Δεξιά: δείγματα από την εκ των υστέρων κατανομή.

Φυσικά, στην πλειονότητα των περιπτώσεων οι παρατηρήσεις συνοδεύονται από ένα σφάλμα (υποθέτουμε προσθετικό, κανονικά κατανεμημένο θόρυβο). Σε αυτήν την περίπτωση, η ύστερη μέση τιμή της $f$ δεν διέρχεται ακριβώς από τα σημεία, όπως στο Διάγραμμα 5.3. Η αντίστοιχη ύστερη παρουσιάζεται στο Διάγραμμα 5.4.

[1] Εδώ χρησιμοποιήσαμε την ιδιότητα σύμφωνα με την οποία η από κοινού σ.π.π των $X, Y$ είναι κανονική αν και μόνο αν η $\alpha X + \beta Y$ είναι κανονικά κατανεμημένη για κάθε $\alpha, \beta \in \mathbb{R}$. Η υπόθεση αυτή ισχύει εδώ λόγω της κανονικότητας του θορύβου.

[2] Εδώ χρησιμοποιούμε την ιδιότητα που αφορά στην δεσμευμένη κατανομή $\mathbf{x}_1 | \mathbf{x}_2 = \alpha$ όταν $\begin{bmatrix} \mathbf{x}_1 \\ \mathbf{x}_2 \end{bmatrix} \sim \mathcal{N}$, η οποία είναι επίσης κανονική με μέση τιμή $\boldsymbol{\mu} = \boldsymbol{\mu}_1 + \Sigma_{12}\Sigma_{22}^{-1}(\alpha - \boldsymbol{\mu}_2)$ και $\Sigma = \Sigma_{11} - \Sigma_{12}\Sigma_{22}^{-1}\Sigma_{21}$.




Διάγραμμα 5.4: Δείγματα από την εκ των υστέρων κατανομή, με γκρι χρώμα. Η μέση τιμή της ύστερης της $f$ σε κάθε σημείο εμφανίζεται με ρόζ χρώμα, ενώ το γραμμοσκιασμένο χωρίο αντικατοπτρίζει την συλλογή των κατά σημείο 95% διαστημάτων εμπιστοσύνης.

## 5.3 Ταξινόμηση με Γκαουσιανές διαδικασίες

Στη προηγούμενη ενότητα ασχοληθήκαμε με προβλήματα παλινδρόμησης, όπου οι στόχοι είναι πραγματικές τιμές. Μια άλλη σημαντική κατηγορία προβλημάτων είναι τα προβλήματα ταξινόμησης, όπου επιθυμούμε να αναθέσουμε ένα μοτίβο εισόδου $\boldsymbol{x}$ σε μία από τις $C$ κλάσεις, $C_1, \ldots, C_C$. Παραδείγματα πρακτικών προβλημάτων ταξινόμησης είναι η αναγνώριση χειρόγραφων ψηφίων και η ταξινόμηση αντικειμένων σε αστρονομικές έρευνες. Αυτά τα παραδείγματα δείχνουν ότι τα προβλήματα ταξινόμησης μπορούν να είναι είτε δυαδικά (με δύο κλάσεις) είτε πολυκλασικά (με περισσότερες από δύο κλάσεις).

Ας ανακαλέσουμε τη μέθοδο της λογιστικής παλινδρόμησης για το πρόβλημα της δυαδικής ταξινόμησης. Δεδομένου ενός συνόλου δεδομένων $\boldsymbol{x}$ και των ετικετών τους $y = +1, y = -1$ η συνάρτηση πιθανοφάνειας είναι

$$p(y = \pm 1 | \boldsymbol{x}, \boldsymbol{w}) = \sigma(\boldsymbol{x}^T \boldsymbol{w}),$$

όπου $\sigma$ μια σιγμοειδής συνάρτηση[3] και $\boldsymbol{w}$ το διάνυσμα των βαρών, το οποίο θέλουμε και να προσδιορίσουμε. Στη λογιστική παλινδρόμηση, η συνάρτηση σύνδεσης είναι η λογιστική συνάρτηση,

$$\sigma(z) = \frac{1}{1 + e^{-z}}.$$

Επειδή οι πιθανότητες των δύο κλάσεων πρέπει να αθροίζουν στη μονάδα, έχουμε

$$p(y = -1 | \boldsymbol{x}, \boldsymbol{w}) = 1 - p(y = +1 | \boldsymbol{x}, \boldsymbol{w}).$$

Άρα για το σημείο $(\boldsymbol{x}_i, y_i)$, η αντίστοιχη πιθανότητα δίνεται από την $\sigma(\boldsymbol{x}_i^T \boldsymbol{W})$, αν ανήκει στην κλάση $y_i = +1$ και $1 - \sigma(\boldsymbol{x}_i^T \boldsymbol{W})$, αν ανήκει στην κλάση $y_i = -1$. Η λογιστική συνάρτηση είναι συμμετρική, με αποτέλεσμα $\sigma(-z) = 1 - \sigma(z)$, επομένως μπορούμε να γράψουμε

$$p(y_i | \boldsymbol{x}_i, \boldsymbol{w}) = \sigma(y_i, f_i),$$

όπου $f_i = f(\boldsymbol{x}_i) = \boldsymbol{x}_i^T \boldsymbol{w}$. Για το διάνυσμα των βαρών, όπως στην περίπτωση της Μπεϋζιανής γραμμικής παλινδρόμησης, θεωρούμε μία κανονική πρότερη κατανομή $\boldsymbol{w} \sim \mathcal{N}(0, \Sigma_p)$.

Για την δυαδική ταξινόμηση, η βασική ιδέα πίσω από την πρόβλεψη με Γκαουσιανές Διαδικασίες (GPs) είναι

[3] Μια συνάρτηση $\sigma : \mathbb{R} \to [0, 1]$ καλείται σιγμοειδής αν είναι αύξουσα και προσομοιάζει το Διάγραμμα του γράμματος S.

απλή: τοποθετούμε μια GP πρότερη πάνω στη λανθάνουσα συνάρτηση $f(\boldsymbol{x})$ και στη συνέχεια τη "συμπιέζουμε" μέσω της λογιστικής συνάρτησης για να αποκτήσουμε μια πρότερη κατανομή στην πιθανότητα κλάσης $\pi(\boldsymbol{x}) = p(y = +1|\boldsymbol{x}) = \sigma(f(\boldsymbol{x}))$. Η $\pi$ είναι μια ντετερμινιστική συνάρτηση του $f$, και επειδή το $f$ είναι στοχαστικό, το ίδιο ισχύει και για την $\pi$.

Η κατασκευή αυτή γενικεύει το γραμμικό λογιστικό μοντέλο και προσομοιάζει στην μετάβαση από γραμμική παλινδρόμηση σε παλινδρόμηση με GP, όπως είδαμε στην προηγούμενη ενότητα. Συγκεκριμένα, αντικαθιστούμε τη γραμμική συνάρτηση $f(\boldsymbol{x})$ του γραμμικού λογιστικού μοντέλου με μια GP, και αντίστοιχα την κανονική πρότερη κατανομή πάνω στα βάρη με μια GP πρότερη κατανομή.

Δεν παρατηρούμε τις τιμές της $f$ απευθείας, αλλά ενδιαφερόμαστε μόνο για την $\pi$, ιδιαίτερα για τις τιμές $\pi(\boldsymbol{x}^*)$ για νέα σημεία $\boldsymbol{x}^*$.

Όσον αφορά την συμπερασματολογία, αρχικά υπολογίζουμε την κατανομή της $f_*$ για τα νέα σημεία

$$p(f^*|X, \boldsymbol{y}, \boldsymbol{x}^*) = \int p(f^*|X, \boldsymbol{x}^*, f)p(f|X, \boldsymbol{y})df,$$

όπου $p(f|X, \boldsymbol{y}) = \frac{p(\boldsymbol{y}|f)p(f|X)}{p(y|X)}$ είναι η ύστερη κατανομή των λανθανουσών μεταβλητών, και στη συνέχεια χρησιμοποιούμε αυτή την κατανομή για να παράγουμε την συνάρτηση πρόβλεψης

$$\pi^* = p(y^* = +1|X, y, x^*) = \int \sigma(f^*)p(f^*|x, y, x^*)df^*.$$

Στην περίπτωση της παλινδρόμησης (με κανονική συνάρτηση πιθανότητας) ο υπολογισμός των προβλέψεων ήταν απλός καθώς τα σχετικά ολοκληρώματα ήταν κανονικά και μπορούσαν να υπολογιστούν αναλυτικά. Στην ταξινόμηση η μη κανονική συνάρτηση πιθανότητας καθιστά το ολοκλήρωμα αναλυτικά δυσεπίλυτο. Έτσι, πρέπει να χρησιμοποιήσουμε είτε αναλυτικές προσεγγίσεις των ολοκληρώσεων είτε λύσεις βασισμένες σε δειγματοληψία Monte Carlo. Η συνηθέστερες αναλυτικές προσεγγίσεις είναι η Laplace (Williams and Barber 1998) και η μέθοδος Expectation Propagation (Minka 2001). Εντούτοις η αναλυτική παρουσίαση των δύο αυτών μεθόδων ξεφεύγει από τους σκοπούς της παρούσας εργασίας.

## 5.4 Επιλογή μοντέλου και ρύθμιση υπερπαραμέτρων

Μία Γκαουσιανή Διαδικασία προσδιορίζεται πλήρως από την συνάρτηση συνδιασποράς $k(x_i, x_j)$ και την μέση τιμή της $m$, με την τελευταία να επιλέγεται συνήθως ως η μηδενική. Όπως αναφέρθηκε στην εισαγωγική ενότητα, μια συνήθης επιλογή για την πρώτη είναι η τετραγωνική εκθετική συνάρτηση,

$$k(x_i, x_j) = \sigma^2 exp\left(-\frac{|x_i - x_j|^2}{l^2}\right),$$

με υπερπαραμέτρους $\sigma, l$. Ωστόσο, υπάρχει πλήθος επιλογών για την συνάρτηση συνδιασποράς. Όσον αφορά την επιλογή της κατάλληλης, αυτή θα πρέπει να συμβαδίζει με την πρότερη γνώση σχετικά με την συνάρτηση που θέλουμε να εκτιμήσουμε. Για παράδειγμα, αν τα δεδομένα μας εμφανίζουν περιοδική συμπεριφορά, είναι λογικό η συνάρτηση συνδιασποράς να είναι επίσης περιοδική. Συνηθίζεται η δοκιμή διάφορων συναρτήσεων πυρήνα και ο προσδιορισμός της βέλτιστης για το πρόβλημα, με τη μέθοδο Cross Validation. Ορισμένες συνήθεις επιλογές για την συνάρτηση συνδιασποράς περιλαμβάνουν:

1. Τετραγωνική εκθετική συνάρτηση (RBF):

$$k(x_i, x_j) = \sigma^2 \exp\left(-\frac{(x_i - x_j)^2}{l^2}\right)$$

2. Συνάρτηση Matérn :

$$k(x_i, x_j) = \sigma^2 \frac{2^{1-\nu}}{\Gamma(\nu)} \left(\frac{\sqrt{2\nu}|x_i - x_j|}{l}\right)^{\nu} K_\nu \left(\frac{\sqrt{2\nu}|x_i - x_j|}{l}\right)$$

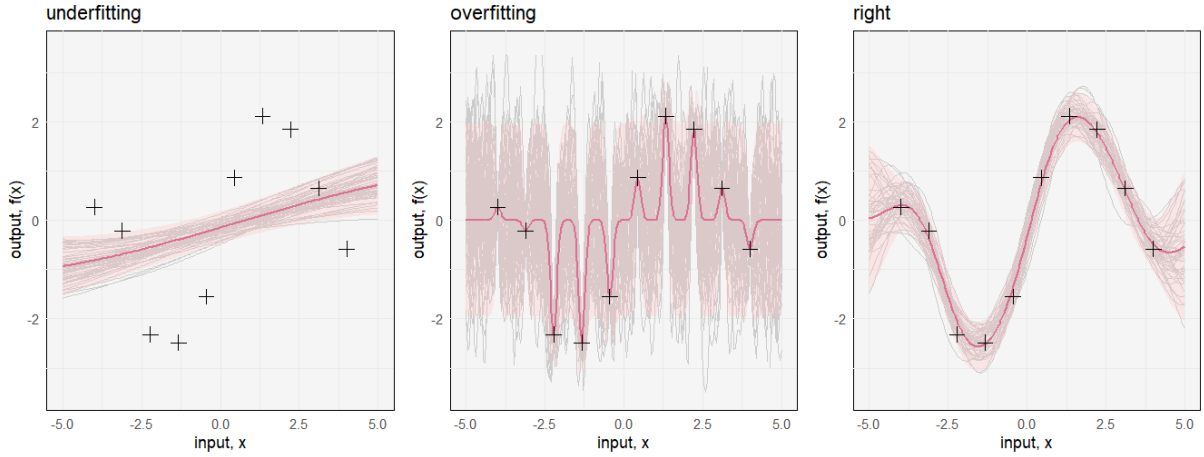
3. Περιοδικός Πυρήνας:

$$k(x_i, x_j) = \sigma^2 \exp\left(-\frac{2\sin^2\left(\frac{\pi|x_i - x_j|}{p}\right)}{l^2}\right)$$

4. Εκθετικός Πυρήνας:
$$k(x_i, x_j) = \sigma^2 \exp\left(-\frac{|x_i - x_j|}{l^2}\right)$$

5. Ρητά Τετραγωνικός Πυρήνας (RQ):
$$k(x_i, x_j) = \sigma^2 \left(1 + \frac{\|x - x'\|^2}{2\alpha\ell^2}\right)^{-\alpha}$$

Η διαδικασία επιλογής της καθώς και ο προσδιορισμός των βέλτιστων υπερπαραμέτρων καλείται επιλογή μοντέλου (model selection) και είναι καθοριστική για την αποφυγή υπερπροσαρμογής ή υποπροσαρμογής του τελικού μοντέλου στα δεδομένα. Αυτό καθίσταται φανερό στο Διάγραμμα 5.5. Στην πρώτη περίπτωση ($l = 10, \sigma^2 = 0.5$), το μοντέλο είναι πολύ απλό για να περιγράψει τη μεταβλητότητα των δεδομένων. Ως αποτέλεσμα έχουμε το φαινόμενο της υποπροσαρμογής (underfitting). Αντίθετα, στην δεύτερη περίπτωση ($l = 0.1, \sigma^2 = 0.01$), συναντάμε το φαινόμενο της υπερπροσαρμογής. Το μοντέλο ουσιαστικά "απομνημονεύει" τα δεδομένα, με αποτέλεσμα να έχει χαμηλή προβλεπτική ικανότητα στην περίπτωση νέων δεδομένων. Μια βελτιωμένη κατάσταση παρουσιάζεται στο τρίτο σχήμα ($l = 1, \sigma^2 = 0.3$) όπου επικρατεί ισορροπία μεταξύ της προσαρμογής στα υπάρχοντα δεδομένα και της γενίκευσης σε νέα.




Διάγραμμα 5.5: Σύγκριση διαφορετικών μοντέλων παλινδρόμησης με Γκαουσιανές Διαδικασίες, με διαφορετικές επιλογές υπερπαραμέτρων.

Αφότου επιλεγεί η κατάλληλη συνάρτηση πυρήνα, σειρά έχει η ρύθμιση των παραμέτρων της. Θα επικεντρωθούμε στην επιλογή μοντέλου μέσω της μεγιστοποίησης της περιθώριας πιθανοφάνειας. Αν συμβολίσουμε με $\boldsymbol{\theta}$ το διάνυσμα των υπερπαραμέτρων της συνάρτησης συνδιασποράς της Γκαουσιανής Διαδικασίας, έχουμε από το Θεώρημα Ολικής Πιθανότητας,

$$p(\boldsymbol{y}|\boldsymbol{\theta}, X) = \int p(\boldsymbol{y}|\boldsymbol{f}, X)p(\boldsymbol{f}|\boldsymbol{\theta}, X)d\boldsymbol{f}.$$

Όμως από την 5.2.2 είναι φανερό ότι

$$\boldsymbol{y}|\boldsymbol{\theta}, X \sim \mathcal{N}(\mathbb{0}, K_{\boldsymbol{\theta}} + \sigma_\varepsilon^2 I),$$

επομένως, κατά τα γνωστά,

$$\log p(\boldsymbol{y}|\boldsymbol{\theta}, X) = -\frac{1}{2}\boldsymbol{y}^T(K_{\boldsymbol{\theta}} + \sigma_\varepsilon^2 I)^{-1}\boldsymbol{y} - \frac{1}{2}\log|K_{\boldsymbol{\theta}} + \sigma_\varepsilon^2 I| - \frac{N}{2}\log(2\pi). \tag{5.4.1}$$

Ο πρώτος όρος της 5.4.1 μπορεί να ερμηνευτεί ως όρος προσαρμογής στα δεδομένα, ενώ ο δεύτερος όρος ερμηνεύεται ως η "ποινή" που αντιστοιχεί στην αύξηση της πολυπλοκότητας του μοντέλου (Bias-Variance Tradeoff). Στην περίπτωση της τετραγωνικής εκθετικής συνάρτησης συνδιασποράς, λόγου χάρη, ο όρος προσαρμογής στα δεδομένα μειώνεται με την αύξηση της υπερπαραμέτρου $l$, ενώ ο όρος πολυπλοκότητας αυξάνεται (Rasmussen and Williams 2006). Για τη μεγιστοποίηση της παραπάνω ποσότητας, είναι

απαραίτητος ο υπολογισμός των μερικών παραγώγων της ως προς τις υπερπαραμέτρους[4]:

$$\begin{aligned}\frac{\partial}{\partial\theta_j}\log p(\boldsymbol{y}|X,\boldsymbol{\theta}) &= \frac{1}{2}\boldsymbol{y}^T K_{\boldsymbol{\theta}}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_j}K^{-1}\boldsymbol{y} - \frac{1}{2}tr\left(K^{-1}\frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_j}\right)\\ &= \frac{1}{2}tr\left((\boldsymbol{a}\boldsymbol{a}^T - K^{-1})\frac{\partial K_{\boldsymbol{\theta}}}{\partial\theta_j}\right),\end{aligned}$$

όπου $\boldsymbol{a} = K^{-1}\boldsymbol{y}$. Η μέθοδος αυτή οδηγεί σε μία σημειακή εκτίμηση των βέλτιστων υπερπαραμέτρων ενώ η διαδικασία αυτή καλείται και εκπαίδευση της Γκαουσιανής Διαδικασίας.

[4]Χρησιμοποιούμε τις εξής ιδιότητες:

1. $\frac{\partial}{\partial\theta}K^{-1} = -K^{-1}\frac{\partial K}{\partial\theta}K^{-1}$
2. $\frac{\partial}{\partial\theta}\log|K| = tr\left(K^{-1}\frac{\partial K}{\partial\theta}\right)$

# Κεφάλαιο 6

# Εφαρμογή: Σύνολο Δεδομένων Old Faithful

Καθ' όλη την έκταση της παρούσας εργασίας, εφαρμόσαμε τις μη-παραμετρικές μεθόδους που παρουσιάστηκαν κυρίως σε συνθετικά-προσομοιωμένα δεδομένα. Παρόλο που οι τετριμμένες αυτές εφαρμογές αρκούν για την απλή κατανόηση των μεθόδων, για να αναδείξουμε πλήρως την αξία τους θα προβούμε στην υλοποίηση των μεθόδων πάνω σε πραγματικά δεδομένα. Συγκεκριμένα, θα εργαστούμε αρχικά με το σύνολο δεδομένων *Old Faithful* (National Park Service 1935), το οποίο περιέχει πληροφορίες σχετικά με τους χρόνους αναμονής και τη διάρκεια των εκρήξεων του διάσημου γκέιζερ Old Faithful στο Εθνικό Πάρκο Yellowstone. Το σύνολο δεδομένων αποτελείται από 272 παρατηρήσεις και περιλαμβάνει δύο μεταβλητές:

1. $waiting$: Χρόνος αναμονής (σε λεπτά) μέχρι την επόμενη έκρηξη του γκέιζερ.
2. $eruptions$: Διάρκεια της έκρηξης (σε λεπτά).

Αυτό το σύνολο δεδομένων χρησιμοποιείται ευρέως στην ανάλυση δεδομένων και αποτελεί ένα ιδανικό παράδειγμα για να καταδείξουμε πώς οι μη-παραμετρικές μέθοδοι μπορούν να εφαρμοστούν σε πραγματικά δεδομένα για την κατανόηση της κατανομής και της δομής των παρατηρούμενων φαινομένων. Θα εστιάσουμε κυρίως στους χρόνους αναμονής, με στόχο να εκτιμήσουμε την κατανομή τους και να εξετάσουμε αν υπάρχουν διαφορετικές συστάδες που να περιγράφουν τη συμπεριφορά του γκέιζερ.

## 6.1 Ανάλυση δεδομένων

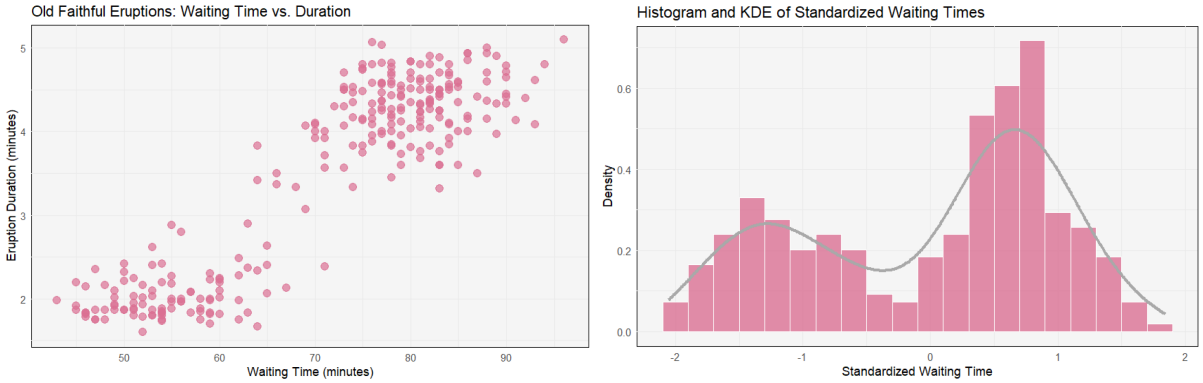
Προτού προβούμε στη μοντελοποίηση των εν λόγω δεδομένων με χρήση της Διαδικασίας Dirichlet, θα κάνουμε μια αρχική περιγραφική ανάλυση για να ανακαλύψουμε την δομή τους. Τα περιγραφικά στοιχεία παρατίθενται στον Πίνακα 6.1.

| **Μέτρο** | **eruptions (λεπτά)** | **waiting (λεπτά)** |
|---|---|---|
| Min | 1.600 | 43.0 |
| 1ο τετ/ριο | 2.163 | 58.0 |
| Διάμεσος | 4.000 | 76.0 |
| Μέσος | 3.488 | 70.9 |
| 3ο τετ/ριο | 4.454 | 82.0 |
| Max | 5.100 | 96.0 |

Πίνακας 6.1: Περιγραφικά στατιστικά του συνόλου δεδομένων Old Faithful

Η διάρκεια των εκρήξεων κυμαίνεται μεταξύ 1.6 και 5.1 λεπτών, με μεση τιμή τα 3.49 λεπτά ενώ ο χρόνος αναμονής μεταξύ δύο διαδοχικών εκρήξεων κυμαίνεται μεταξύ 43 και 96 λεπτών με τη μέση τιμή να είναι 70.9 λεπτά.

Στο Διάγραμμα 6.1 απεικονίζονται τα ζεύγη σημείων $(waiting, eruption)$ σε ένα καρτεσιανό σύστημα συντεταγμένων, και καθίσταται ήδη φανερό ότι υπάρχουν δύο ομάδες στα δεδομένα: εκρήξεις με διάρκεια




Διάγραμμα 6.1: Αριστερά: Διάγραμμα διασποράς των δεδομένων Old Faithful. Δεξιά: Ιστόγραμμα των τυποποιημένων χρόνων αναμονής μεταξύ διαδοχικών εκρήξεων. Η γκρι καμπύλη αντιστοιχεί στην KDE (Kernel Density Estimator) εκτίμηση της πυκνότητας.

μικρότερη από 3 λεπτά και εκρήξεις με διάρκεια μεγαλύτερη από 3 λεπτά. Μάλιστα, παρατηρώντας το ιστόγραμμα των τυποποιημένων χρόνων αναμονής στο ίδιο Διάγραμμα, φαίνεται λογική η υπόθεση ότι αυτοί προέρχονται από μία μίξη δύο κανονικών κατανομών.

## 6.2 Μοντελοποίηση του χρόνου αναμονής

Στην συνέχεια του παρόντος κεφαλαίου θα χρησιμοποιήσουμε το πακέτο *dirichletprocess* (Merritt et al. 2022) της R για την προσαρμογή ενός Μοντέλου Μίξης Διαδικασίας Dirichlet στα δεδομένα Old Faithful.

Έστω $\{y_i\}_{i=1,\dots,272}$ το σύνολο των παρατηρούμενων χρόνων αναμονής. Θα μοντελοποιήσουμε τις παρατηρήσεις αυτές με το εξής DPMM:

$$y_i \sim F \tag{6.2.1}$$

$$F = \sum_{i=1}^{n} \pi_i \mathcal{N}(y_i|\boldsymbol{\theta}_i), \boldsymbol{\theta}_i = (\mu_i, \sigma_i^2) \tag{6.2.2}$$

$$\boldsymbol{\theta}_i \sim G \tag{6.2.3}$$

$$G \sim DP(\alpha, G_0), \tag{6.2.4}$$

όπου ως μέτρο βάσης $G_0$ θα χρησιμοποιηθεί η συζυγής κατανομή για την Κανονική, γνωστή ως Κανονική-Γάμμα:

$$G_0(\theta|\boldsymbol{\gamma}) = \mathcal{N}\left(\mu|\mu_0, \frac{\sigma^2}{k_0}\right) \text{Inv-Gamma}(\sigma^2|\alpha_0, \beta_0),$$

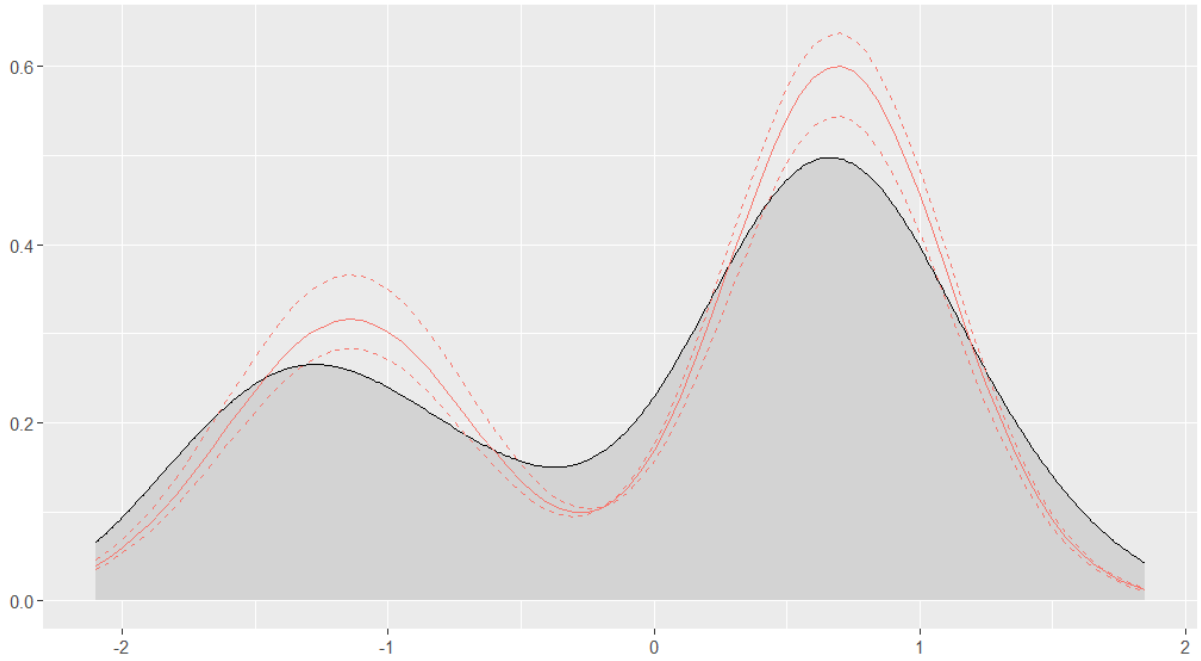
με τις προκαθορισμένες υπερπαραμέτρους του πακέτου, $\boldsymbol{\gamma} = (\mu_0, k_0, \alpha_0, \beta_0) = (0, 1, 1, 1)$, οι οποίες έχουν ως αποτέλεσμα μια μή-πληροφοριακή πρότερη, αφού πρώτα γίνει τυποποίηση των δεδομένων.

Η τελική προσαρμογή του μοντέλου παρουσιάζεται στο Διάγραμμα 6.2, με τον ύστερο μέσο (δηλαδή, το μοντέλο μίξης της 6.2.2 όπου οι παράμετροι $\mu_i, \sigma_i^2$ έχουν αντικατασταθεί με την μέση τιμή της αντίστοιχης ύστερης κατανομής) και το αντίστοιχο 95% διάστημα πιθανότητας. Εξερευνώντας το αντικείμενο $dp$ που δημιουργήθηκε, μπορούμε να ελέγξουμε την ύστερη μέση τιμή των παραμέτρων των συστάδων που ανακαλύφθηκαν από το μοντέλο, καθώς και τις αντίστοιχες παραμέτρους μίξης. Οι τιμές αυτές παρουσιάζονται στον Πίνακα 6.2.

| $\pi_i$ | **Μέσος ($\mu$)** | **Τυπική Απόκλιση ($\sigma$)** |
|---|---|---|
| 0.63 | 0.78 | 0.46 |
| 0.37 | -1.14 | 0.45 |

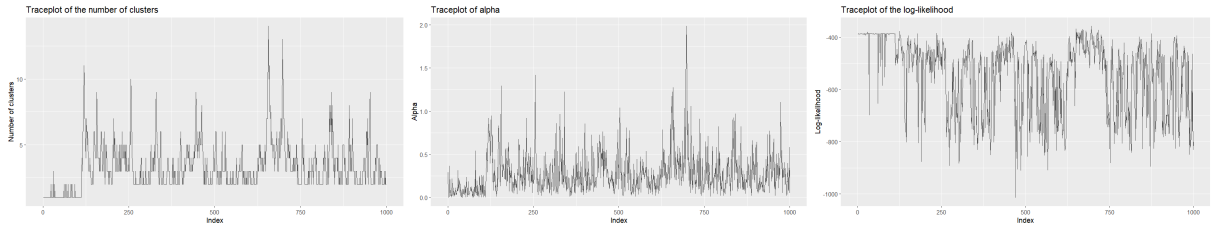
Πίνακας 6.2: Εκ των υστέρων μέσοι των παραμέτρων των δύο συστάδων.

Όσον αφορά την υπερπαράμετρο συγκέντρωσης $\alpha$ της Διαδικασίας Dirichlet, υπενθυμίζουμε ότι αυτή ελέγχει το πόσο πιθανό είναι νέα σημεία να σχηματίσουν νέες συστάδες αντί να τοποθετηθούν στις ήδη υπάρχουσες.




Διάγραμμα 6.2: Εκτίμηση πυκνότητας πιθανότητας του συνόλου δεδομένων. Με κόκκινο χρώμα σχεδιάζεται ο ύστερος μέσος του DPMM καθώς και το 95% διάστημα πιθανότητας.

Το πακέτο *dirichletprocess* αναθέτει στην $\alpha$ πρότερη Γάμμα κατανομή ενώ δείγματα από την ύστερη της κατανομή λαμβάνονται σε κάθε επανάληψη με τρόπο που σκιαγραφείται εκτενέστερα στις οδηγίες του πακέτου (West 1992). Εδώ χρησιμοποιήσαμε για την παράμετρο $\alpha$ πρότερη $Gamma(1, 6)$. Ακόμα, δίνεται η δυνατότητα για τύπωση διαγνωστικών διαγραμμάτων, όπως τα διαγράμματα προσομοιωμένων τιμών των παραμέτρων (traceplots), αλλά και για την εξέλιξη της τιμής της λογαριθμοποιημένης πιθανοφάνεια, τα οποία παρέχουν σημαντική πληροφορία σχετικά με την μίξη της αλυσίδας MCMC κάθε παραμέτρου. Αυτά παρατίθενται στο Διάγραμμα 6.3. Παρατηρούμε ότι η μίξη των αλυσίδων είναι σχετικά ικανοποιητική, ιδιαίτερα για την παράμετρο $\alpha$ αν και το πλήθος των συστάδων φαίνεται να ταλαντώνεται αρκετά με την πάροδο των επαναλήψεων, ενδεικτικό λιγότερο ικανοποιητικής μίξης της αλυσίδας. Για να επιλύσουμε το πρόβλημα αυτό μπορούμε να αυξήσουμε το πλήθος των επαναλήψεων (εδώ χρησιμοποιήσαμε 1000). Ωστόσο σημειώνεται ότι παρόλο που σε ορισμένες επαναλήψεις "ανοίγουν" νέες συστάδες, αυτές έχουν τετριμμένο πλήθος σημείων και ως εκ τούτου είναι στην ευχέρεια μας να τις αγνοήσουμε.

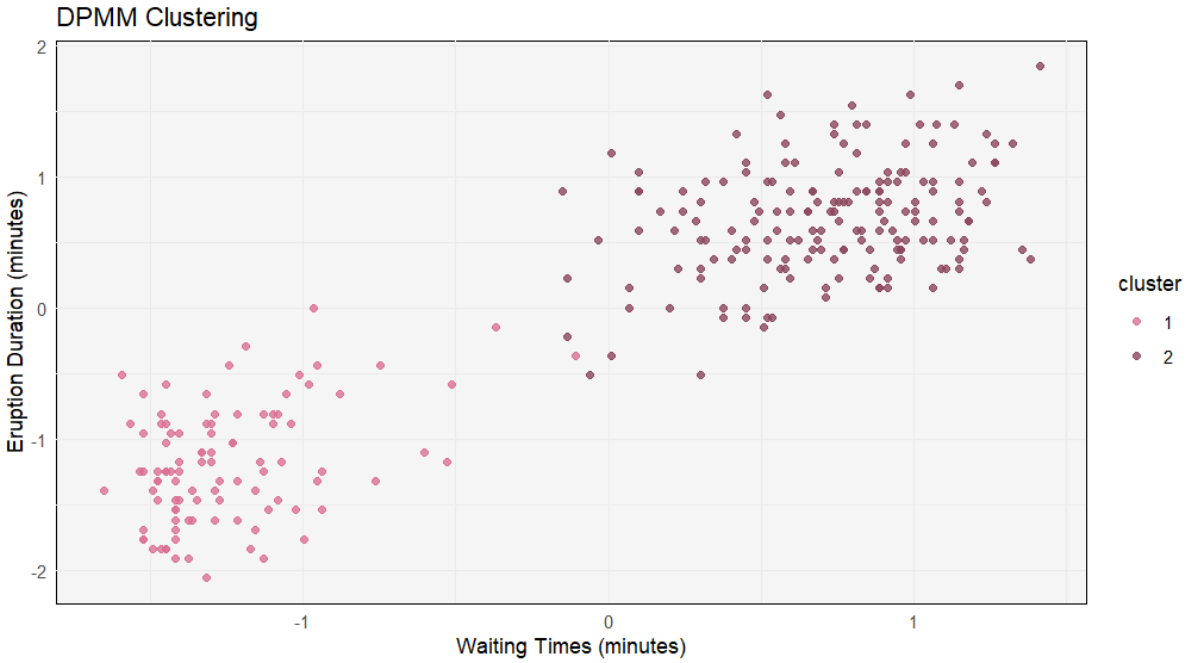



Διάγραμμα 6.3: Διαγνωστικά διαγράμματα από την προσαρμογή του μοντέλου DPMM.

## 6.3 Συσταδοποίηση των εκρήξεων με DPMM

Εκτός από τη μοντελοποίηση του χρόνου αναμονής ως μία μίξη κανονικών κατανομών, μπορούμε να μοντελοποιήσουμε τα ζεύγη των δεδομένων ως μίξη δισδιάστατων κανονικών κατανομών. Αν θεωρήσουμε $\boldsymbol{y}_i = (waiting, eruption)$, μπορούμε να μοντελοποιήσουμε τα δεδομένα ως

$$\boldsymbol{y}_i \sim \mathcal{N}(\boldsymbol{y}|\boldsymbol{\theta}_i)$$
$$\boldsymbol{\theta}_i = (\boldsymbol{\mu}_i, \Sigma_i)$$



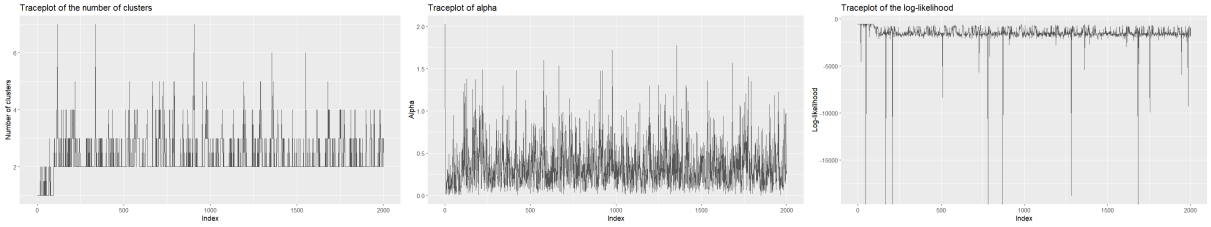

Διάγραμμα 6.4: Με διαφορετικά χρώματα συμβολίζονται οι δύο συστάδες που ανακαλύφθηκαν μέσω της προσαρμογής του μοντέλου DPMM.

$$\boldsymbol{\theta}_i \sim G$$
$$G \sim DP(\alpha, G_0),$$

όπου για το μέτρο βάσης $G_0$ ισχύουν τα ίδια με την προηγούμενη ενότητα και για την υπερπαράμετρο $\alpha$ χρησιμοποιείται πρότερη Γάμμα κατανομή.

Χρησιμοποιώντας και πάλι το πακέτο *dirichletprocess*, θα προσαρμόσουμε το παραπάνω μοντέλο μίξης. Στόχος είναι η ανάθεση μιας ετικέτας σε κάθε σημείο του συνόλου δεδομένων, που θα συμβολίζει σε ποια συστάδα ανήκει. Η τελική ανάθεση σε συστάδες φαίνεται στο Διάγραμμα 6.4, από όπου είναι φανερό ότι η μέθοδος εντόπισε τις δύο συστάδες διάρκειας των εκρήξεων, ανάλογα με τον χρόνο αναμονής από την προηγούμενη.

Έγιναν 2000 επαναλήψεις του αλγορίθμου, ενώ για την παράμετρο $\alpha$ χρησιμοποιήθηκε πρότερη $Gamma(1,1)$. Στο Διάγραμμα 6.5 παρουσιάζονται τα διαγράμματα των προσομοιωμένων τιμών της καθώς και η πορεία της λογαριθμοποιημένης πιθανοφάνειας. Από το πρώτο, φαίνεται η μίξη της αλυσίδας να μην είναι αρκετά ικανοποιητική.




Διάγραμμα 6.5: Διαγνωστικά διαγράμματα από την προσαρμογή του μοντέλου DPMM.

Στην περίπτωση αυτή φαίνεται πως το μοντέλο τείνει να υπερεκτιμήσει το πλήθος των συστάδων, ανοίγοντας νέες συστάδες με λίγα σημεία, παρόλο που για την παράμετρο συγκέντρωσης $\alpha$ χρησιμοποιήθηκε πρότερη $Gamma(1,1)$ (με μέσο και διασπορά 1). Το πρόβλημα αυτό έχει διερευνηθεί από τους Miller και Harrison, οι οποίοι απέδειξαν την ασυνέπεια της μεθόδου όσον αφορά τον προσδιορισμό του πλήθους των συστάδων, αν υπάρχουν ακόμα και μικρές αποκλίσεις από την προτεινόμενη οικογένεια κατανομών (εδώ, την Κανονική) (Miller and Harrison 2014). Για αυτό, συνιστούν τα DPMM περισσότερο ως ένα εργαλείο για εκτίμηση πυκνότητας, παρά για συσταδοποίηση.

# Κεφάλαιο 7

# Εφαρμογή: Σύνολο Δεδομένων Air Quality

Το κεφάλαιο αυτό αφορά στην εφαρμογή των Γκαουσιανών Διαδικασιών σε πραγματικά δεδομένα, και πιο συγκεκριμένα στο σύνολο δεδομένων *airquality* που περιλαμβάνει δεδομένα σχετικά με την ποιότητα του αέρα στη Νέα Υόρκη, από το Μάϊο έως τον Σεπτέμβριο του 1973. Το σύνολο δεδομένων αποτελείται από 153 παρατηρήσεις και περιλαμβάνει τρεις επεξηγηματικές μεταβλητές μεταβλητές:

1. $ozone$: Μέσος όρος του όζοντος σε μέρη ανά δισεκατομμύριο (ppb) από τη 1 έως τις 3 το μεσημέρι, στο Roosevelt Island.
2. $solar.R$: Ηλιακή ακτινοβολία σε Langleys (lang) στο φάσμα συχνοτήτων 4000–7700 Ångström από τις 8 το πρωί έως τις 12 το μεσημέρι, στο Central Park.
3. $wind$: Μέση ταχύτητα ανέμου σε μίλια ανά ώρα (mph) στις 7 και στις 10 το πρωί, στο αεροδρόμιο LaGuardia.
4. $temp$: Μέγιστη ημερήσια θερμοκρασία σε βαθμούς Φαρενάιτ (F), στο αεροδρόμιο La Guardia.

Σκοπός είναι η κατασκευή ενός μοντέλου GP για την πρόβλεψη της παρουσίας όζοντος στον αέρα, σε ppb, με βάση τις μετρήσεις των τριών επεξηγηματικών μεταβλητών $solar.R$,$wind$ και $temp$.

## 7.1 Ανάλυση δεδομένων

Αρχικά, θα κάνουμε μια περιγραφική ανάλυση ώστε να ανακαλύψουμε την δομή του συνόλου δεδομένων Τα περιγραφικά στοιχεία παρατίθενται στον Πίνακα 7.1.

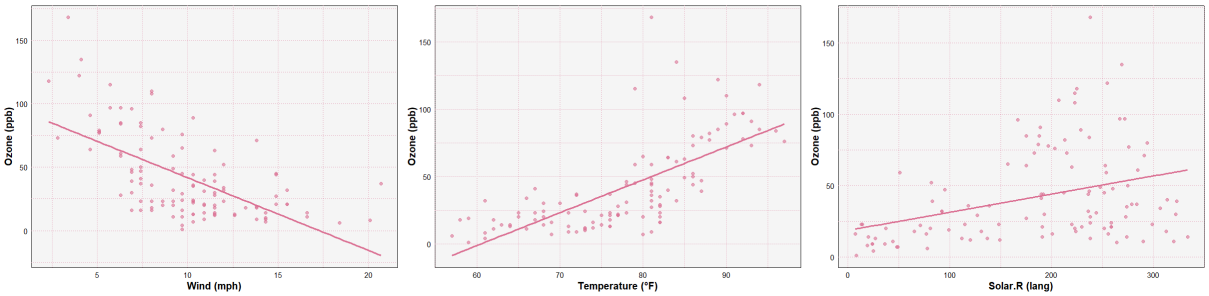
| **Μέτρο** | **Ozone (ppb)** | **Solar.R (lang)** | **Wind (mph)** | **Temp (°F)** |
|---|---|---|---|---|
| Min | 1.00 | 7.0 | 1.70 | 56.00 |
| 1ο τετ/ριο | 18.00 | 115.8 | 7.40 | 72.00 |
| Διάμεσος | 31.50 | 205.0 | 9.70 | 79.00 |
| Μέσος | 42.13 | 185.9 | 9.96 | 77.88 |
| 3ο τετ/ριο | 63.25 | 258.8 | 11.50 | 85.00 |
| Max | 168.00 | 334.0 | 20.70 | 97.00 |

Πίνακας 7.1: Περιγραφικά στατιστικά του συνόλου δεδομένων airquality

Για να γίνει πιο εύκολα αντιληπτή η σχέση μεταξύ του όζοντος σε ppb και των τριών επεξηγηματικών μεταβλητών, παρατίθενται στο Διάγραμμα 7.1 τα τρία διαγράμματα διασποράς. Είναι φανερό ότι η συγκέντρωση του όζοντος σε ppb είναι αρνητικά συσχετισμένη με την ταχύτητα του αέρα, ενώ αυξάνεται με την αύξηση της θερμοκρασίας και της ακτινοβολίας του ήλιου.

## 7.2 Μη-γραμμική παλινδρόμηση με GP

Στην συνέχεια του παρόντος κεφαλαίου θα χρησιμοποιήσουμε το πακέτο *GauPro* (Zhao and Wood 2022) της R για την προσαρμογή ενός μοντέλου Μη-Γραμμικής Παλινδρόμησης με Γκαουσιανές Διαδικασίες στα




Διάγραμμα 7.1: Ανά δύο διαγράμματα διασποράς των ποσοτήτων $wind$-$ozone$, $temperature$-$ozone$,$solar.R$-$ozone$.

δεδομένα Air Quality. Κατά την προεπεξεργασία του συνόλου δεδομένων, αφαιρέθηκαν οι γραμμές με ελλιπή δεδομένα (NA) με αποτέλεσμα να έχουμε τελικά 111 παρατηρήσεις. Στη συνέχεια ακολούθησε η τυποποίηση του συνόλου δεδομένων και τέλος αυτό χωρίστηκε σε train set (70%) και test set (30%).

Όσον αφορά την διαδικασία επιλογής μοντέλου, αυτό προσαρμόστηκε με χρήση τριών διαφορετικών πυρήνων: του τετραγωνικού εκθετικού, του πυρήνα Matern, και του ρητά τετραγωνικού πυρήνα. Για την βελτιστοποίηση των υπερπαραμέτρων χρησιμοποιήθηκε πλέγμα τιμών, με βάση το οποίο προσαρμόστηκε το μοντέλο με όλους τους συνδυασμούς και ως τελικό μοντέλο επιλέχθηκε αυτό με την μέγιστη τιμή της περιθώριας πιθανοφάνειας. Ο σχετικός κώδικας σε R παρατίθεται στο Παράρτημα 9.

Για την αξιολόγηση των μοντέλων, κάνουμε προβλέψεις με βάση αυτά χρησιμοποιώντας ως νέα δεδομένα τα δεδομένα του test set, και υπολογίζουμε τη μετρική RMSE,

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \hat{y}_i)^2}.$$

Τα τελικά αποτελέσματα και η σχετική σύγκριση παρουσιάζονται στον Πίνακα 7.2.

| **Πυρήνας** | $l$ | $\sigma$ | **RMSE** |
|---|---|---|---|
| RBF | 1.2 | 2.0 | 16.69370 |
| Matern | 0.2 | 1.7 | 16.72632 |
| RQ | 1.7 | 0.5 | 16.84276 |

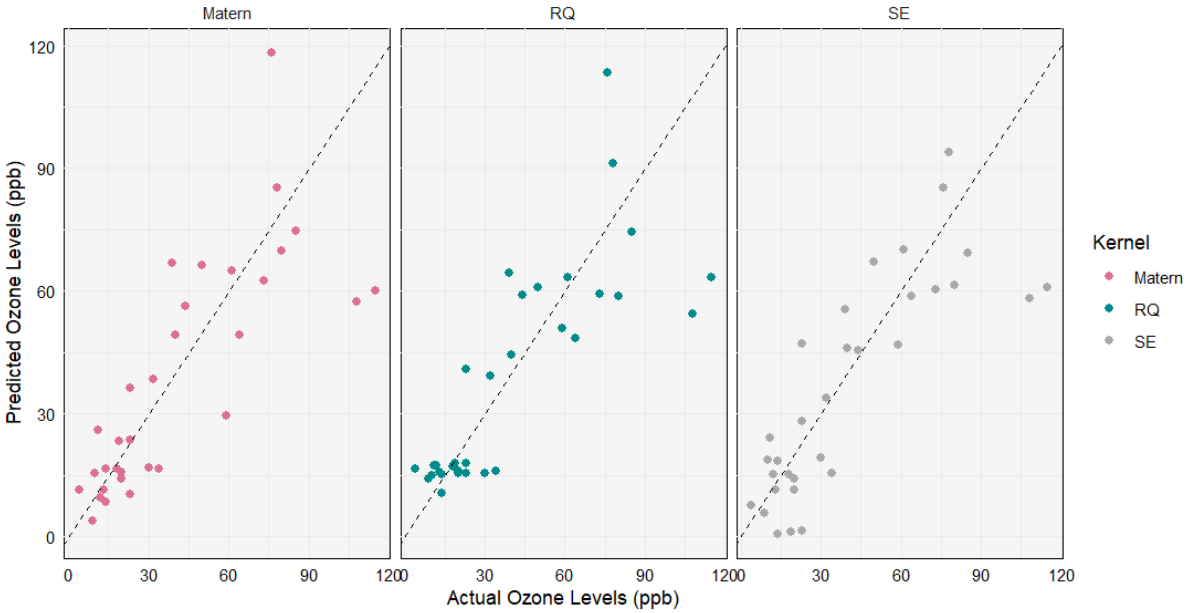
Πίνακας 7.2: Σύγκριση της προβλεπτικής ικανότητας στο test set μοντέλων GP με διαφορετικούς πυρήνες και βέλτιστες υπερπαραμέτρους.

Η διαφορά των RMSE μεταξύ των διαφορετικών μοντέλων φαίνεται να είναι τετριμμένη. Για λόγους πληρότητας, προσαρμόζουμε και ένα κλασικό μοντέλο πολλαπλής γραμμικής παλινδρόμησης και αξιολογούμε την επίδοση του στο test set. Το RMSE που προκύπτει είναι 30.47, σαφώς μεγαλύτερο από αυτό του μοντέλου GP (με οποιονδήποτε πυρήνα). Στο Διάγραμμα 7.2 αντιπαρατίθενται οι πραγματικές τιμές έναντι των προβλέψεων του μοντέλου, για τα δεδομένα του test set και για τις βέλτιστες ρυθμίσεις των τριών πυρήνων.

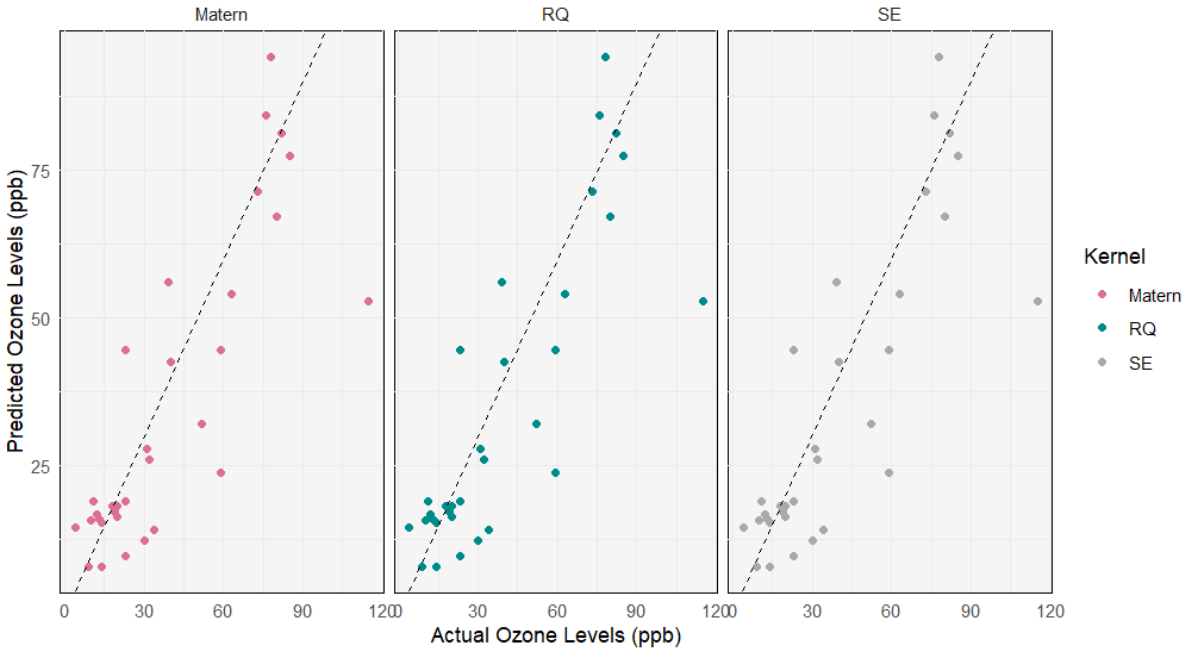
Η επίδοση του μοντέλου βελτιώνεται ελαφρώς αν αφαιρέσουμε από το σύνολο δεδομένων ακραίες παρατηρήσεις οι οποίες θεωρούμε πως είναι αποτέλεσμα εξωτερικών παραγόντων. Επαναλαμβάνουμε την ανάλυση χωρίς τις τιμές αυτές και λαμβάνουμε τα αποτελέσματα του Πίνακα 7.3. Η αντίστοιχη σύγκριση πραγματικών και προβλεπόμενων τιμών παρατίθεται στο Διάγραμμα 7.3.

| **Πυρήνας** | $l$ | $\sigma$ | **RMSE** |
|---|---|---|---|
| RBF | 0.7 | 0.9 | 16.15591 |
| Matern | 1.7 | 0.5 | 16.14543 |
| RQ | 0.7 | 1.3 | 16.16893 |

Πίνακας 7.3: Σύγκριση της προβλεπτικής ικανότητας στο test set μοντέλων GP με διαφορετικούς πυρήνες και βέλτιστες υπερπαραμέτρους, μετά την αφαίρεση των ακραίων τιμών.




Διάγραμμα 7.2: Διαγράμματα των πραγματικών και προβλεπόμενων τιμών του μοντέλου καθώς και η ευθεία $y = x$, για τις βέλτιστες ρυθμίσεις των τριών πυρήνων.




Διάγραμμα 7.3: Διαγράμματα των πραγματικών και προβλεπόμενων τιμών του μοντέλου καθώς και η ευθεία $y = x$, για τις βέλτιστες ρυθμίσεις των τριών πυρήνων μετά την αφαίρεση των ακραίων τιμών.

# Κεφάλαιο 8

# Συμπεράσματα και Μελλοντικές Επεκτάσεις

## 8.1 Συμπεράσματα

Στην παρούσα διπλωματική εργασία, διερευνήσαμε τη χρήση μη-παραμετρικών Μπεϋζιανών μεθόδων, με έμφαση στις Διαδικασίες Dirichlet (DP) και τις επεκτάσεις τους, όπως οι Ιεραρχικές Διαδικασίες Dirichlet (HDP) και οι Εξαρτημένες Διαδικασίες Dirichlet (DDP). Οι Διαδικασίες Dirichlet αναδείχθηκαν ως ένα ευέλικτο και ισχυρό εργαλείο για χρήση σε προβλήματα συσταδοποίησης, όπου το πλήθος των ομάδων των δεδομένων δεν είναι εκ των προτέρων γνωστό.

Μέσω των εφαρμογών που αναπτύχθηκαν, αποδείχθηκε ότι οι Ιεραρχικές Διαδικασίες Dirichlet είναι ικανές να παρέχουν ισχυρά μοντέλα για την ανάλυση και την μοντελοποίηση θεμάτων σε συλλογές κειμένων, αναδεικνύοντας την ικανότητά τους να διαχειρίζονται σύνθετες δομές δεδομένων και να προσαρμόζονται στην πολυπλοκότητα των δεδομένων.

Επιπλέον, αποδείχθηκε η ευελιξία των Γκαουσιανών Διαδικασίων όσον αφορά προβλήματα μη-γραμμικής παλινδρόμησης αλλά και προβλήματα ταξινόμησης.

## 8.2 Μελλοντικές επεκτάσεις

Το ερευνητικό μέλλον των μη-παραμετρικών Μπεϋζιανών μοντέλων επιφυλάσσει πολλά υποσχόμενες κατευθύνσεις για την ενσωμάτωσή τους στα βαθιά νευρωνικά δίκτυα (DNNs), ακόμα και για την αντικατάστασή τους. Ένα τέτοιο παράδειγμα είναι η κατασκευή υβριδικών μοντέλων που συνδυάζουν τα πλεονεκτήματα τόσο των μη-παραμετρικών Μπεϋζιανών μεθόδων όσο και των DNNs. Αυτά τα υβριδικά μοντέλα θα μπορούσαν να αξιοποιήσουν την ευελιξία και την ερμηνευσιμότητα των μη-παραμετρικών Μπεϋζιανών μεθόδων για συγκεκριμένα μέρη του μοντέλου, ενώ θα εκμεταλλεύονται και τις ισχυρές δυνατότητες αναπαράστασης που προσφέρουν τα DNNs για άλλες πτυχές τους. Συνδυάζοντας αυτές τις δύο προσεγγίσεις, τέτοια υβριδικά μοντέλα θα μπορούσαν να αποτελέσουν ένα ισχυρό εργαλείο για την αντιμετώπιση σύνθετων προβλημάτων που απαιτούν τόσο προσαρμοστικότητα στα δεδομένα όσο και την αποτελεσματική αναπαράσταση τους. Αυτά τα προβλήματα μπορούν να πηγάζουν από ένα ευρύ φάσμα αντικειμένων, όπως η επεξεργασία φυσικής γλώσσας, η όραση υπολογιστών και η ενισχυτική μηχανική μάθηση (Moraffah 2024).

# Παράρτημα Α

# Κώδικας R για την εφαρμογή του Metropolis-Hastings σε Γραμμική Παλινδρόμηση

Ακολουθεί υλοποίηση σε R του αλγορίθμου Metropolis-Hastings για την εκ των υστέρων κατανομή των παραμέτρων ενός απλού γραμμικού μοντέλου, όπως περιγράφηκε στο Παράδειγμα 1.4.1.

```r
library(ggplot2)

# Generate synthetic data
set.seed(123)
n <- 100
x <- rnorm(n)
beta_true <- 2
alpha_true <- 3
sigma_true <- 1
y <- beta_true + alpha_true * x + rnorm(n, 0, sigma_true)

# Prior hyperparameters
mu_beta <- 0
sigma_beta <- 10
mu_alpha <- 0
sigma_alpha <- 10
alpha_sigma <- 2
beta_sigma <- 2
# Metropolis algorithm parameters
n_iter <- 5000
chain <- matrix(NA, n_iter, 3)
chain[1, ] <- c(0, 0, 2)  # initial values for beta, alpha, and sigma

for (i in 2:n_iter) {
  current <- chain[i - 1, ]

  # Propose new values
  proposal <- current + rnorm(3, 0, 0.1)

  # Calculate log likelihood for current and proposal
  ll_current <- sum(dnorm(y, current[1] + current[2] * x, current[3],
  log = TRUE))
  ll_proposal <- sum(dnorm(y, proposal[1] + proposal[2] * x, proposal[3],
  log = TRUE))

  # Calculate log prior for current and proposal
```

```
  lp_current <- sum(dnorm(current[1], mu_beta, sigma_beta,
  log = TRUE),
                    dnorm(current[2], mu_alpha, sigma_alpha,
                    log = TRUE),
                    dgamma(current[3]^2, alpha_sigma,
                    beta_sigma, log = TRUE))
  lp_proposal <- sum(dnorm(proposal[1], mu_beta, sigma_beta, log = TRUE),
                     dnorm(proposal[2], mu_alpha, sigma_alpha, log = TRUE),
                     dgamma(proposal[3]^2, alpha_sigma,
                     beta_sigma, log = TRUE))

  # acceptance test
  if (log(runif(1)) < (ll_proposal + lp_proposal - ll_current - lp_current))
  {
    chain[i, ] <- proposal
  } else {
    chain[i, ] <- current
  }
}

# Burn-in
burn_in <- 1000
chain <- chain[-(1:burn_in), ]

chain_df <- as.data.frame(chain)
colnames(chain_df) <- c("beta", "alpha", "sigma")

# Plot for beta0
p_beta <- ggplot(chain_df, aes(x = beta)) +
  geom_histogram(fill = "lightpink3", color = "black", bins = 30) +
  geom_vline(aes(xintercept = mean(beta)), color = "black",
  linetype = "dashed", size = 1) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  labs(title = "Posterior distribution of beta", x = "beta",
  y = "Frequency")
print(p_beta)

# Plot for beta1
p_alpha <- ggplot(chain_df, aes(x = alpha)) +
  geom_histogram(fill = "lightpink3", color = "black", bins = 30) +
  geom_vline(aes(xintercept = mean(alpha)), color = "black",
  linetype = "dashed", size = 1) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  labs(title = "Posterior distribution of alpha", x =
  "alpha", y = "Frequency")
print(p_alpha)

# Plot for sigma
p_sigma <- ggplot(chain_df, aes(x = sigma)) +
  geom_histogram(fill = "lightpink3", color = "black", bins = 30) +
  geom_vline(aes(xintercept = mean(sigma)), color = "black",
  linetype = "dashed", size = 1) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  labs(title = "Posterior distribution of sigma",
  x = "sigma", y = "Frequency")
print(p_sigma)
```

```r
# Trace plots
trace_plot <- ggplot() +
  geom_line(data = chain_df, aes(x = seq_len(nrow(chain_df)), y = beta,
  color = "beta")) +
  geom_line(data = chain_df, aes(x = seq_len(nrow(chain_df)), y = alpha,
  color = "alpha")) +
  geom_line(data = chain_df, aes(x = seq_len(nrow(chain_df)), y = sigma,
  color = "sigma")) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96")) +
  labs(title = "Trace Plots", x = "Iteration", y = "Value", color
  = "Parameter")
print(trace_plot)
```

# Παράρτημα Β

# Κώδικας R για προσομοίωση σημείων από δυσδιάστατη κανονική κατανομή με γνωστή παράμετρο συσχέτισης ρ

```
library(ggplot2)
library(MASS)  # for mvrnorm
library(gridExtra)

# Set the parameters
rho <- 0.8
n_iter <- 500
mu <- c(0, 0)
Sigma <- matrix(c(1, rho, rho, 1), ncol = 2)

# Initialize the chains
set.seed(123)
initial_points <- matrix(c(2.5, 2.5, -2.5,
-2.5, -2.5, 2.5, 2.5, -2.5),
ncol = 2, byrow = TRUE)
chains <- list()

for (i in 1:4) {
  theta <- initial_points[i, ]
  chain <- matrix(NA, nrow = n_iter, ncol = 2)
  for (iter in 1:n_iter) {
    theta[1] <- rnorm(1, mean = mu[1] + rho * (theta[2] - mu[2]),
    sd = sqrt(1 - rho^2))
    theta[2] <- rnorm(1, mean = mu[2] + rho * (theta[1] - mu[1]),
    sd = sqrt(1 - rho^2))
    chain[iter, ] <- theta
  }
  chains[[i]] <- chain
}

# Convert to data frames
chains_df <- do.call(rbind, lapply(1:4, function(i)
data.frame(x = c(initial_points[i, 1], chains[[i]][1:10, 1]),
y = c(initial_points[i, 2], chains[[i]][1:10, 2]), chain =
factor(i), iter = 0:10)))

# Highlight the starting points
starting_points_df <-
```

```
data.frame(x = initial_points[,1], y = initial_points[,2], chain =
factor(1:4))

# Plot the first 10 iterations
p1 <- ggplot(chains_df, aes(x = x, y = y, group = chain)) +
  geom_path(colour="palevioletred") +
  geom_point(data = starting_points_df, aes(x = x, y = y), size = 3,
  shape = 15) +
  xlim(-4, 4) + ylim(-4, 4) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+

  ggtitle("(a) First 10 iterations")

# Plot after 500 iterations
chains_df_full <- do.call(rbind, lapply(1:4, function(i)
data.frame(x = chains[[i]][,1],
y = chains[[i]][,2], chain = factor(i), iter = 1:n_iter)))
p2 <- ggplot(chains_df_full, aes(x = x, y = y, group = chain)) +
  geom_path(alpha = 0.5,colour="palevioletred") +
  xlim(-4, 4) + ylim(-4, 4) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  ggtitle("(b) After 500 iterations")

# Plot the second halves of the sequences
second_half <- chains_df_full[chains_df_full$iter > (n_iter / 2),]
p3 <- ggplot(second_half, aes(x = x, y = y)) +
  geom_point(alpha = 0.5, size = 0.5,colour="palevioletred") +
  xlim(-4, 4) + ylim(-4, 4) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  ggtitle("(c) Second halves of the sequences")

# Arrange the plots in a single row
grid.arrange(p1, p2, p3, ncol = 3)
```

# Παράρτημα C

# Κώδικας R για την υλοποίηση του αλγορίθμου EM

Ακολουθεί υλοποίηση σε R του αλγορίθμου EM για την εκτίμηση των παραμέτρων των δύο συστάδων από τις οποίες προέρχονται τα δεδομένα του Διαγράμματος 2.1.

```
library(mvtnorm)

# Number of  components

K <- 2

# Initialize component probabilites uniformly

pi <- rep(1/K, K)

# Randomly initialize means from the data points

set.seed(1)

m1 <- coords[sample(nrow(coords), 1), ]
m2 <- coords[sample(nrow(coords), 1), ]

# Initialize covariance matrices as identity matrices

sigma1 <- diag(ncol(coords))
sigma2 <- diag(ncol(coords))


# Function to compute the r's

r <- function(data, params) {
  K <- length(params$pi)
  N <- nrow(data)
  r <- matrix(0, nrow = N, ncol = K)

  for (i in 1:N) {
    p <- params$pi[1] * dmvnorm(data[i, ], params$m1, params$sigma1) +
      params$pi[2] * dmvnorm(data[i, ], params$m2, params$sigma2)

    r[i, 1] <- params$pi[1] * dmvnorm(data[i, ], params$m1,
    params$sigma1) / p
    r[i, 2] <- params$pi[2] * dmvnorm(data[i, ], params$m2,
    params$sigma2) / p
  }
```

```
  return(r)
}

# Function to update parameters

update_parameters <- function(data, r) {
  N <- nrow(data)
  K <- ncol(r)

  # Update component probabilites

  pi <- colSums(r) / N

  # Update means

  m1 <- colSums(r[, 1] * data) / sum(r[, 1])
  m2 <- colSums(r[, 2] * data) / sum(r[, 2])

  # Update covariance matrices

  diff1 <- t(t(data) - m1)
  diff2 <- t(t(data) - m2)
  sigma1 <- t(diff1) %*% (r[, 1] * diff1) / sum(r[, 1])
  sigma2 <- t(diff2) %*% (r[, 2] * diff2) / sum(r[, 2])

  return(list(pi = pi, m1 = m1, m2 = m2, sigma1 = sigma1, sigma2 = sigma2))
}

# EM algorithm

em_algorithm <- function(data, K, threshold = 1e-6) {

  params <- list(pi = pi, m1 = m1, m2 = m2, sigma1 = sigma1, sigma2 = sigma2)

  all_iterations<-list(params)

  converged <- FALSE

  i<-0
  while (!converged) {

    old_params <- params
    all_iterations<-append(all_iterations,old_params)
    # E-step

    r <- r(data, params)

    # M-step

    params <- update_parameters(data, r)

    # Check for convergence

    diff <- abs(unlist(params) - unlist(old_params))
    if (sum(diff) < threshold) {
      converged <- TRUE
    }
  }
```

```
  return(list(params,all_iterations))
}

# Apply EM algorithm

results<-em_algorithm(coords, K)
result<-results[[1]]
print("The final estimates for the parameters are the following ")
result
```

# Παράρτημα D

# Κώδικας R για την προσομοίωση δεδομένων από ένα HMM τριών καταστάσεων

Ακολουθεί προσομοίωση σε R δεδομένων από ένα HMM τριών καταστάσεων. Οι παρατηρήσεις προέρχονται από το εξής μοντέλο:

$$p(x_t|\theta_t = 1) \sim \mathcal{N}_1$$
$$p(x_t|\theta_t = 2) \sim \mathcal{N}_2$$
$$p(x_t|\theta_t = 3) \sim \mathcal{N}_3,$$

όπου

$$\mathcal{N}_1$$
$$\Sigma_1 = 10 \begin{pmatrix} 1 & \frac{1}{2} \\ \frac{1}{2} & 1 \end{pmatrix}$$
$$\mathcal{N}_2$$
$$\Sigma_2 = 10 \begin{pmatrix} 1 & -\frac{1}{2} \\ -\frac{1}{2} & 1 \end{pmatrix}$$
$$\mathcal{N}_3$$
$$\Sigma_3 = 10 \begin{pmatrix} 3 & 0 \\ 0 & \frac{1}{2} \end{pmatrix}.$$

Ο πίνακας μετάβασης της μαρκοβιανής αλυσίδας είναι ο

$$A = \begin{pmatrix} 0.8 & 0.1 & 0.1 \\ 0.1 & 0.8 & 0.1 \\ 0.1 & 0.1 & 0.8 \end{pmatrix}.$$

```r
library(MASS)   # for mvrnorm
library(ggplot2)   # for plotting
library(car)   # for dataEllipse

# Function to create HMM model
hmmCreate <- function(type, pi, A, emission) {
  list(type = type, pi = pi, A = A, emission = emission)
}

# Function to sample from HMM
hmmSample <- function(model, T, N) {
```

```
  K <- nrow(model$A)
  D <- model$emission$d
  observed <- array(0, dim = c(D, T, N))
  hidden <- matrix(0, nrow = T, ncol = N)
  for (n in 1:N) {
    hidden[1, n] <- which(rmultinom(1, 1, model$pi)
    == 1)
    observed[, 1, n] <- mvrnorm(1, model$emission$mu[, hidden[1, n]],
    model$emission$Sigma[, , hidden[1, n]])
    for (t in 2:T) {
      hidden[t, n] <- which(rmultinom(1, 1, model$A[hidden[t - 1, n], ]) == 1)
      observed[, t, n] <- mvrnorm(1, model$emission$mu[, hidden[t, n]],
      model$emission$Sigma[, , hidden[t, n]])
    }
  }
  list(observed = observed, hidden = hidden)
}

# Parameters
K <- 3
D <- 2
mu <- 10 * matrix(c(-1, 0, 1, 0, 0, 1), nrow = D, ncol = K)
sf <- 10
Sigma <- array(0, dim = c(2, 2, K))
Sigma[, , 1] <- sf * matrix(c(1, 1/2, 1/2, 1), nrow = 2)
Sigma[, , 2] <- sf * matrix(c(1, -1/2, -1/2, 1), nrow = 2)
Sigma[, , 3] <- sf * matrix(c(3, 0, 0, 1/2), nrow = 2)
A <- matrix(c(0.8, 0.1, 0.1, 0.1, 0.8, 0.1, 0.1, 0.1, 0.8), nrow = K,
byrow = TRUE)
pi <- c(1, 0, 0)

emission <- list(mu = mu, Sigma = Sigma, d = D)
model <- hmmCreate('gauss', pi, A, emission)

# Sample from HMM
T <- 100
sampled <- hmmSample(model, T, 1)
observed <- sampled$observed
hidden <- sampled$hidden

# Convert observed data to data frame
df <- data.frame(
  x = as.vector(observed[1, , ]),
  y = as.vector(observed[2, , ]),
  hidden = factor(hidden)
)



# Plotting the observed data and hidden states with ellipses
ggplot(df, aes(x, y, color = hidden, shape = hidden)) +
  geom_point(size = 2) +
  scale_color_manual(values = c("palevioletred", "limegreen", "darkcyan")) +
  scale_shape_manual(values = c(16, 17, 18)) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  ggtitle("Observed Data with Hidden States") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(color = "Hidden State", shape = "Hidden State") +
```

```
  geom_path(aes(group = 1), color = "black") +
  stat_ellipse(level = 0.95, linetype = 2)

# Plotting the hidden states over time
df_hidden <- data.frame(
  time = 1:T,
  hidden = factor(hidden)
)

ggplot(df_hidden, aes(time, hidden, color = hidden)) +
  geom_point(size = 2) +
  scale_color_manual(values = c("palevioletred", "limegreen", "darkcyan")) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  ggtitle("Hidden States Over Time") +
  theme(plot.title = element_text(hjust = 0.5)) +
  labs(color = "Hidden State")
```

# Παράρτημα Ε

# Κώδικας R για την προσομοίωση μονοπατιών της Γκαουσιανής Διαδικασίας

```r
# Load required libraries
library(ggplot2)
library(MASS)
library(reshape2)
library(gridExtra)

# Function to generate Gaussian process draws
gp_draws <- function(x, length_scale, amplitude, num_draws = 3) {
  n <- length(x)
  cov_matrix <- amplitude^2 *
  exp(-outer(x, x, function(a, b) (a - b)^2) / length_scale^2)
  draws <- mvrnorm(num_draws, mu = rep(0, n), Sigma = cov_matrix)
  return(draws)
}

# Define the x values and parameters
x <- seq(-3, 3, length.out = 100)
params <- list(
  list(tau = 1/2, l = 2),
  list(tau = 1/4, l = 1/2),
  list(tau = 1/2, l = 1/2)
)

# Generate the draws
draws <- lapply(params, function(p) {
  gp_draws(x, length_scale = p$l, amplitude = p$tau)
})

# Function to convert draws to data frame for ggplot
draws_to_df <- function(draws, x) {
  df <- data.frame(x = rep(x, nrow(draws)))
  for (i in 1:nrow(draws)) {
    df[paste0("draw_", i)] <- draws[i, ]
  }
  return(df)
}

# Create data frames for each parameter set
```

```r
dfs <- lapply(draws, draws_to_df, x = x)

# Plotting function
plot_gp_draws <- function(df, params) {
  df_long <- melt(df, id.vars = "x", variable.name = "draw",
  value.name = "mu_x")
  ggplot(df_long, aes(x = x, y = mu_x, group = draw)) +
    geom_line(color = "palevioletred") +
    theme_minimal() +
    theme(panel.background = element_rect(fill = "gray96"))+
    labs(x = "x", y = expression(mu(x))) +
    ggtitle(bquote(tau == .(params$tau) ~ ", " ~ l == .(params$l)))
}

# Create plots
plots <- mapply(plot_gp_draws, dfs, params, SIMPLIFY = FALSE)

# Arrange plots in a grid
grid.arrange(grobs = plots, ncol = 3)
```

# Παράρτημα F

# Υλοποίηση σε Python: Topic Modelling με HDP

## F.1 Συλλογή και επεξεργασία κειμένων

```
import requests
from bs4 import BeautifulSoup
import time

def get_article_urls(base_url, num_pages=10):
    article_urls = []
    for i in range(1, num_pages + 1):
        url = f"{base_url}?page={i}"
        response = requests.get(url)
        soup = BeautifulSoup(response.content, 'html.parser')

        # Find all 'a' tags within 'media-content' divs
        article_tags = soup.find_all('div', class_='media-content')

        # Extract the href attribute (URL) from each 'a' tag
        for tag in article_tags:
            a_tag = tag.find('a')
            if a_tag and 'href' in a_tag.attrs:
                article_urls.append(a_tag['href'])

        time.sleep(2)  # To avoid overwhelming the server with requests
    return article_urls

def get_article_text(url):
    response = requests.get(url)
    soup = BeautifulSoup(response.content, 'html.parser')

    # Extract article text
    paragraphs = soup.find_all('p')
    article_text = ' '.join([para.get_text() for para in paragraphs])
    return article_text

base_url = 'https://www.kathimerini.gr/politics'
article_urls = get_article_urls(base_url, num_pages=6)
print(f"Collected {len(article_urls)} article URLs")

documents = [get_article_text(url) for url in article_urls]
print(f"Collected {len(documents)} articles")
import gensim
```

```
from gensim.utils import simple_preprocess
from gensim.parsing.preprocessing import STOPWORDS
from nltk.corpus import stopwords
import re
import nltk

nltk.download('stopwords')
greek_stopwords = set(stopwords.words('greek'))
# Define additional Greek stopwords
additional_stopwords = set([
    "αλλα", "αν", "αντι", "απο", "αυτα", "αυτεσ", "αυτη", "αυτο", "αυτοι",
    "αυτοσ", "αυτουσ", "αυτων", "εγω", "ειμαι", "ειμαστε", "εισαι", "ειστε",
    "εχει", "εχουν", "εχω", "η", "θα", "ισωσ", "κατα", "με", "μετα", "μη",
    "μην", "να", "ο", "οι", "οπωσ", "οτι", "παρα", "πρεπει", "πωσ", "σε",
    "στη", "στην", "στο", "στον", "τα", "τη", "την", "τι", "την", "τισ",
    "το", "τον", "του", "των", "υπαρχει", "ωσ", "μονο", "παντωσ", "χθεσ",
    "σημερα", "μασ", "οταν", "ακομη", "ακομα", "βασει", "πηγεσ", "πηγων",
    "δηλωσε", "ακομη", "οποιον", "ηταν", "ανεφερε", "τετοιο", "καποια",
    "βρεθει", "βρισκεται", "αναφερουν", "φερεται", "πληροφοριεσ", "καθωσ",
    "ρεπορταζ", "δηλωσεισ", "πρωι", "ωστοσο", "υπο", "κυρ", "υπαρχουν",
    "προκειται", "μαλιστα", "μιασ", "ωστε", "ενοσ", "στουσ", "παντα", "στισ",
    "μια", "ειπε", "μεταξυ", "στα", "αλλων", "ενα", "ωσ", "εωσ", "τουσ",
    "οπωσ", "εχουμε", "στισ", "ωσ", "απ", "οποια", "ετσι", "μπορει", "οποιο",
    "οποιοσ", "πολυ", "πριν", "μπορει", "κατι", "ωσ", "μιλησε", "οποιοσ",
    "γινει", "κανει", "μεχρι", "εκει", "ειτε", "θελει"
])

# Combine the two sets of stopwords
greek_stopwords.update(additional_stopwords)
def preprocess(text):
    text = re.sub(r'\s+', ' ', text)  # Replace multiple
    spaces with single space
    return [word for word in
    simple_preprocess(text, deacc=True) if word not in greek_stopwords]

texts = [preprocess(doc) for doc in documents]
```

## F.2 Εκπαίδευση του μοντέλου HDP

```
from gensim.corpora import Dictionary
from gensim.models import HdpModel
from gensim.utils import simple_preprocess
from gensim.corpora import Dictionary

dictionary = Dictionary(texts)
dictionary.filter_extremes(no_below=2,no_above=0.8)  # Filter out words
that occur in less than 2 documents or more than 80% of the documents
corpus = [dictionary.doc2bow(text) for text in texts]

from gensim.models import HdpModel
from gensim.models import CoherenceModel

gamma_values = [0.1, 1, 10]
alpha_values = [0.1,5,7,10,15]

best_coherence = 0
```

```
best_params = (0, 0, 0)
best_model = None

for gamma in gamma_values:
    for alpha in alpha_values:
            print(f'Training model with gamma={gamma}, alpha={alpha}')
            hdp_model = HdpModel(corpus=corpus, id2word=dictionary,
            T=150, gamma=gamma, alpha=alpha)
            coherence_model_hdp = CoherenceModel(model=hdp_model,
            texts=texts, dictionary=dictionary, coherence='c_v')
            coherence_hdp = coherence_model_hdp.get_coherence()
            print(f'Coherence Score: {coherence_hdp}')

            if coherence_hdp > best_coherence:
                best_coherence = coherence_hdp
                best_params = (gamma, alpha)
                best_model = hdp_model

print(f'Best Coherence Score: {best_coherence}')
print(f'Best Parameters: Gamma={best_params[0]}, Alpha={best_params[1]}')
```

## F.3 Οπτικοποίηση των αποτελεσμάτων

```
from wordcloud import WordCloud
import matplotlib.pyplot as plt

topics = hdp_model.print_topics(num_topics=6)
for idx, topic in topics:
    print(f"Topic #{idx}:")
    for term in topic.split(' + '):
        weight, word = term.split('*')
        print(f"  {word.strip()} ({weight.strip()})")
    print("\n")

colormap = 'PuRd'

# Get topics from the HDP model
topics = hdp_model.show_topics(num_topics=6, formatted=False)

# Loop through the topics and generate word clouds
for i, (topic_id, topic) in enumerate(topics):
    word_freq = dict(topic)
    wordcloud = WordCloud(width=800, height=400,
    background_color='white', colormap=colormap).generate_from_frequencies(word_freq)

    plt.figure()
    plt.imshow(wordcloud, interpolation="bilinear")
    plt.axis("off")
    plt.title(f'Topic {i + 1}')
    plt.savefig(f'Topic_{i + 1}.png', format='png')
    plt.close()
```

# Παράρτημα G

# Κώδικας R για δείγματα από την Single-p και Single-Atom DDP

```r
# Load necessary libraries
library(ggplot2)
library(gridExtra)

# Define the stick-breaking process function
stick_breaking <- function(gamma, num_atoms) {
  v <- rbeta(num_atoms, 1, gamma)  # Beta distributed proportions
  beta <- v * c(1, cumprod(1 - v[-length(v)]))  # Compute the weights
  return(beta)
}

# Define the function to generate the pi distributions conditioned on beta
generate_pi <- function(alpha, beta, num_groups) {
  pi_list <- list()
  for (i in 1:num_groups) {
    v <- rbeta(length(beta), alpha * beta, alpha * (1 - cumsum(beta)))
    pi <- v * c(1, cumprod(1 - v[-length(v)]))
    pi_list[[i]] <- pi
  }
  return(pi_list)
}

# Function to generate atom locations
generate_atoms <- function(num_atoms, num_groups) {
  atom_list <- list()
  for (i in 1:num_groups) {
    atoms <- runif(num_atoms, 0, 10)
    atom_list[[i]] <- atoms
  }
  return(atom_list)
}

# Set parameters
set.seed(123)
gamma <- 2
alpha <- 1
num_atoms <- 5  # Number of atoms for smoother histogram
num_groups <- 3  # Number of groups for the hierarchical process

# Generate beta
beta <- stick_breaking(gamma, num_atoms)
```

```
# Generate different atom locations for each group
atoms_list <- generate_atoms(num_atoms, num_groups)
atoms1 <- atoms_list[[1]]
atoms2 <- atoms_list[[2]]
atoms3 <- atoms_list[[3]]

# Generate pi distributions conditioned on beta for the hierarchical DP
pi_list <- generate_pi(alpha, beta, num_groups)
pi1 <- pi_list[[1]]
pi2 <- pi_list[[2]]
pi3 <- pi_list[[3]]

# Convert to data frames for Single-p Dependent DP
data_single_p <- data.frame(
  atom = c(atoms1, atoms2, atoms3),
  weight = rep(beta, num_groups),

  group = factor(rep(c("G1", "G2", "G3"), each = num_atoms),
  levels = c("G1", "G2", "G3"))
)

# Generate shared atom locations for Hierarchical (Single Atom) DP
shared_atoms <- runif(num_atoms, 0, 10)

# Convert to data frames for Hierarchical (Single Atom) DP
data_single_atom <- data.frame(
  atom = rep(shared_atoms, num_groups),
  weight = c(pi1, pi2, pi3),
  group = factor(rep(c("G1", "G2", "G3"),
  each = num_atoms),
  levels = c("G1", "G2", "G3"))
)

# Custom theme
custom_theme <- theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))

# Plot for Single-p Dependent DP
plot_single_p <- ggplot(data_single_p, aes(x = atom, y = weight)) +
  geom_segment(aes(xend = atom, yend = 0), color = "darkcyan") +
  geom_point(color = "darkcyan") +
  facet_grid(group ~ .) +
  custom_theme +
  labs(title = "Single-p Dependent DP", x = "Atom Location", y = "Weights")

# Plot for Hierarchical (Single Atom) DP
plot_single_atom <- ggplot(data_single_atom, aes(x = atom, y = weight)) +
  geom_segment(aes(xend = atom, yend = 0), color = "palevioletred") +
  geom_point(color = "palevioletred") +
  facet_grid(group ~ .) +
  custom_theme +
  labs(title = "Hierarchical (Single Atom) DDP", x = "Atom Location",
  y = "Weights")

# Align the atom locations for the Hierarchical (Single Atom) DP
plot_single_atom_aligned <- plot_single_atom +
  geom_segment(data = data_single_atom,
  aes(x = atom, xend =
```

```
  atom, y = 0, yend = weight),
               linetype = "dotted", color = "black")

# Combine the plots
grid.arrange(plot_single_atom_aligned, plot_single_p, ncol = 2)
```

# Παράρτημα Η

# Κώδικας R για τον υπολογισμό της ύστερης μέσης τιμής της $f$ για τα δεδομένα του Διαγράμματος 5.2

```r
# Load required libraries
library(ggplot2)
library(tidyr)
library(dplyr)
library(ggplot2)
library(kernlab)
library(MASS)
library(reshape2)


# Generate sample data points
set.seed(12)
x <- seq(-4, 4, length.out = 10)
y <- 2*sin(x)+0.2*x + rnorm(length(x), sd = 0.3) #with noise
y <- 2*sin(x)+0.2*x #without noise
data <- data.frame(x = x, y = y)

# Fit models
linear_model <- lm(y ~ x, data = data)
quadratic_model <- lm(y ~ poly(x, 2), data = data)
fourth_order_model <- lm(y ~ poly(x, 4), data = data)
fourier_model <- lm(y ~ sin(x) + cos(x), data = data)

# Generate data for fitted curves
x_fit <- seq(min(x), max(x), length.out = 100)
data_fit <- data.frame(
  x = x_fit,
  Linear = predict(linear_model, newdata = data.frame(x = x_fit)),
  Quadratic = predict(quadratic_model, newdata = data.frame(x = x_fit)),
  FourthOrder = predict(fourth_order_model, newdata = data.frame(x = x_fit)),
  Fourier = predict(fourier_model, newdata = data.frame(x = x_fit))
)

data_fit_long <- data_fit %>%
  pivot_longer(cols = -x, names_to = "Model", values_to = "y")

# Plotting
ggplot(data) +
```

```
  geom_point(aes(x, y), color = "black", size = 3, shape = 3) +
  geom_line(data = data_fit_long, aes(x, y, color = Model), size = 0.6) +
  scale_color_manual(
    values = c(
      "Linear" = "lightgoldenrod4",
      "Quadratic" = "palevioletred",
      "FourthOrder" = "darkcyan",
      "Fourier" = "deepskyblue4"
    )
  ) +
  labs(
    x = "input, x",
    y = "output, y",
    color = ""
  ) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  theme(
    legend.position = "top",
    legend.title = element_blank(),
    legend.text = element_text(size = 10),
    plot.margin = margin(10, 10, 10, 10)
  )


# Define the squared exponential kernel function
calcSigma <- function(X1, X2, l = 1, sigma_f = 1) {
  Sigma <- matrix(0, nrow = length(X1), ncol = length(X2))
  for (i in 1:nrow(Sigma)) {
    for (j in 1:ncol(Sigma)) {
      Sigma[i, j] <- sigma_f^2 * exp(-0.5 * (abs(X1[i] - X2[j]) / l)^2)
    }
  }
  return(Sigma)
}

# 1. Plot some sample functions from the Gaussian process
x.star <- seq(min(x) - 1, max(x) + 1, len = 200)
sigma <- calcSigma(x.star, x.star)
n.samples <- 3
values <- matrix(0, nrow = length(x.star), ncol = n.samples)
for (i in 1:n.samples) {
  values[, i] <- mvrnorm(1, rep(0, length(x.star)), sigma)
}
values <- cbind(x = x.star, as.data.frame(values))
values <- melt(values, id = "x")

# Plot the result
ggplot(values, aes(x = x, y = value)) +
  geom_line(aes(group = variable),colour = "palevioletred") +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  scale_y_continuous(lim = c(-2.5, 2.5), name = "output, f(x)") +
  xlab("input, x")

# 2. Fit the Gaussian process to the known data points
f <- data.frame(x = x, y = y)

# Calculate the covariance matrices
```

```
x <- f$x
k.xx <- calcSigma(x, x)
k.xxs <- calcSigma(x, x.star)
k.xsx <- calcSigma(x.star, x)
k.xsxs <- calcSigma(x.star, x.star)

# Mean and covariance functions for the prediction
f.star.bar <- k.xsx %*% solve(k.xx) %*% f$y
cov.f.star <- k.xsxs - k.xsx %*% solve(k.xx) %*% k.xxs

# Generate sample functions from the GP posterior
n.samples <- 50
values <- matrix(0, nrow = length(x.star), ncol = n.samples)
for (i in 1:n.samples) {
  values[, i] <- mvrnorm(1, f.star.bar, cov.f.star)
}
values <- cbind(x = x.star, as.data.frame(values))
values <- melt(values, id = "x")

# Data frame for the mean function
mean_values <- data.frame(x = x.star, y = f.star.bar)

# Plot the results including the mean function
ggplot(values, aes(x = x, y = value)) +
  geom_line(aes(group = variable), colour = "rosybrown", alpha = 0.3) +
  geom_line(data = mean_values, aes(x = x, y = y), colour = "palevioletred",
  size = 1) +
  geom_point(data = f, aes(x = x, y = y)) +
  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  scale_y_continuous(lim = c(-3, 3), name = "output, f(x)") +
  xlab("input, x")

# 3. Include noise in the observations
sigma.n <- 0.5  # Standard deviation of the noise
#sigma.n <- 0  #no noise

# Recalculate the mean and covariance functions with noise
f.bar.star <- k.xsx %*% solve(k.xx + sigma.n^2 * diag(ncol(k.xx))) %*% f$y
cov.f.star <- k.xsxs - k.xsx %*% solve(k.xx + sigma.n^2 * diag(ncol(k.xx)))
%*% k.xxs

# Generate sample functions from the GP posterior with noise
values <- matrix(0, nrow = length(x.star), ncol = n.samples)
for (i in 1:n.samples) {
  values[, i] <- mvrnorm(1, f.bar.star, cov.f.star)
}
values <- cbind(x = x.star, as.data.frame(values))
values <- melt(values, id = "x")

# Data frame for the mean function with noise
mean_values_with_noise <- data.frame(x = x.star, y = f.bar.star)
# Data frame for the confidence intervals
conf_intervals <- data.frame(x = x.star,
                             ymin = f.bar.star - 1.96 *
                             sqrt(diag(cov.f.star)),

                             ymax = f.bar.star + 1.96 *
                             sqrt(diag(cov.f.star)))
```

```
# Plot the results including the mean function and noise
ggplot() +
  geom_line(data = values, aes(x = x, y = value, group = variable),
  colour = "grey80") +
  geom_ribbon(data = conf_intervals, aes(x = x, ymin = ymin, ymax = ymax),
  fill = "rosybrown1", alpha = 0.3) +
  geom_line(data = mean_values_with_noise, aes(x = x, y = y),
  colour = "palevioletred", size = 0.7) +
  geom_point(data = f, aes(x = x, y = y), colour = "black", size = 3,
  shape = 3) +

  theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))+
  scale_y_continuous(lim = c(-3.5, 3.5), name = "output, f(x)") +
  xlab("input, x")
```

# Παράρτημα I

# Κώδικας R για το σύνολο δεδομένων Old Faithful

## I.1 Ανάλυση δεδομένων

```r
library(ggplot2)
library(dplyr)

# Load the Old Faithful dataset
data("faithful")

# Custom theme
custom_theme <- theme_minimal() +
  theme(panel.background = element_rect(fill = "gray96"))

# Summary of the dataset
summary(faithful)

# Scatter plot: Eruptions vs. Waiting time
ggplot(faithful, aes(x = waiting, y = eruptions)) +
  geom_point(color = "palevioletred", size = 3, alpha = 0.7) +
  labs(title = "Old Faithful Eruptions: Waiting Time vs. Duration",
       x = "Waiting Time (minutes)",
       y = "Eruption Duration (minutes)") +
  custom_theme

# Histogram of Waiting Times
ggplot(faithful, aes(x = waiting)) +
  geom_histogram(binwidth = 1, fill = "darkcyan", color = "white",
  alpha = 0.8) +
  labs(title = "Histogram of Waiting Times",
       x = "Waiting Time (minutes)",
       y = "Frequency") +
  custom_theme

# Density plot of Waiting Times
ggplot(faithful, aes(x = waiting)) +
  geom_density(fill = "palevioletred", color = "darkcyan", alpha = 0.7) +
  labs(title = "Density Plot of Waiting Times",
       x = "Waiting Time (minutes)",
       y = "Density") +
  custom_theme

# Boxplot: Waiting time distribution
```

```
ggplot(faithful, aes(y = waiting)) +
  geom_boxplot(fill = "palevioletred", color = "darkcyan", alpha = 0.7) +
  labs(title = "Boxplot of Waiting Times",
       y = "Waiting Time (minutes)") +
  custom_theme
```

## I.2 Μοντελοποίηση του χρόνου αναμονής

```
library(dirichletprocess)
faithfulTrans <- (faithful$waiting - mean(faithful$waiting))/sd(faithful$waiting)
dp <- DirichletProcessGaussian(faithfulTrans)
dp <- Fit(dp, 1000)
data.frame(Weights=dp$weights,
           mu=c(dp$clusterParameters[[1]]),
           sigma=c(dp$clusterParameters[[2]]))
AlphaTraceplot(dp, gg = TRUE)
AlphaPriorPosteriorPlot(
  dp,
  prior_color = "palevioletred",
  post_color = "gray1",
  gg = TRUE
)
ClusterTraceplot(dp, gg = TRUE)
LikelihoodTraceplot(dp, gg = TRUE)
```

## I.3 Συσταδοποίηση των εκρήξεων με χρήση DPMM

```
faithfulTrans <- scale(faithful)
dp <- DirichletProcessMvnormal(faithfulTrans)
dp <- Fit(dp, 500)
# Custom plot function with specified colors
plot_custom_dp <- function(dp) {
  cluster_labels <- dp$clusterLabels
  data <- as.data.frame(dp$data)
  colnames(data) <- c("Feature1", "Feature2")
  data$cluster <- factor(cluster_labels)

  ggplot(data, aes(x = Feature1, y = Feature2, color = cluster)) +
    geom_point(size = 3, alpha = 0.8) +
    scale_color_manual(values = c("palevioletred", "darkcyan", "gold",
    "steelblue")) +
    theme_minimal() +
    labs(title = "DPMM Clustering",
         x = "Waiting Times (minutes)",
         y = "Eruption Duration (minutes)") +
    theme(panel.background = element_rect(fill = "gray96"))
}

# Plot the result
plot_custom_dp(dp)
```

# Παράρτημα J

# Κώδικας R για το σύνολο δεδομένων Air Quality

## J.1 Ανάλυση δεδομένων

```
library(ggplot2)
library(dplyr)
library(gridExtra)

# Load the airquality dataset
data("airquality")

# Remove NAs
airquality_clean <- na.omit(airquality)

# Summary statistics
summary(airquality_clean)

custom_theme <- theme_minimal() +
    theme(
        plot.title = element_text(hjust = 0.5, color = "palevioletred", size =
        16, face = "bold"),
        axis.title = element_text(color = "palevioletred", size = 12, face =
        "bold"),
        axis.text = element_text(color = "palevioletred"),
        panel.grid.major = element_line(color = "palevioletred", linetype =
        "dotted"),
        panel.grid.minor = element_line(color = "palevioletred", linetype =
        "dotted"),
        strip.background = element_rect(fill = "palevioletred", color =
        "palevioletred"),
        strip.text = element_text(color = "white", face = "bold")
    )

# Pairwise scatter plots
pairwise_plot <- ggpairs(airquality_clean,
                         mapping = ggplot2::aes(color = "palevioletred"),
                         lower = list(continuous = "smooth"),
                         upper = list(continuous = wrap("cor", size = 4)),
                         diag = list(continuous = wrap("densityDiag", color =
                         "palevioletred"))
) + custom_theme

# Scatter plots of Ozone vs other variables
```

```
scatter_plots <- airquality_clean %>%
    gather(key = "Predictor", value = "Value", -Ozone, -Month, -Day) %>%
    ggplot(aes(x = Value, y = Ozone, color = "palevioletred")) +
    geom_point(alpha = 0.6) +
    facet_wrap(~ Predictor, scales = "free_x") +
    geom_smooth(method = "lm", se = FALSE, color = "palevioletred") +
    labs(title = "Scatter Plots: Ozone vs Other Variables", x = "Value", y =
    "Ozone") +
    custom_theme +
    theme(legend.position = "none")

# Arrange all plots in a grid layout
grid.arrange(hist_plot, boxplot_plot, scatter_plots, nrow = 3)
```

## J.2 Προσαρμογή μοντέλου GP

```
X <- airquality_clean[, c("Solar.R", "Wind", "Temp")]
Y <- airquality_clean$Ozone

# Split data into training and testing sets
set.seed(123)  # For reproducibility
train_index <- createDataPartition(Y, p = 0.7, list = FALSE)
X_train <- X[train_index, ]
Y_train <- Y[train_index]
X_test <- X[-train_index, ]
Y_test <- Y[-train_index]

# Normalize the data (important for GPR)
X_train_scaled <- scale(X_train)
X_test_scaled <- scale(X_test, center = attr(X_train_scaled, "scaled:center"),
scale = attr(X_train_scaled, "scaled:scale"))

# Define grids for hyperparameters
length_scales <- seq(0.1, 2, by = 0.1)
sigma_values <- seq(0.1, 2, by = 0.1)
kernels <- c("SE", "Matern", "RQ")

# Initialize variables to store RMSE results
rmse_results <- data.frame(Kernel = character(), LengthScale = numeric(),
Sigma
= numeric(), RMSE = numeric(), stringsAsFactors = FALSE)

# Function to train and evaluate GPR model with different l and sigma
evaluate_hyperparameters <- function(kernel, length_scales, sigma_values,
X_train_scaled, Y_train, X_test_scaled, Y_test) {
  best_rmse <- Inf
  best_length_scale <- NULL
  best_sigma <- NULL

  for (length_scale in length_scales) {
    for (sigma in sigma_values) {
      tryCatch({
        gp_model <- GauPro(X_train_scaled, Y_train, kernel = kernel,
        lengthScale = length_scale, sigma = sigma, parallel = FALSE)

        # Make predictions
```

```
        predictions <- predict(gp_model, X_test_scaled)

        # Compute RMSE
        rmse <- sqrt(mean((predictions - Y_test) ^ 2))

        # Update the best parameters if the current RMSE is lower
        if (rmse < best_rmse) {
          best_rmse <- rmse
          best_length_scale <- length_scale
          best_sigma <- sigma
        }
      }, error = function(e) {
        # Handle errors
        cat("Error with kernel:", kernel, "length scale:", length_scale,
        "sigma:", sigma, "\n")
        cat("Error message:", e$message, "\n")
      })
    }
  }

  # Store results
  return(data.frame(Kernel = kernel, LengthScale = best_length_scale, Sigma =
  best_sigma, RMSE = best_rmse, stringsAsFactors = FALSE))
}

# Evaluate each kernel
for (kernel in kernels) {
  results <- evaluate_hyperparameters(kernel, length_scales, sigma_values,
  X_train_scaled, Y_train, X_test_scaled, Y_test)
  rmse_results <- rbind(rmse_results, results)
}

# Print RMSE results for each kernel
print(rmse_results)

# Train final models with best hyperparameters for each kernel and make
predictions
final_predictions <- list()
for (i in 1:nrow(rmse_results)) {
  kernel <- rmse_results$Kernel[i]
  length_scale <- rmse_results$LengthScale[i]
  sigma <- rmse_results$Sigma[i]

  tryCatch({
    # Train final model with best hyperparameters
    gp_model_final <- GauPro(X_train_scaled, Y_train, kernel = kernel,
    lengthScale = length_scale, sigma = sigma, parallel = FALSE)

    # Make predictions
    final_predictions[[paste(kernel, length_scale, sigma, sep = "_")]] <-
    predict(gp_model_final, X_test_scaled)
  }, error = function(e) {
    cat("Error with final model for kernel:", kernel, "\n")
    cat("Error message:", e$message, "\n")
  })
}

# Combine results
results_df <- do.call(rbind, lapply(names(final_predictions), function(name) {
```

```
  kernel <- strsplit(name, "_")[[1]][1]
  data.frame(
    Actual = Y_test,
    Predicted = final_predictions[[name]],
    Kernel = kernel
  )
}))

# Plot predictions vs. actual values for each kernel and
hyperparameter combination
ggplot(results_df, aes(x = Actual, y = Predicted, color = Kernel)) +
  geom_point() +
  geom_abline(intercept = 0, slope = 1, color = "black", linetype = "dashed") +
  facet_wrap(~ Kernel) +
  ggtitle("Gaussian Process Regression: Actual vs Predicted by Kernel and
  Hyperparameters") +
  xlab("Actual Ozone Levels") +
  ylab("Predicted Ozone Levels") +
  theme_minimal()

# Fit a Linear Regression Model
lm_model <- lm(Ozone ~ Solar.R + Temp + Wind, data =
airquality_clean[train_indices, ])
lm_pred <- predict(lm_model, newdata = airquality_clean[-train_indices, ])

# Calculate RMSE for Linear Regression Model
lm_rmse <- sqrt(mean((Y_test - lm_pred)^2))
print(paste("Linear Regression RMSE:", lm_rmse))
```

# Παράρτημα Κ

# Υπολογιστικά Πακέτα

## Κ.1 Βιβλιοθήκη *gensim* (Python)

Το gensim είναι ένα δημοφιλές πακέτο στην Python για την επεξεργασία φυσικής γλώσσας (Natural Language Processing - NLP) και ειδικότερα για τη μοντελοποίηση θεμάτων (topic modeling) και τη διανυσματική αναπαράσταση λέξεων. Το Gensim έχει βελτιστοποιηθεί για να επεξεργάζεται μεγάλους όγκους κειμένων αποδοτικά, καθιστώντας το κατάλληλο για εφαρμογές σε δεδομένα μεγάλης κλίμακας. Το Gensim ενσωματώνει μεθόδους που χρησιμοποιούν αποδοτικούς αλγορίθμους για την εκπαίδευση μοντέλων με χαμηλές απαιτήσεις σε μνήμη, παρέχοντας παράλληλα ευέλικτες δυνατότητες για την επεξεργασία κειμένων και τη μοντελοποίηση δεδομένων.

## Κ.2 Πακέτο *dirichletprocess* (R)

Το πακέτο dirichleprocess της R επιτρέπει την εφαρμογή μη παραμετρικών μεθόδων βασισμένων στη Διαδικασία Dirichlet. Χρησιμοποιείται για μοντέλα που επιτρέπουν έναν απεριόριστο αριθμό παραμέτρων ή ομάδων, παρέχοντας έτσι μεγαλύτερη ευελιξία στη μοντελοποίηση δεδομένων σε σύγκριση με τις παραδοσιακές παραμετρικές προσεγγίσεις. Το πακέτο υποστηρίζει δειγματοληψία μέσω αλγορίθμων όπως ο Gibbs sampling και ο Metropolis-Hastings, διευκολύνοντας την εκτίμηση της εκ των υστέρων κατανομής των παραμέτρων. Επίσης, προσφέρει την δυνατότητα για adaptive hierarchical modeling, επιτρέποντας τη χρήση μοντέλων που αυξάνουν τη χωρητικότητά τους με την εισαγωγή περισσότερων δεδομένων, χωρίς προαπαιτούμενο όριο για τον αριθμό των clusters.

## Κ.3 Πακέτο *GauPro* (R)

Το πακέτο GauPro της R χρησιμοποιείται για παλινδρόμηση με Γκαουσιανές Διεργασίες (Gaussian Process Regression), προσφέροντας μια μη παραμετρική, στοχαστική προσέγγιση για την παλινδρόμηση. Υποστηρίζει πολλούς τύπους πυρήνων όπως ο Radial Basis Function (RBF), ο Matern, ο Exponential. Η ρύθμιση των υπερπαραμέτρων των πυρήνων (π.χ., length scale, variance) πραγματοποιείται μέσω μεγίστοποίησης της περιθώριας πιθανοφάνειας, όπως αυτή περιγράφηκε προηγουμένως (σ.σ: Το πακέτο *GauPro* της R αρχειοθετήθηκε από το CRAN repository μετά την ολοκλήρωση της παρούσας εργασίας, λόγω ελλιπούς συντήρησης του. Προηγούμενες εκδόσεις είναι διαθέσιμες εδώ, ενώ ένα μοντέλο με αντίστοιχες δυνατότητες είναι το *GPy* της Python).

# Βιβλογραφικές Αναφορές

[1]L. E. Baum, T. Petrie, G. Soules, and N. Weiss, "Maximization techniques occurring in the statistical analysis of probabilistic functions of markov chains", The Annals of Mathematical Statistics **41**, 164–171 (1970).

[2]M. Beal, Z. Ghahramani, and C. Rasmussen, "The infinite hidden markov model", in Advances in neural information processing systems, Vol. 14, edited by T. Dietterich, S. Becker, and Z. Ghahramani (2001).

[3]D. Blackwell and J. B. MacQueen, "Ferguson distributions via pólya urn schemes", The annals of statistics **1**, 353–355 (1973).

[4]M. De Iorio, P. Müller, G. L. Rosner, and S. N. MacEachern, "An anova model for dependent random measures", Journal of the American Statistical Association **99**, 205–215 (2004).

[5]M. D. Escobar and M. West, "Bayesian density estimation and inference using mixtures", Journal of the American Statistical Association **90**, 577–588 (1995).

[6]W. J. Ewens, "Population genetics theory - the past and the future", in (1990).

[7]T. S. Ferguson, "A Bayesian Analysis of Some Nonparametric Problems", The Annals of Statistics **1**, 209 –230 (1973).

[8]B. J. Frey, "Extending factor graphs so as to unify directed and undirected graphical models", in Proceedings of the 19th conference on uncertainty in artificial intelligence (uai2003) (2003), pp. 257–264.

[9]A. Gelman, J. Carlin, H. Stern, D. Dunson, A. Vehtari, and D. Rubin, *Bayesian data analysis: texts in statistical science*, 3η έκδοση (CRC Press, 2013).

[10]W. Hastings, "Monte carlo sampling methods using markov chains and their applications", Biometrika **57**, 97–109 (1970).

[11] Jeffreys, "An invariant form for the prior probability in estimation problems", Proceedings of the Royal Society of London. Series A, Mathematical and Physical Sciences **186**, 186 (1946).

[12]A. Kottas and M. Krnjajić, *Bayesian nonparametric mixture modelling: methods and applications*, Presentation, Presented at the National University of Ireland, Galway, Galway, Ireland: National University of Ireland, Galway, 2009.

[13]A. Kottas, P. Müller, and F. Quintana, "Nonparametric bayesian modeling for multivariate ordinal data", Journal of Computational and Graphical Statistics **14**, 610–625 (2005).

[14]A. Kottas, Z. Wang, and A. Rodríguez, "Spatial modeling for risk assessment of extreme values from environmental time series: a bayesian nonparametric approach", Environmetrics **23**, 649–662 (2012).

[15]S. N. MacEachern, "Dependent dirichlet processes", in Proceedings of the workshop on nonparametric bayesian methods (1999).

[16]T. Merritt, P. R. Buerkner, and M. Galili, *Dirichletprocess: fit dirichlet process objects*, R package version 0.5.0 (2022).

[17]N. Metropolis and S. Ulam, "The monte carlo method", Journal of the American Statistical Association **44**, 335–341 (1949).

[18]N. Metropolis, A. W. Rosenbluth, M. N. Rosenbluth, A. H. Teller, and E. Teller, "Equation of state calculations by fast computing machines", The Journal of Chemical Physics **21**, 1087–1092 (1953).

[19]J. W. Miller and M. T. Harrison, "Inconsistency of pitman–yor process mixtures for the number of components", Journal of Machine Learning Research **15**, 3333–3370 (2014).

[20]D. Mimno, H. Wallach, E. Talley, M. Leenders, and A. McCallum, "Optimizing semantic coherence in topic models", in Proceedings of the 2011 conference on empirical methods in natural language processing (Association for Computational Linguistics, 2011), pp. 262–272.

[21]T. P. Minka, "Expectation propagation for approximate bayesian inference", Proceedings of the 17th Conference on Uncertainty in Artificial Intelligence, 362–369 (2001).

[22]B. Moraffah, "Bayesian nonparametrics: an alternative to deep learning", arXiv preprint arXiv:2404.00085 (2024).

[23]K. P. Murphy, *Probabilistic machine learning: an introduction* (MIT Press, 2022).

[24]National Park Service, *Old faithful geyser data*, Available in R datasets package, 1935.

[25]R. M. Neal, "Markov chain sampling methods for dirichlet process mixture models", Journal of Computational and Graphical Statistics **9**, 249–265 (2000).

[26]A. O'Hagan and J. F. C. Kingman, "Curve fitting and optimal design for prediction", Journal of the Royal Statistical Society: Series B (Methodological) **40**, 1–42 (1978).

[27]C. E. Rasmussen and C. K. I. Williams, *Gaussian processes for machine learning* (MIT Press, Cambridge, MA, USA, 2006).

[28]C. Rudin, *Gaussian mixture models and expectation maximization - duke course notes*, 2020.

[29]J. Sethuraman, "A constructive definition of dirichlet priors", Statistica sinica, 639–650 (1994).

[30]M. A. Taddy, "Autoregressive mixture models for dynamic spatial poisson processes: application to tracking intensity of violent crime", Journal of the American Statistical Association **105**, 1403–1417 (2010).

[31]Y. Teh, K. Kurihara, and M. Welling, "Collapsed variational inference for hdp", in Advances in neural information processing systems, Vol. 20, edited by J. Platt, D. Koller, Y. Singer, and S. Roweis (2007).

[32]Y. W. Teh, M. I. Jordan, M. J. Beal, and D. M. Blei, "Hierarchical dirichlet processes", en, J. Am. Stat. Assoc. **101**, 1566–1581 (2006).

[33]A. J. Viterbi, "Error bounds for convolutional codes and an asymptotically optimum decoding algorithm", IEEE Transactions on Information Theory **13**, 260–269 (1967).

[34]C. Wang, J. Paisley, and D. M. Blei, "Online variational inference for the hierarchical dirichlet process", in Proceedings of the fourteenth international conference on artificial intelligence and statistics, Vol. 15, edited by G. Gordon, D. Dunson, and M. Dudík, Proceedings of Machine Learning Research (2011), pp. 752–760.

[35]M. West, *Hyperparameter estimation in dirichlet process mixture models*, Discussion Paper 92-A03 (Duke University, Institute of Statistics and Decision Sciences, Durham, NC, 1992).

[36]C. K. Williams and D. Barber, "Bayesian classification with gaussian processes", IEEE Transactions on Pattern Analysis and Machine Intelligence **20**, 1342–1351 (1998).

[37]X. Zhao and S. N. Wood, *Gaupro: gaussian process regression for r*, R package version 1.6.1, 2022.