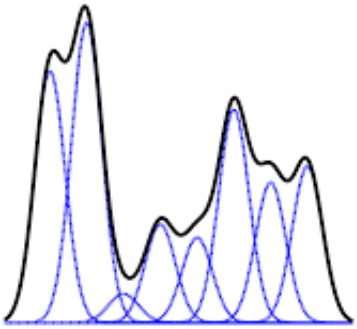
National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Learning Mixtures of Selective Mallows Models

Diploma Thesis

of

POLLATOS VASILEIOS



Supervisor: Dimitris Fotakis
Associate Professor

Athens, March 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

Learning Mixtures of Selective Mallows Models

Diploma Thesis

of

POLLATOS VASILEIOS

Supervisor: Dimitris Fotakis
Associate Professor

Approved by the examination committee on 14th March 2022.

(Signature)	*(Signature)*	*(Signature)*
.....	.....	.....
Dimitris Fotakis	Aris Pagourtzis	Nikolaos Papaspyrou
Associate Professor	Professor	Professor

Athens, March 2022

National Technical University of Athens
School of Electrical and Computer Engineering
Division of Computer Science

DISCLAIMER ON ACADEMIC ETHICS AND INTELLECTUAL PROPERTY RIGHTS

Being fully aware of the implications of copyright laws, I expressly state that this diploma thesis, as well as the electronic files and source codes developed or modified in the course of this thesis, are solely the product of my personal work and do not infringe any rights of intellectual property, personality and personal data of third parties, do not contain work / contributions of third parties for which the permission of the authors / beneficiaries is required and are not a product of partial or complete plagiarism, while the sources used are limited to the bibliographic references only and meet the rules of scientific citing. The points where I have used ideas, text, files and / or sources of other authors are clearly mentioned in the text with the appropriate citation and the relevant complete reference is included in the bibliographic references section. I fully, individually and personally undertake all legal and administrative consequences that may arise in the event that it is proven, in the course of time, that this thesis or part of it does not belong to me because it is a product of plagiarism.

(Signature)

.....
Pollatos Vasileios

14th March 2022

Περίληψη

Στην παρούσα εργασία μελετάμε το πρόβλημα της εκμάθησης μιγμάτων κατανομών κατάταξης με χρήση θορυβωδών ελλιπών δειγμάτων. Οι κατανομές κατάταξης έχουν κερδίσει ενδιαφέρον στα πεδία της θεωρίας των κοινωνικών επιλογών και της θεωρητικής μηχανικής μάθησης εδώ και δεκαετίες. Πέρα από την εκτεταμένη θεωρητική έρευνα, οι κατανομές κατάταξης έχουν διάφορες εφαρμογές στον πραγματικό κόσμο, όπως στο crowdsourcing, στα συστήματα ψηφοφορίας, τα recommendation systems και την αναζήτηση στο διαδίκτυο. Κατά τη συνάθροιση δειγμάτων κατάταξης χρησιμοποιούμε μια συλλογή δειγμάτων που προέρχονται από έναν συγκεκριμένο πληθυσμό και προσπαθούμε να εκτιμήσουμε το υποκείμενο "ground truth" σχετικά με τις προτιμήσεις του πληθυσμού πάνω σε ένα σύνολο n στοιχείων. Τα δείγματα είναι είτε πλήρη είτε ελλιπή. Στην πρώτη περίπτωση κάθε δείγμα είναι μια μετάθεση του πλήρους συνόλου των n στοιχείων ενώ στη δεύτερη περίπτωση κάθε δείγμα είναι μια μετάθεση κάποιου υποσυνόλου του πλήρους συνόλου των n στοιχείων. Το πρώτο μας βήμα είναι να θεωρήσουμε ένα είδος μοντέλου που παράγει τα δείγματα, ώστε το πρόβλημά μας να είναι καλά διατυπωμένο και να είναι δυνατή η βελτιστοποίηση. Δημοφιλή μοντέλα είναι τα Mallows, Plackett Luce και το Repeated Insertion Model. Στην εργασία μας εστιάζουμε στο μοντέλο Mallows και ιδιαίτερα στην selective εκδοχή του, όπου τα δείγματα είναι ελλιπή. Μια περαιτέρω γενίκευση του κλασικού μοντέλου Mallows είναι να υποθέσουμε ότι η κατανομή κατάταξης που παράγει τα δείγματα είναι ένα μείγμα k μοντέλων Mallows και όχι ένα μεμονωμένο. Αυτή η υπόθεση μοντελοποιεί την ετερογένεια των προτιμήσεων ενός πληθυσμού διαιρώντας τον σε πολλές ομάδες (π.χ. γυναίκες και άνδρες). Σε αυτή την εργασία μελετάμε την εφικτότητα εκμάθησης του selective μοντέλου Mallows και προτείνουμε αλγόριθμους για την εκτίμηση της κατανομής και (όποτε είναι δυνατόν) την εκτίμηση των παραμέτρων. Προτείνουμε αλγόριθμους που λειτουργούν στη γενική περίπτωση και για την ειδική περίπτωση όπου τα κέντρα είναι καλά διαχωρισμένα δείχνουμε ότι υπάρχουν πολύ πιο αποδοτικοί αλγόριθμοι. Παρέχουμε εγγυήσεις για τη συμπεριφορά των προτεινόμενων αλγορίθμων καθώς και πειραματικά αποτελέσματα.

Λέξεις Κλειδιά

Συνάθροιση Δειγμάτων Κατάταξης, Μίγματα Κατανομών Κατάταξης, Μάθηση από Ελλιπή Δείγματα, Εκμάθηση Κατανομών, Εκτίμηση Παραμέτρων, Μέθοδος Ροπών, Συσταδοποίηση.

Abstract

In this thesis we study the problem of learning mixtures of rankings using noisy incomplete samples. Ranking distributions have drawn interest in the fields of social choice theory and theoretical machine learning for many decades. Apart from the extensive theoretical research ranking distributions have various real world applications including crowdsourcing, voting and recommendation systems and web search. Ranking aggregation is about using a collection of ranking samples drawn from a certain population in order to estimate the underlying ground truth about the preferences of the population on a set of n items. The samples are either complete or incomplete. In the first case each sample is a permutation of the full range of n items whereas in the second case each sample is a permutation of some subset of the full set of n items. The first step in our setting is to assume a generative model so our problem is well formulated and optimisation is possible. Popular generative models are the Mallows Model, the Plackett Luce Model and The Repeated Insertion Model. In our work we focus on the Mallows Model and particularly on its selective variation, where samples are incomplete. A further generalisation of the classical Mallows model is to assume that the underlying ranking distribution is a mixture of k Mallows models rather than a single one. This assumption models the heterogeneity of the preferences of a population by dividing it into several clusters (e.g women and men). In this work we study the identifiability of the Selective Mallow Mixture Model and suggest algorithms for distribution estimation and (when possible) parameter estimation. We suggest algorithms that work in the general case and for the specific case where centers are well separated we show that there exist much more efficient ones. We provide provable guarantees for the behavior of the suggested algorithms as well as experimental results.

Keywords

Ranking Aggregation, Mallow Mixture Model, Selective Mallows Model, Distribution Learning , Parameter Estimation, Method Of Moments, Clustering.

to my parents

Ευχαριστίες

Με την εργασία αυτή ολοκληρώνεται ο κύκλος των προπτυχιακών σπουδών μου στο ΕΜΠ και ξεκινά μια νέα πορεία, αυτή του διδακτορικού. Θα ήθελα να ευχαριστήσω τον κύριο Φωτάκη ως επιβλέποντα της διπλωματικής αλλά και ως καθηγητή που μέσα απο τα μαθήματα του με έκανε να αγαπήσω τη θεωρητική επιστήμη υπολογιστών. Η στήριξη και καθοδήγηση του κου Φωτάκη αλλά και του Κώστα Σταυρόπουλου και Άλκη Καλαβάση κατά την εκπόνηση της διπλωματικής ήταν καθοριστικές για την επιτυχή ολοκλήρωση αυτού του εγχειρήματος που μου έμαθε πολλά πράγματα και με μύησε στη θεωρητική έρευνα. Θα ήθελα να ευχαριστήσω τους παραπάνω καθώς και τον Αργύρη Μουζάκη και για το κομμάτι των αιτήσεων, στο οποίο με βοήθησαν να προσανατολιστώ και να πάρω σημαντικές αποφάσεις. Θα ήθελα επίσης να εκφράσω την ευγνωμοσύνη μου σε όλα τα μέλη του Corelab για όσα κάνουν για να κρατάνε ζωντανή την κοινότητα και την έρευνα στη θεωρητική πληροφορική έχοντας δημιουργήσει ένα πολύ όμορφο και επικοιδομητικό πλαίσιο (συν)εργασίας. Ευχαριστώ τα μέλη της επιτροπής, τον κο Παγουρτζή για όσα μου έδωσε μέσα από τα μαθήματα του αλλά και για τον κεντρικό ρόλο του στο εργαστήριο και τον κο Παπασπύρου για την εξαιρετική δουλειά του στα μαθήματα που διδάσκει αλλά και όσα προσφέρει σε τεχνικό και συντονιστικό επίπεδο σε διάφορα θέματα της σχολής. Γενικότερα, θα ήθελα να ευχαριστήσω το σύνολο των καθηγητών του ΕΜΠ για το υγιές ακαδημαϊκό πνεύμα που καλλιεργούν και τη συνεισφορά τους στην εκπαιδευτική διαδικασία αλλά και των φοιτητών που ανταποκρίνονται επάξια στο κάλεσμα αυτό. Τέλος, ένα μεγάλο ευχαριστώ στους φίλους και την οικογένεια μου για τη συμπαράστασή τους σε όλη τη διάρκεια των σπουδών και τις όμορφες στιγμές που περάσαμε μαζί αυτά τα χρόνια.

Athens, March 2022

Pollatos Vasileios

Table of Contents

List of Figures

Εκτεταμένη Ελληνική Περίληψη

Συνάθροιση Δειγμάτων Κατάταξης

Η θεωρία κοινωνικής επιλογής είναι ένα πλαίσιο για την ανάλυση της συνάθροισης ατομικών απόψεων, προτιμήσεων, ενδιαφερόντων ή συμφερόντων για την επίτευξη μιας συλλογικής απόφασης. Η θεωρία της κοινωνικής επιλογής χρονολογείται από τη διατύπωση του Μαρκήσιου ντε Κοντορσέ για το παράδοξο της ψήφου (τέλη 18ου αιώνα). Το παράδοξο Condorcet είναι μια κατάσταση στην οποία οι κοινωνικές προτιμήσεις μπορεί να είναι κυκλικές, ακόμα κι αν οι προτιμήσεις των ατόμων είναι άκυκλες. Σε εκλογές με δύο μόνο υποψηφίους, όπου κάθε ψηφοφόρος έχει προτίμηση για έναν υποψήφιο έναντι του άλλου, ο κανόνας επιλογής της πλειοψηφίας λειτουργεί σωστά, δίνοντας μια κατάταξη των δύο υποψηφίων που συμφωνεί με τις προτιμήσεις της πλειοψηφίας των ψηφοφόρων και είναι συνεπής με τον εαυτό της. Ωστόσο, αυτό δεν είναι πάντα δυνατό όταν ο αριθμός των υποψηφίων υπερβαίνει τους δύο. Ένα παράδειγμα του παραδόξου είναι το εξής:

Ας υποθέσουμε ότι έχουμε τους υποψήφιους Α, Β και Γ και τρεις ψηφοφόρους. Ο παρακάτω πίνακας παρουσιάζει τις ατομικές προτιμήσεις των ψηφοφόρων.

Ατομικές προτιμήσεις			
Ψηφοφόρος	Πρώτη προτίμηση	Δεύτερη προτίμηση	Τρίτη προτίμηση
Ψηφοφόρος 1	A	B	Γ
Ψηφοφόρος 2	B	C	A
Ψηφοφόρος 3	C	A	B

Η πλειοψηφία των ψηφοφόρων προτιμά το Α από το Β, το Β από το Γ και το Γ από το Α. Η προκύπτουσα συλλογική προτίμηση Α>Β>Γ>Α είναι κυκλική και επομένως ασυνεπής. Αυτό το παράδοξο αναδεικνύει την ανάγκη για πιο σύνθετους και ισχυρούς μηχανισμούς ψηφοφορίας, όπως η ψηφοφορία με σκορ. Ένα άλλο ενδιαφέρον ερώτημα που προκύπτει είναι εάν ο μηχανισμός ψηφοφορίας είναι φιλαλήθης, δηλαδή εάν οι ψηφοφόροι έχουν κίνητρο να δώσουν ψήφο που δεν συμφωνεί πλήρως με τις ατομικές τους πεποιθήσεις προκειμένου να προωθήσουν ένα συγκεκριμένο αποτέλεσμα των εκλογών. Ωστόσο, αυτή η οπτική του προβλήματος της ψηφοφορίας σχετίζεται περισσότερο με τη θεωρία παιγνίων και είναι εκτός του πεδίου αυτής της εργασίας.

Ο κανόνας του Kemeny είναι ένας πιο ουσιαστικός και αποτελεσματικός τρόπος για τη συνάθροιση δειγμάτων κατάταξης. Δεδομένου ενός δείγματικού προφίλ $\{\sigma_1, \sigma_2, ...\sigma_N\} \in \mathbb{S}_n^N$,

ο κανόνας του Kemeny επιλέγει την ακόλουθη κατάταξη τ ως εκτίμηση της συλλογικής προτίμησης: $\tau = argmin_{\tau \in \mathbb{S}_n} \sum_{i=1}^{N} d_{KT}(\tau, \sigma_i)$. Αυτή η πράξη μπορούμε να πούμε ότι βρίσκει τη διάμεσο των δειγμάτων στον μετρικό χώρο του συνόλου $\mathbb{S}_n$ με την απόσταση Kendall Tau ως την l_1-νόρμα. Έχει αποδειχθεί ότι το συγκεκριμένο πρόβλημα είναι NP-Hard. Επιπλέον, ο κανόνας του Kemeny είναι ισοδύναμος με την εύρεση μιας εκτίμησης μέγιστης πιθανότητας υποθέτοντας ότι οι παρατηρήσεις μας δημιουργήθηκαν από ένα μοντέλο Mallows. Το $d_{KT}(\sigma, \pi)$ είναι μια μέτρηση απόστασης στο $\mathbb{S}_n$, το σύνολο των μεταθέσεων n στοιχείων, και ισούται με τον αριθμό των διμελών συγκρίσεων που είναι ασύμφωνες μεταξύ του σ και του π. Το μοντέλο Mallows είναι μια κατανομή κατάταξης στο $\mathbb{S}_n$ παραμετροποιημένη από μια κεντρική μετάθεση π^* που αποδίδει σε κάθε μετάθεση σ μια πιθανότητα εκθετική ως προς το $-d_{KT}(\sigma, \pi^*)$ (φθίνουσα).

Ένα άλλο σημαντικό ζήτημα στη συνάθροιση δειγμάτων κατάταξης είναι η μη πληρότητα των ατομικών προτιμήσεων. Για παράδειγμα ας θεωρήσουμε ένα σύνολο ταινιών και μια ομάδα ατόμων που τις κατατάσσουν σε μια διαδικτυακή πλατφόρμα. Κάθε χρήστης πρέπει να δώσει μια σειρά προτίμησης των ταινιών σύμφωνα με το προσωπικό του γούστο. Ωστόσο, ορισμένοι χρήστες μπορεί να μην έχουν ξεκάθαρη γνώμη για ορισμένες ταινίες ή να μην τις έχουν δει καθόλου, με αποτέλεσμα να μην μπορούν να συμπεριλάβουν αυτές τις ταινίες στη λίστα προτιμήσεών τους. Αυτό έχει ως αποτέλεσμα ημιτελείς λίστες προτιμήσεων των ατόμων. Επιπλέον, γίνεται όλο και πιο δύσκολο για τους χρήστες να κατασκευάσουν μια ενιαία κατάταξη των ταινιών καθώς ο αριθμός των ταινιών αυξάνεται. Αντίθετα, θα προτιμούσαν να σπάσουν τις αποφάσεις τους σε μικρότερες συγκρίσεις (κατά ζεύγη, τριμελείς συγκρίσεις, κ.λπ.). Και πάλι, μπορεί να είναι αδύνατο για τους χρήστες να αποφασίσουν για ορισμένες από τις συγκρίσεις των ταινιών. Τέλος, δεν μπορούμε να απαιτήσουμε από τους χρήστες να δώσουν μια κατάταξη κάθε ταινίας που γνωρίζουν, καθώς αυτό θα ήταν πολύ κουραστικό για αυτούς. Αυτοί οι περιορισμοί υπογραμμίζουν την ανάγκη θεώρησης του λεγόμενου selective μοντέλου, όπου κάθε δείγμα είναι μια μετάθεση κάποιου τυχαία επιλεγμένου υποσυνόλου του πλήρους συνόλου στοιχείων ή ένα σύνολο συγκρίσεων ανά ζεύγη μεταξύ στοιχείων του πλήρους συνόλου.

Στη συνέχεια, θα εξηγήσουμε τη σημασία της υπόθεσης μίγματος κατανομών κατάταξης και όχι μεμονωμένων μοντέλων για ολόκληρο τον πληθυσμό. Οι πληθυσμοί μπορεί να είναι ετερογενείς, πράγμα που σημαίνει ότι πρέπει να εκτιμηθούν περισσότερες από μία συλλογικές προτιμήσεις, μία για κάθε ομάδα. Επιστρέφοντας στο παράδειγμα των ταινιών, οι γυναίκες μπορεί να έχουν παρόμοιες προτιμήσεις ταινιών μεταξύ τους, όπως και οι άνδρες, αλλά οι προτιμήσεις των ανδρών μπορεί να είναι σημαντικά διαφορετικές από τις προτιμήσεις των γυναικών. Σε αυτή την περίπτωση, η εκτίμηση μιας κοινής συλλογικής προτίμησης για ολόκληρο τον πληθυσμό με κάποια μέθοδο όπως ο κανόνας του Kemeny θα αποτύχει να εκφράσει το ground truth του πληθυσμού και σε πιο τεχνικό επίπεδο η εκτιμώμενη κατανομή που θα μοντελοποιήσει τη συμπεριφορά του πληθυσμού θα είναι υπερβολικά απλοϊκή και έτσι θα αποτυγχάνει να κάνει fit στα δείγματα με επαρκή ακρίβεια. Η ιδέα των μιγμάτων έχει χρησιμοποιηθεί ευρέως και σε άλλα είδη δεδομένων, για παράδειγμα στα μείγματα Γκαουσιανών για διανύσματα χαρακτηριστικών. Ένας αριθμός βασικών μοντέλων

υπερτίθεται για την κατασκευή μιας πιο σύνθετης κατανομής, με περισσότερες ελεύθερες παραμέτρους (και επομένως μεγαλύτερη εκφραστικότητα) που θα κάνουν fit στα δεδομένα του δείγματος. Θεωρώντας ένα μείγμα αντί για ένα μεμονωμένο μοντέλο αυξάνει δραστικά τη δυσκολία του προβλήματός μας, επειδή για κάθε δείγμα πρέπει να μαντέψουμε την «ετικέτα» του, δηλαδή τη συνιστώσα του μείγματος από την οποία προήλθε, για να το αντιστοιχίσουμε στη σωστή ομάδα παρόμοιων δείγματα. Όσο πιο κοντά είναι τα κέντρα και όσο μεγαλύτερη διακύμανση έχουν τα δείγματα γύρω από αυτά τα κέντρα, τόσο πιο δύσκολο γίνεται η ταξινόμηση των δειγμάτων σε ομάδες. Το γεγονός ότι τα δείγματα είναι ελλιπή παίζει επίσης σημαντικό ρόλο στη δυνατότητα διαχωρισμού καθιστώντας ακόμη και αδύνατη την αναγνώριση των κρυφών κέντρων εάν τα δείγματα είναι πολύ μικρά, δηλαδή λύσεις με διαφορετικές παραμέτρους θα ήταν ισοδύναμες ως προς το ιστόγραμμά τους.

Το μοντέλο Mallows και οι γενικεύσεις του

Το μοντέλο Mallows μοιάζει με την κανονική κατανομή αλλά αντί για διανύσματα ορίζεται σε στοιχεία του $\mathbb{S}_n$, του συνόλου δηλαδή των μεταθέσεων n αντικειμένων. Όπως η κανονική κατανομή, το μοντέλο Mallows περιγράφεται από μια κεντρική παράμετρο και μια παράμετρο διακύμανσης (spread). Η πιθανότητα που αποδίδεται σε κάθε στοιχείο του συνόλου support είναι αντιστρόφως ανάλογη με μια εκθετική απόσταση μεταξύ του στοιχείου και της κεντρικής παραμέτρου και η βάση του εκθετικού εξαρτάται από την παράμετρο spread. Πιο συγκεκριμένα, εάν μια τυχαία μετάθεση $\pi \in \mathbb{S}_n$ ακολουθεί την κατανομή Mallows $\mathcal{M}(\pi_0, \phi)$, τότε $\mathbb{P}[\pi = \sigma] = \frac{\phi^{d(\pi_0,\sigma)}}{Z(\phi,n)}$.

- Το $\pi_0 \in \mathbb{S}_n$ είναι η κεντρική μετάθεση του μοντέλου. Εκφράζει το λανθάνον ground truth σχετικά με τις προτιμήσεις του πληθυσμού και είναι η πιο πιθανή μετάθεση στο σύνολο support.

- Το $\phi \in (0, 1)$ είναι η παράμετρος spread. Όσο υψηλότερη είναι η τιμή του, τόσο πιο διασκορπισμένα είναι τα δείγματα γύρω από την κεντρική μετάθεση. Στην ακραία περίπτωση όπου το ϕ πλησιάζει το μηδέν, το μόνο δείγμα με μη μηδενική πιθανότητα εμφάνισης είναι η κεντρική μετάθεση, οπότε σχηματίζεται μια σταθερή κατανομή. Στην αντίθετη ακραία περίπτωση, όπου το ϕ πλησιάζει το ένα, όλες οι μεταθέσεις στο $\mathbb{S}_n$ έχουν την ίδια πιθανότητα να εμφανιστούν, έτσι το μοντέλο Mallows εκφυλίζεται σε ομοιόμορφη κατανομή σε $\mathbb{S}_n$.

- Το $d : \mathbb{S}_n \times \mathbb{S}_n \to \mathbb{R}$ είναι κάποια μετρική απόστασης, για παράδειγμα η απόσταση KT, ο κανόνας του Spearman ή η απόσταση Hamming. Σε αυτή την εργασία εστιάζουμε αποκλειστικά στην απόσταση KT.

- Το $Z(\phi, n)$ είναι η σταθερά κανονικοποίησης, η οποία κάνει τη συνάρτηση πυκνότητας να αθροίζεται σε 1 έτσι ώστε να εκφράζει πιθανότητα. Στην περίπτωσή μας, όπου d είναι η απόσταση KT, έχουμε $Z(\phi, n) = \prod_{i=1}^{n} Z_i(\phi) = \prod_{i=1}^{n} \left(\sum_{j=0}^{i-1} \phi^i \right) = \frac{1}{(1-\phi)^{n-1}} \prod_{i=2}^{n} (1 - \phi^i)$.

Το μείγμα μοντέλων Mallows $\mathcal{M}$ παραμετροποιείται από το σύνολο των κεντρικών μεταθέσεων π_i, τα βάρη w_i και τις παραμέτρους διακύμανσης ϕ_i που αντιστοιχούν στα κέντρα π_i. Η

συνάρτηση μάζας πιθανότητας του μείγματος Mallows είναι η εξής:

$$M(\pi = \sigma) = \sum_{i=1}^{k} w_i \cdot \frac{\phi_i^{d_{KT}(\pi_i,\sigma)}}{Z(\phi_i, n)}$$

Κάθε κεντρική μετάθεση είναι μια μετάθεση n στοιχείων $\pi_i \in \mathbb{S}_n$ και υποθέτουμε ότι τα κέντρα διαφέρουν μεταξύ τους ανά δύο ($\pi_i \neq \pi_j$ για $i \neq j$). Τα βάρη w_i δεν είναι αρνητικά και αθροίζονται στο ένα ($\sum_{i=1}^{k} w_i = 1$). Η διαδικασία δειγματοληψίας έχει δύο στάδια. Αρχικά, επιλέγεται ένα κέντρο $i \in [n]$ με πιθανότητα w_i. Στη συνέχεια, γίνεται δειγματοληψία μιας μετάθεσης από το μεμονωμένο μοντέλο Mallows $\mathcal{M}(\pi_i, \phi_i)$. Στο πλαίσιο αυτής της εργασίας, συχνά θεωρούμε ότι όλες οι παράμετροι εξάπλωσης είναι ίσες ($\phi_i = \phi \forall i \in [n]$).

Μια γενίκευση του Mallows μοντέλου είναι το selective μοντέλο Mallows. Η συνάρτηση μάζας πιθανότητας αυτού του μοντέλου είναι η εξής: $\mathbb{P}[\pi = \sigma] = f(s) \cdot \frac{\phi^{d_{KT}(\pi_0,\sigma)}}{Z(\phi,|s|)}$, όπου s είναι το σύνολο των στοιχείων που βρίσκονται στο σ. Κάθε παρατήρηση π είναι μια μετάθεση των στοιχείων που εμφανίζονται στο αντίστοιχο σύνολο επιλογής s. Το f(s) είναι ο μηχανισμός επιλογής, μια συνάρτηση πιθανότητας που αποδίδει μια πιθανότητα επιλογής σε κάθε υποσύνολο s του πλήρους συνόλου των στοιχείων $[n]$. Το π_0 είναι η κεντρική μετάθεση του μοντέλου και είναι πλήρες (περιέχει όλα τα στοιχεία στο [n]). Το $d_{KT}(\pi_0, \pi)$ είναι η απόσταση Tau του Kendall μεταξύ της κεντρικής μετάθεσης και του δείγματος. Χρειάζεται να επαναπροσδιοριστεί γιατί το δείγμα π είναι πιθανόν ελλιπές. Μια φυσική γενίκευση του κλασικού ορισμού είναι η εξής:

$$d_{KT}(\pi_0, \pi) = \sum_{a,b \in s \wedge a<b} \mathbb{1}\{(\pi_0(a) - \pi_0(b)) \cdot (\pi(a) - \pi(b)) < 0\}$$

Αυτό που διαφέρει είναι ότι το άθροισμα μετράει τα ασύμφωνα ζεύγη (a,b) όπου $a, b \in s$, αντί για $a, b \in [n]$, όπου s είναι το σύνολο επιλογής.

Ορισμός 1. *Ένας μηχανισμός επιλογής f(s) λέγεται ότι είναι p-frequent ως προς τις l-μελείς συγκρίσεις, εάν για όλα τα σύνολα $x \subseteq \{1, ..n\}$ με μήκος μικρότερο ή ίσο του l $\mathbb{P}\{x \subseteq s\} \geq p \Leftrightarrow \forall x \sum_{x \in s} f(s) \geq p$.*

Το selective μείγμα Mallows συνδυάζει τις ιδιότητες του selective μοντέλου Mallows και του Mixture Mallows μοντέλου. Είναι ένα μοντέλο μείγματος, επειδή υποτίθεται ότι μια συλλογή από διακριτά κέντρα $\{\pi_1, ...\pi_k\}$ και όχι μια μεμονωμένη κεντρική κατάταξη. Είναι επίσης selective επειδή τα δείγματα που παράγονται από αυτό το μοντέλο δεν περιέχουν όλες τις πιθανές εναλλακτικές αλλά ένα τυχαίο υποσύνολο J αυτών, το οποίο δίνεται από έναν μηχανισμό επιλογής $f(J)$ για κάθε δείγμα. Η συνάρτηση μάζας πιθανότητας του μοντέλου είναι η εξής:

$$M(\pi = \sigma) = f(J) \cdot \sum_{i=1}^{k} w_i \cdot \frac{\phi^{d_{KT}(\pi_i||_J,\sigma)}}{Z(\phi, |J|)}$$

Η διαδικασία παραγωγής δείγματος αποτελείται από τρία στάδια. Στο πρώτο βήμα, ο μηχανισμός επιλογής $f(J)$ επιλέγει ένα τυχαίο υποσύνολο J στοιχείων του $[n]$ με πιθανότητα $f(J)$. Έπειτα, μια από τις k συνιστώσες του μείγματος ενεργοποιείται με πιθανότητα που

δίνεται από τα βάρη ανάμειξης. Η συνιστώσα i έχει πιθανότητα w_i να ενεργοποιηθεί κάθε φορά που παράγεται δείγμα. Τέλος, μια τυχαία μετάθεση π των στοιχείων στο J λαμβάνεται από το μοντέλο Mallows $M_i(\pi) = \frac{\phi^{d_{KT}(\pi_i||_J,\pi)}}{Z(\phi,|J|)}$, όπου i είναι ο δείκτης της ενεργοποιημένης συνιστώσας. Σημειώνουμε ότι το κέντρο π_i περιορίζεται στο J $(\pi_i||_J)$ και η συνάρτηση απόστασης KT μετράει ασύμφωνα ζεύγη μόνο σε στοιχεία που εμφανίζονται στο J.

Εκμάθηση Κατανομής και Εκτίμηση Παραμέτρων

Μια κλάση κατανομών C χαρακτηρίζεται efficiently learnable εάν για κάθε $\epsilon > 0$ και $0 < \delta \leq 1$ έχοντας πρόσβαση σε ένα μαντείο $GEN(D)$ που επιστρέφει δείγματα από μια άγνωστη κατανομή $D \in C$, υπάρχει ένας πολυωνυμικός αλγόριθμος A, που ονομάζεται αλγόριθμος εκμάθησης της C, παίρνει ως είσοδο τα δείγματα και δίνει μια εκτίμηση D' της D έτσι ώστε $Pr[d(D, D') \leq \epsilon]$ $geq1 - \delta$, όπου d είναι κάποια μετρική απόστασης μεταξύ των κατανομών D και D', όπως για παράδειγμα η απόσταση TV ή η απόκλιση KL, που θα συζητήσουμε αργότερα σε αυτό το κεφάλαιο. Στην εργασία αυτή χρησιμοποιούμε την απόσταση TV, που στην περίπτωση δύο διακριτών κατανομών P και Q πάνω σε ένα δειγματικό χώρο Ω γράφεται ως $d_{TV}(P, Q) = \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)|$. Σε ορισμένες περιπτώσεις, κάθε κατανομή $D \in C$ προσδιορίζεται μοναδικά από ένα σύνολο παραμέτρων. Για παράδειγμα, η κλάση μονοδιάστατων Γκαουσιανών κατανομών $N(\mu, \sigma^2)$ παραμετροποιείται από το ζεύγος (μ, σ). Διαφορετικές τιμές του (μ, σ) δίνουν διαφορετικές κατανομές $D \in C$, που όλες μαζί καλύπτουν ολόκληρη την κλάση C. Σε αυτήν την περίπτωση, ο αλγόριθμος A θα πρέπει να μπορεί να εκτιμήσει τις παραμέτρους (μ, σ) και τον ονομάζαμε αλγόριθμο εκμάθησης παραμέτρων.

Ένα εξαιρετικά χρήσιμο εργαλείο για τη συγκεκριμένη εργασία και την εκμάθηση κατανομών γενικότερα είναι οι ανισότητες συγκέντρωσης κατανομών, που φράσουν την πιθανότητα μια τυχαία μεταβλητή να λάβει τιμές μακριά από τη μέση τιμή της.Το θεώρημα 2 του [1] παρέχει εκθετικά tail bounds για αθροίσματα ανεξάρτητων φραγμένων μεταβλητών.
Έστω ανεξάρτητες μεταβλητές $X_1, ...X_n$ και κάθε X_i φράσεται στο διάστημα $[a_i, b_i]$. Έστω $\bar{X}$ ο εμπειρικός μέσος όρος αυτών των μεταβλητών, $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$. Τότε, για $t > 0$ έχουμε:

$$\mathbb{P}\{\bar{X} - \mathbb{E}[\bar{X}] \geq t\} \leq exp\left(-\frac{2n^2t^2}{sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$\mathbb{P}\left\{\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq t\right\} \leq 2exp\left(-frac2n^2t^2\sum_{i=1}^{n}(b_i - a_i)^2\right)$$

Σε αυτή την εργασία κάνουμε εκτενή χρήση των φραγμάτων Hoeffding για διωνυμικές κατανομές.
Έστω $X \sim Bin(n, p)$. Τότε έχουμε:

$$\mathbb{P}[X \leq k] \leq exp\left(-2n\left(p - \frac{k}{n}\right)^2\right)$$

Τα φραγματα Hoeffding αξιοποιούνται για τον υπολογισμό της δειγματικής πολυπλοκότητας εκτίμησης της κεντρικής κατάταξης του μοντέλου Mallows. Παραθέτουμε τα αποτέσματα των Caragiannis et al. στο [2]. Έστω ότι μας δίνεται ένα σύνολο N δειγμάτων $\sigma_1, ...\sigma_N$, που προέρχονται από ένα μοντέλο Mallows. Τα δείγματα είναι πιθανώς ελλιπή. Θέλουμε να χρησιμοποιήσουμε αυτά τα δείγματα για να εκτιμήσουμε την κρυφή κεντρική κατάταξη με μεγάλη πιθανότητα. Για το σκοπό αυτό θα χρησιμοποιήσουμε έναν εκτιμητή που αποφασίζει πλειοψηφικά για κάθε διμελή σύγκριση. Ο Εκτιμητής Θέσης $\hat{\pi}$ υπολογίζει τη θέση κάθε αντικειμένου στην κρυφή κατάταξη ως εξής:

$$\hat{\pi}[i] = 1 + \sum_{j \in [n] \setminus \{i\}} \mathbb{1}\left\{ \sum_{k=1}^{N} \mathbb{1}\{j > i \text{ in } \sigma_k\} > \sum_{k=1}^{N} \mathbb{1}\{i > j \text{ in } \sigma_k\} \right\}, \ \forall i \in [n]$$

Αν προκύψουν ισοπαλίες τις σπάμε ομοιόμορφα από αριστερά προς τα δεξιά. Εάν το N είναι αρκετά μεγάλο, τότε ο εκτιμητής θέσης ανακτά τη σωστή κεντρική κατάταξη π_0 με μεγάλη πιθανότητα, όπως θα δούμε στο επόμενο θεώρημα.

Θεώρημα 1. *Έστω $\mathcal{M}(\pi_0, \phi)$ μια κατανομή Mallows με κεντρική κατάταξη $\pi_0 \in \mathbb{S}_n$ και παράμετρο spread $\phi \in (0, 1)$. Για οποιοδήποτε $\epsilon > 0$, δεδομένου ενός δειγματικού προφίλ που προέρχεται από την από κατανομή $\mathcal{M}(\pi_0, \phi)^N$ για οποιοδήποτε N τουλάχιστον ίσο με κάποια τιμή $O\left(\frac{\log(n/\epsilon)}{(1-\phi)^2}\right)$, ο εκτιμητής θέσης ανακτά την κεντρική κατάταξη π_0 με πιθανότητα τουλάχιστον $1 - \epsilon$.*

Το παραπάνω άνω φράγμα για τη δειγματική πολυπλοκότητα είναι tight, όπως μας δείχνει το ακόλουθο θεώρημα.

Θεώρημα 2. *Για κάθε $\epsilon \in (0, 1/2]$ και οποιουδήποτε εκτιμητή κεντρικής κατάταξης, υπάρχει μια κεντρική κατάταξη $\pi_0 \in \mathbb{S}_n$ έτσι ώστε, για κάθε $\phi \in (0, 1)$, ο εκτιμητής, δεδομένου ενός δειγματικού προφίλ που προέρχεται από την από κατανομή $\mathcal{M}(\pi_0, \phi)^N$, ανακτά το π_0 με πιθανότητα τουλάχιστον $1 - \epsilon$, μόνο εάν $N = \Omega\left(\frac{\log(n/\epsilon)}{\log(1/\phi)}\right)$.*

Οι Busa Fekete et al. στο [3] δίνουν έναν αλγόριθμο για την εκτίμηση της παραμέτρου ϕ. Αρχικά με τον εκτιμητή θέσης ανακτούμε την κεντρική διάταξη, όπως δείξαμε παραπάνω. Στη συνέχεια καθίσταται εφικτή η ανάκτηση του ϕ με αυθαίρετα μικρό απόλυτο σφάλμα.

Θεώρημα 3. *Αν η κεντρική κατάταξη π_0 είναι γνωστή, τότε με $N = \Omega\left(\frac{\log(1/\delta)}{n\epsilon^2}\right)$ δείγματα μπορούμε να βρούμε σε πολυωνυμικό χρόνο μια εκτίμηση $\hat{\phi}$ του άγνωστου ϕ^* τέτοια ώστε:*

$$\mathbb{P}_{\Pi \sim \mathcal{M}^N_{\phi^*, \pi_0}} \left[\left| \hat{\phi}(\Pi) - \phi^* \right| \leq \epsilon \right] \geq 1 - \delta$$

Μάθηση του μείγματος Mallows

Στο paper του Zagier [4] εμφανίζεται ένα αποτέλεσμα καθοριστικής σημασίας για τη μελέτη της μάθησης μειγμάτων Mallows. Θεωρούμε τον $n! \times n!$ πίνακα $A_n(\phi)$, του οποίου οι γραμμές και στήλες δεικτοδοτούνται απο τις διάφορες μεταθέσεις π, σ της συλλογής n αντικειμένων $[n]$ και του οποίου τα στοιχεία $A_{\pi\sigma}$ είναι ίσα με $\phi^{d_{KT}(\pi,\sigma)}$. Μπορούμε εύκολα να δούμε ότι

κάθε γραμμή του πίνακα αυτού αντιστοιχεί στο vectorization ενός Mallows μοντέλου. Κάθε κυρτός συνδυασμός k γραμμών αντιστοιχεί σε ένα μείγμα Mallows. Ο Zagier υπολόγισε την ορίζουσα αυτού του πίνακα και κατέληξε σε ένα κλειστό τύπο που παίρνει πάντοτε μη μηδενική τιμή. Το αποτέλεσμα αυτό είναι σημαντικό για το πρόβλημα μας γιατί δείχνει ότι οποιεσδήποτε k γραμμές είναι γραμμικά ανεξάρτητες μεταξύ τους και άρα αν δύο γραμμικοί συνδυασμοί k στηλών (δηλαδή δύο μείγματα Mallows) είναι ίσα σαν διανύσματα (δηλαδή αν τα δυο μείγματα έχουν ίδια ιστογράμματα) τότε οι δύο συνδυασμοί αποτελούνται από τις ίδιες γραμμές του πίνακα $A_n(\phi)$ με τα ίδια αντίστοιχα βάρη (δηλαδή τα δύο μείγματα έχουν τα ίδια κέντρα και τα ίδια αντίστοιχα βάρη πρόσμιξης). Αυτή η ιδιότητα εξασφαλίζει το identifiability του μείγματος Mallows πάνω σε πλήρη δείγματα, δηλαδή τη δυνατότητα να συμπεράνουμε με μοναδικό τρόπο τις παραμέτρους του αν γνωρίζουμε τη μάζα πιθανότητας σε κάθε σημείο. Επίσης, η τιμή της ορίζουσας μπορεί να χρησιμοποιηθεί για να δειχθεί ότι το μέτρο της προβολής μιας γραμμής πάνω στο ορθογώνιο συμπλήρωμα άλλων k-1 γραμμών δε μπορεί να είναι πολύ μικρό. Με βάση αυτό το κάτω φράγμα αποδεικνύονται κάτω φράγματα για την TV απόσταση μεταξύ δύο μειγμάτων με διαφορετικές παραμέτρους. Αυτό αποτελεί μια εύρωστη και ποσοτική διατύπωση του identifiability που μπορεί να χρησιμοποιηθεί στο learning.

Οι Liu και Moitra στο [5] χρησιμοποιώντας τις παραπάνω ιδέες δίνουν ένα πολυωνυμικό αλγόριθμο για τη μάθηση μειγμάτων Mallows, κάνοντας μόνο αναγκαίες υποθέσεις για το μείγμα (οι συνιστώσες να διαφέρουν ανά δύο μεταξύ τους και κάθε μια να διαφέρει από την ομοιόμορφη κατανομή)

Θεώρημα 4. *Έστω ότι το μείγμα είναι μ-μη εκφυλισμένο, δηλαδή* $\forall i,j \in [k]\ i \neq j \Rightarrow d_{TV}(M_i,M_j) > \mu$ *και* $\forall i \in [k]\ d_{TV}(M_i, Uniform) > \mu$. *Τότε υπάρχει αλγόριθμος με χρονική και δειγματική πολυπλοκότητα* $poly_k(n, \frac{1}{\mu}, \frac{1}{w_{min}}, \frac{1}{,\partial}, log(\frac{1}{\delta}))$ *που μαθαίνει τα κέντρα του μείγματος επακριβώς και τις παραμέτρους* ϕ_i, w_i *με απόλυτο σφάλμα το πολύ* ∂.

Οι Mao et al. στο [6] βελτιώνουν την εξάρτηση από τον αριθμό n των αντικειμένων κάνοντάς τη λογαριθμική και γεφυρώνοντας το κενό που υπήρχε μεταξύ της περίπτωσης του ενός κέντρου και της περίπτωσης μείγματος. Επίσης δίνουν ένα αλγόριθμο μάθησης που χρησιμοποιεί queries ελάχιστου μήκους. Ένας περιορισμός βέβαια της συγκεκριμένης δουλειάς είναι ότι υποθέτει πως όλες οι παράμετροι spread είναι ίσες και γνωστές εκ των προτέρων.

Ας δούμε κάποιες βασικές έννοιες που χρησιμοποιούνται στο paper. Καταρχάς, γίνεται χρήση της μεθόδου των ροπών. Στα μείγματα Mallows ένας φυσικός τρόπος να οριστούν οι ροπές τάξης l είναι οι ομάδες l διμελών συγκρίσεων ή παρόμοια οι l-μελείς συγκρίσεις. Θα εστιάσουμε στην περίπτωση των δεύτερων, οι οποίες υπερκαλύπτουν την πρώτη περίπτωση από άποψη προσεφερόμενης πληροφορίας. Οι συγκρίσεις αυτές μας δίνουν τις περιθώριες κατανομές του μοντέλου πάνω σε υποσύνολα των αντικειμένων.

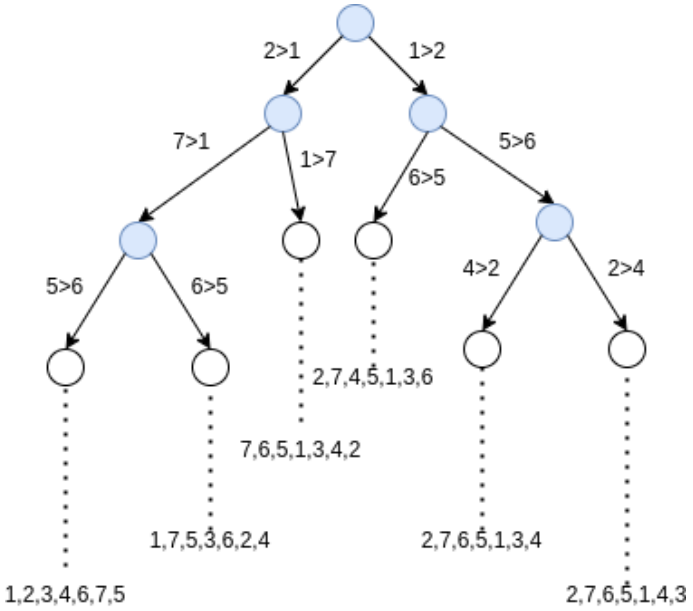
Για μια μετάθεση π n αντικειμένων θεωρούμε δύο είδη περιορισμού της πάνω σε ένα υποσύνολο J των n αντικειμένων. Πρώτον τον injective (1-1 αλλά όχι επί) που τον συμβολίζουμε με $\pi|J$ και δεύτερον τον bijective (1-1 και επί) που τον συμβολίζουμε με $\pi||J$. Στον injective

διατηρείται η πληροφορία για τη θέση των αντικειμένων στο πλήρες ranking π, ενώ στον bijective μόνο η κατάταξη μεταξύ των επιλεγμένων αντικειμένων. Για παράδειγμα, για το ranking $\pi = (3, 2, 4, 6, 1, 5)$ και το σύνολο επιλογής $J = \{1, 4, 5\}$ έχουμε για το injection $\pi|J(1) = 5$, $\pi|J(4) = 3$ και $\pi|J(5) = 6$ και για το bijection $\pi||J = (4, 1, 5)$.

Μια άλλη σημαντική έννοια είναι τα μαντεία (oracles). Αυτά προσφέρουν πληροφορίες χωρίς θόρυβο για το μοντέλο, σε αντίθεση με τα δείγματα που είναι θορυβώδη. Απώτερος στόχος είναι να χρησιμοποιηθούν τα δείγματα από κάποιο αλγόριθμο που με μεγάλη πιθανότητα κάνει simulate τα μαντεία, χρησιμοποποιώντας πολυωνυμικό αριθμό δειγμάτων.

Ορισμός 2. *Έστω ένα μείγμα* M *με κέντρα* $\{c_1, c_2, ...c_k\}$. *Το "ασθενές" μαντείο με είσοδο ένα ερώτημα για κάποιο σύνολο J επιστρέφει το σύνολο των περιορισμών των κέντρων του μείγματος πάνω στο J:* $\{c_1, c_2, ...c_k\}$. *Το σύνολο αυτό περιέχει μόνο διαφορετικά μεταξύ τους στοιχεία, οπότε ενδέχεται ο πληθάριθμός του να είναι μικρότερος του* k. *Το "ισχυρό" μαντείο επιστρέφει την κατανομή του* $\pi|J$, *όπου* π *τυχαίο ranking που ακολουθεί την κατανομή* M.

Μια τελευταία έννοια που θα χρειαστεί να δούμε πριν προχωρήσουμε στους αλγορίθμους μάθησης είναι η "υπογραφή" (signature) των κέντρων του μείγματος. Signature ονομάζουμε μια ομάδα διμελών συγκρίσεων που απομονώνει ένα κέντρο. Μπορούμε να βρούμε ένα μοναδικό signature set που ξεχωρίζει ταυτόχρονα όλα τα κέντρα και χρησιμοποιεί $k - 1$ συγκρίσεις. Επίσης υπάρχει πάντα τουλάχιστον ένα κέντρο που μπορεί να απομονωθεί χρησιμοποιώντας $O(log(k))$ συγκρίσεις. Ωστόσο κάποια κέντρα μπορεί να χρειάζονται $O(k)$ συγκρίσεις για να απομονωθούν. Τα signatures μπορούμε να τα δούμε σαν decision trees που στα φύλλα τους έχουν τα κέντρα που απομονώνονται.



Στη συνέχεια παρουσιάζουμε έναν αλγόριθμο μάθησης των κέντρων και των αντίστοιχων βαρών του μείγματος που χρησιμοποιεί κλήσεις στο ισχυρό μαντείο και χρησιμοποιεί τα signatures που είδαμε παραπάνω. Ο αλγόριθμος που παρουσιάζεται εφαρμόζεται σε μείγμα με κοινά spread parameters, ωστόσο εύκολα γενικεύεται στην περίπτωση των διαφορετικών

spread parameters, αρκεί να έχουμε κι εκεί διαθέσιμο μαντείο. Επίσης, υπάρχει μια παρόμοια παραλλαγή του αλγορίθμου που χρησιμοποιεί κλήσεις στο ασθενές μαντείο και μαθαίνει μόνο τα κέντρα του μείγματος. Ο αλγόριθμος που χρησιμοποιεί το ισχυρό μαντείο κάνει queries ελάχιστου μήκους ($O(log(k))$, ενώ ο αλγόριθμος που χρησιμοποιεί το ασθενές μαντείο κάνει queries μήκους $O(k)$.

Ο αλγόριθμος μαθαίνει το μείγμα επαγωγικά ως προς το πλήθος των αντικειμένων. Για n=2 αντικείμενα λαμβάνονται άμεσα οι παράμετροι της περιθώριας κατανομής πάνω στα αντικείμενα αυτά με μια κλήση στο ισχυρό μαντείο.

for n in $[3, n_{max}]$:

∗ C := το σύνολο των διαφορετικών στοιχείων του (πολυ)συνόλου $\{\pi_1|_{[n-1]}, \ldots, \pi_k|_{[n-1]}\}$.

∗ l=0

do{

∗ Υπάρχει $\pi_{s^*}|_{[n-1]} = [e_1, e_2, \ldots e_{n-1}] \in C$ που μπορεί να απομονωθεί από ένα signature *sig* μήκους το πολύ $l \leq \lfloor log_2(k) \rfloor$

for r in $[2, n-1]$:

∗ J := το σύνολο των αντικειμένων του *sig* συνένωση με το $\{e_{r-1}, e_r, n\}$

∗ Παίρνουμε από το ισχυρό μαντείο τις παραμέτρους της κατανομής $M_J(\pi) = \sum_{j=1}^{k'} w'_j \cdot \frac{\phi^{d_{KT}(\pi'_j, \pi)}}{Z(\phi, |J|)}$

∗ Ένα από τα διαφορετικά κέντρα π'_{j^*} ισούται με $\pi_{s^*|J}$.

∗ Αν το π'_{j^*} περιέχει τη διατεταγμένη τριπλέτα (e_{r-1}, n, e_r), έχουμε μάθει ένα νέο κέντρο c_l στο $\mathbb{S}_n$, που ισούται με $[e_1, e_2, \ldots, e_{r-1}, n, e_r, \ldots, e_{n-1}]$, και το αντίστοιχο βάρος του wt_l=w'_{j^*} - $\sum_{m:c_m|J=c_l|J} wt_m$.

∗ Πράττουμε όμοια για κέντρα που αρχίζουμε με ή τελειώνουν σε n

∗ Διγράφουμε το $\pi_{s^*}|_{[n-1]}$ από το C.

∗ l+=1

}**while** Το C δεν είναι κενό

Τώρα μένει να υλοποιήσουμε κάπως το μαντείο. Αυτό γίνεται μέσω της μεθόδου SubOrder, η οποία κάνει εξαντλητική αναζήτηση πάνω στο χώρο των υποψήφιων περιθώριων κατανομών. Κάθε υποψήφιο μοντέλο συγκρίνεται με το το εμπειρικό μοντέλο που προκύπτει από τα διαθέσιμα δείγματα και αν η απόσταση είναι αρκετά μικρή επιστρέφονται οι παράμετροι του υποψήφιου μοντέλου. Ο χώρος των υποψήφιων μοντέλων είναι πολυωνυμικά (και όχι εκθετικά) μεγάλος αν θεωρήσουμε το k σταθερά, καθώς τα υποψήφια μοντέλα ορίζονται πάνω σε ένα υποσύνολο μεγέθους $O(k)$ του πλήρους συνόλου των n αντικειμένων. Παρακάτω παρουσιάζεται η SubOrder που κάνει simulate το ασθενές μαντείο, ωστόσο μπορεί να φτιαχτεί όμοια και ρουτίνα που κάνει simulate το ισχυρό μαντείο, επιστρέφοντας ρητές προσεγγίσεις των βαρών με κάποιο βαθμό ακρίβειας, μιας και τα βάρη είναι συνεχείς μεταβλητές που δε μπορούμε να τις εκτιμήσουμε επακριβώς. Έτσι επιτυγχάνεται η μάθηση, καθώς υπάρχουν guarantees ότι με μεγάλη πιθανότητα η SubOrder κάνει σωστά simulate το μαντείο.

Συνάρτηση SubOrder

Είσοδος: Οι παρατηρήσεις $\sigma_1, \ldots, \sigma_N \in \mathcal{S}_n$, ένα υποσύνολο $J \subset [n]$, $\ell := |J|$ και $L = \lceil 3k/\eta \rceil$ όπου $\eta = poly_{k,\ell,}(\phi, \gamma)$

$* \ \mathscr{M} := \left\{ \sum_{i=1}^{k} \frac{r_i}{L} M\left(\pi_{\rho_i}, \phi\right) : \rho_i \in \mathcal{S}_{n,J}, r_i \in [L], r_i \geq \gamma L, \sum_{i=1}^{k} r_i = L \right\}$

$*$ Για κάθε μείγμα Mallows $\mathcal{M}' \in \mathscr{M}$, πάραξε N' i.i.d.
τυχαία permutations $\sigma'_1, \ldots, \sigma'_{N'}$ από το $\mathcal{M}'$.
Υπολόγισε την εμπειρική περιθώρια κατανομή $\mathcal{M}'_{N'}\big|_J = \frac{1}{N'} \sum_{m=1}^{N'} \delta_{\sigma'_m|_J}$.

$*$ Αν για κάποιο $\mathcal{M}' = \sum_{i=1}^{k} \frac{r_i}{L} M\left(\pi_{\rho_i}, \phi\right) \in \mathscr{M}$ ισχύει ότι
$\mathrm{TV}\left(\mathcal{M}'_{N'}\big|_J, \mathcal{M}_N|_J\right) \leq \eta/2$,
επίστρεψε το σύνολο των σχετικών κατατάξεων $\left\{\pi_{\rho_i}||_J : i \in [k]\right\}$

.

Μάθηση Selective μειγμάτων Mallows

Παραπάνω είδαμε πώς γίνεται η μάθηση μειγμάτων Mallows από πλήρη δείγματα. Όταν τα δείγματα είναι ελλιπή το πρόβλημα δυσκολεύει και μάλιστα μπορεί να γίνει μη επιλύσιμο. Συγκεκριμένα μπορεί να παραβιαστεί το identifiability που ίσχυε στην περίπτωση των πλήρων δειγμάτων. Επιστρέφουμε στην ορίζουσα του Zagier με ένα παράδειγμα που αναδεικνύει το πρόβλημα αυτό.
Θεωρούμε την περίπτωση τριών αντικειμένων. Πρώτα υποθέτουμε ότι τα δείγματα είναι πλήρη. Τότε έχουμε:

$$A = \begin{bmatrix} 1 & \phi & \phi & \phi^2 & \phi^2 & \phi^3 \\ \phi & 1 & \phi^2 & \phi^3 & \phi & \phi^2 \\ \phi & \phi^2 & 1 & \phi & \phi^3 & \phi^2 \\ \phi^2 & \phi^3 & \phi & 1 & \phi^2 & \phi \\ \phi^2 & \phi & \phi^3 & \phi^2 & 1 & \phi \\ \phi^3 & \phi^2 & \phi^2 & \phi & \phi & 1 \end{bmatrix}$$

$$det(A) = -(\phi^2 - 1)^7(\phi^2 - \phi + 1)(\phi^2 + \phi + 1) \neq 0 \ \forall \phi \in (0, 1)$$

Στη συνέχεια υποθέτουμε ότι τα δείγματα είναι ελλιπή. Δεδομένου ότι έχουμε μόνο 3 αντικείμενα, τα ημιτελή δείγματα μπορούν να είναι μόνο συγκρίσεις ανά ζεύγη. Κατασκευάζουμε τον πίνακα A με γραμμές που αντιστοιχούν σε μεταθέσεις στο $\mathbb{S}_n$ και στήλες σε συγκρίσεις κατά ζεύγη στοιχείων του 1,2,3.

$$A = \begin{bmatrix} 1 & 1 & \phi & \phi & 1 & \phi \\ 1 & 1 & 1 & \phi & \phi & \phi \\ \phi & \phi & 1 & 1 & \phi & 1 \\ 1 & \phi & 1 & 1 & \phi & \phi \\ \phi & \phi & \phi & 1 & 1 & 1 \\ \phi & 1 & \phi & \phi & 1 & 1 \end{bmatrix}$$

$det(A) = 0\ \forall \phi \in (0,1)\ \Rightarrow$ το identifiability παραβιάζεται.

Το επόμενο θεώρημα συνοψίζει τα αποτελέσματά μας ως προς το identifiability στο σενάριο ελλιπών δειγμάτων. Κάνουμε την υπόθεση των Mao et al. ότι τα spread parameters είναι όλα ίσα μεταξύ τους.

Θεώρημα 5. *(Identifiability)*

- *Αν για όλα τα ελλιπή rankings π μήκους $2 \cdot \lfloor log_2(k) \rfloor + 3$ ισχύει ότι*

  1. *$f(J) \neq 0$, όπου J το σύνολο των αντικειμένων του π και*
  2. *$M_1(\pi) = M_2(\pi) \Leftrightarrow \sum_{i=1}^{k} w_{1,i} \cdot \frac{\phi^{d_{KT}(\pi_{1,i}|_J, \pi)}}{Z(\phi, |J|)} = \sum_{i=1}^{k} w_{2,i} \cdot \frac{\phi^{d_{KT}(\pi_{2,i}|_J, \pi)}}{Z(\phi, |J|)}$*

  τότε το $\{(w_{1,1}, \pi_{1,1}), (w_{1,2}, \pi_{1,2}), \ldots (w_{1,k}, \pi_{1,k})\}$ και το $\{(w_{2,1}, \pi_{2,1}), (w_{2,2}, \pi_{2,2}), \ldots (w_{2,k}, \pi_{2,k})\}$ είναι ίσα σαν σύνολα.

- *Αν $l < 2(\lfloor log_2(k) \rfloor + 1)$, τότε υπάρχουν δύο μείγματα M_1, M_2 με διαφορετικά σύνολα κεντρικών rankings και $M_1(\pi) = M_2(\pi), \forall \pi$ με μήκος μικρότερο ίσο του l.*

Το παραπάνω θεώρημα μας εγγυάται ότι μπορούμε να μάθουμε το μείγμα παρατηρώντας δείγματα λογαριθμικού μήκους ως προς k. Ιδανικά θα θέλαμε να αρκούσαν οι απλές διμελείς συγκρίσεις αλλά αυτό δεν είναι πάντα εφικτό. Στην απλή αλλά ενδιαφέρουσα περίπτωση των δύο κέντρων είναι η εφικτή η μάθηση από διμελείς συγκρίσεις.

Θεώρημα 6. *(Μάθηση Μείγματος Δύο Κέντρων Από Διμελείς Συγκρίσεις)*
Έστω μείγμα δύο κέντρων με κοινή παράμετρο ϕ. Υποθέτουμε ότι το μείγμα είναι $\alpha-$μη εκφυλισμένο, δηλαδή $|w_i - 0.5| > \alpha$, $w_i > \alpha$, για i=1,2 και $\phi < 1 - \alpha$. Τότε μπορούμε να μάθουμε τα δύο κεντρικά rankings επακριβώς με πιθανότητα τουλάχιστον $1-\epsilon$, χρησιμοποιώντας $O\left(\frac{n log(n) \cdot log(n/\epsilon)}{\alpha^4}\right)$ δείγματα διμελών συγκρίσεων.

Επιστρέφουμε στην περίπτωση των k κέντρων. Μια εναλλακτική μέθοδος του parameter cover που επιχειρεί η suborder μέθοδος που είδαμε παραπάνω είναι να δοκιμάζουμε εξαντλητικά διαφορετικούς τρόπους να συσταδοποιήσουμε τα δείγματα σε k ομάδες. Αν η κάθε ομάδα περιέχει τον αριθμό δειγμάτων που απαιτεί ο εκτιμητής θέσης των Caragiannis et al. τότε κάποια από τις υποψήφιες συσταδοποιήσεις θα δώσει το σωστό σύνολο κεντρικών rankings αν εφαρμόσουμε εσωτερικά σε κάθε συστάδα τον εκτιμητή θέσης. Η μέθοδος αυτή είναι μάλιστα φιλική ως προς τα ελλιπή δείγματα αφού ο positional estimator λειτουργεί και με είσοδο ελλιπή δείγματα. Το παρακάτω θεώρημα συνοψίζει το αποτέλεσμά μας.

Θεώρημα 7. *(Sample Grouping vs Parameter Cover)*
Έστω ένα μείγμα M *με κέντρα* $\{c_1, c_2, ...c_k\}$. *Δεδομένων* $N = O\left(\frac{r}{\gamma} + \frac{ln(k/\epsilon)}{\gamma^2}\right)$ *δειγμάτων του* M, *όπου* $r = O\left(\frac{log(k \cdot n/\epsilon)}{(1-\phi)^2}\right)$ *και* $\gamma = min\{w_i\}$, *μπορούμε να κατασκευάσουμε ένα σύνολο* $C \subseteq \mathbb{S}_n^k$ *σε χρόνο* $O\left(\frac{N^{k(r+1)}}{(r!)^k}\right)$, *το οποίο με πιθανότητα τουλάχιστον* $1 - \epsilon$ *περιέχει το σωστό συνδυασμό κέντρων* $\{c_1, c_2, ...c_k\}$.

Μια μέθοδος cover θα έκανε αντ'αυτού $(n!)^k$ ελέγχους υποψήφιων μοντέλων. Η εξάρτηση της δικής μας μεθόδου από το n είναι πολύ πιο ήπια, ενώ ο αριθμός των ελέγχων παραμετροποιείται και από άλλες παραμέτρους εκτός του n και του k. Αν για παράδειγμα το ϕ είναι μικρό ο χώρος αναζήτησης της μεθόδου μας συρρικνώνεται, ενώ η μέθοδος cover των Mao et al. δεν προσαρμόζει την πολυπλοκότητα της στο ϕ.

Τέλος, παρουσιάζουμε δύο αποτελέσματα μας που αφορούν διαχωρίσιμα μείγματα. Το πρώτο θεώρημα αξιοποιεί τη διαχωρισιμότητα που επιφέρει η μικρή διασπορά (μικρό spread parameter) ενώ το δεύτερο τη διαχωρισιμότητα με την έννοια ότι οι αποστάσεις μεταξύ διαφορετικών συνιστωσών είναι επαρκώς μικρές σε σχέση με τις "ακτίνες" των συνιστωσών του μείγματος.

Θεώρημα 8. *Έστω μείγμα με βάρη ανάμιξης* $w_1, ...w_k$ *και παράμετρο* ϕ *τέτοια ώστε* $min\{w_i\} - \phi = g > 0$, *με το* g *να θεωρείται γνωστό. Τότε εκτελώντας* $n^2 \cdot k$ *adaptive queries πάνω σε υποσύνολα μήκους* $O(k)$ *των αντικειμένων και δεδομένου ότι το σύνολο επιλογής του κάθε query εκπροσωπείται σε τουλάχιστον* N *δείγματα, όπου* $N = O(log(n \cdot k/\epsilon)/(g^2))$, *μαθαίνουμε τα κέντρα του μείγματος επακριβώς με πιθανότητα τουλάχιστον* $1 - \epsilon$.

Θεώρημα 9. *Έστω ένας μηχανισμός επιλογής που αφαιρεί* m *αντικείμενα με πιθανότητα* $p(m)$. *Υποθέτουμε ότι διαθέτουμε* $N = O\left(\frac{r}{\gamma} + \frac{L}{\gamma^2}\right)$ *δείγματα, όπου* $L = ln\left(\frac{k}{\epsilon}\right), \gamma = min_i\{w_i\}$ *και* $r = O\left(\frac{log(k \cdot n/\epsilon)}{(1-\phi)^2}\right)$. *Επίσης, θεωρούμε ένα πιθανοτικό άνω φράγμα* m_{cr} *για τον αριθμό των αφαιρούμενων αντικειμένων, για το οποίο* $\sum_{m=0}^{m_{cr}} p(m) > \epsilon/N$. *Αν η ελάχιστη KT απόσταση* a *μεταξύ δύο κέντρων του μείγματος ικανοποιεί τη συνθήκη* $a > (2n - m_{cr} + 1)m_{cr}/2 + 4d_{max}$, *όπου* $d_{max} = O\left([log(N) + nlog(n) - log(\epsilon)]\ /log(\frac{1}{\phi})\right)$, *τότε μπορούμε να μάθουμε τα κέντρα του μείγματος επακριβώς με πιθανότητα τουλάχιστον* $1 - \epsilon$.

Chapter 1

Introduction

Social choice theory is a framework for analysis of combining individual opinions, preferences, interests, or welfares to reach a collective decision or social welfare in some sense. Social choice theory dates from Marquis de Condorcet's formulation of the voting paradox (late 18th century). The Condorcet paradox is a situation in which societal preferences can be cyclic (conflicting) , even if individuals' preferences are acyclic (transitive). In an election with only two candidates, where each voter has a preference for one candidate over the other, the majority selection rule works fine, giving an order of the two candidates that agrees with the majority of the voters preferences and is self-consistent. However this is not always possible when the number of candidates exceeds two. An example of the paradox is the following:

Suppose we have candidates A,B and C and three voters. The following table presents the individual preferences of the voters.

Individual Preferences			
Voter	First preference	Second preference	Third preference
Voter 1	A	B	C
Voter 2	B	C	A
Voter 3	C	A	B

The majority of the voters prefer A to B, B to C and C to A. The resulting collective preference A>B>C>A is cyclic and thus inconsistent. This paradox indicates the need for more complex and robust voting mechanisms like score voting. Another interesting question arising is whether the voting mechanism is truthful, that is whether the voters have an incentive to give a vote that doesn't fully agree with their individual beliefs in order to promote a specific outcome of the election. However, this perspective of the voting problem is rather game theoretic and is out of the scope of this work.

Kemeny's rule is a more meaningful and effective way of aggregating ranking samples. Given a sample profile $\{\sigma_1, \sigma_2, \ldots \sigma_N\} \in \mathbb{S}_n^N$, Kemeny's rule chooses the following ranking τ as an estimation of the collective preference: $\tau = argmin_{\tau \in \mathbb{S}_n} \sum_{i=1}^{N} d_{KT}(\tau, \sigma_i)$. This calculation can be viewed as finding the median of the samples in the metric space of the set

$\mathbb{S}_n$ with the Kendall Tau distance as the l_1 norm and has been proved to be an NP-Hard problem. Moreover Kemeny's rule is equivalent to finding a maximum likelihood estimation assuming that our observations were generated by a Mallows Model. $d_{KT}(\sigma, \pi)$ is a distance metric in $\mathbb{S}_n$, the set of permutations of n items, and is equal to the number of pairwise comparisons that are discordant between σ and π. Mallows Model is a ranking distribution in $\mathbb{S}_n$ parametrised by a central permutation π^* that assigns to each permutation σ a probability exponentially proportional (decreasing) to $d_{KT}(\sigma, \pi^*)$.

Another important issue in ranking aggregation is the incompleteness of the individual preferences. Take for example a set of movies and a group of people ranking them on an online platform. Each user has to give a preference order of the movies according to their personal taste. However, some users might not have a clear opinion about some movies or they might have not seen them at all, making them unable to include these movies in their preference list. This results in incomplete preference lists of the individuals. Apart from that, it gets increasingly difficult for the users to construct a single ranking of the movies as the number of the movies increases. Instead they would rather break their decisions into smaller comparisons (pairwise, 3-wise, etc). Again, it might be impossible for the users to decide for some of the movie comparisons. Finally we simply can't demand the users to give a ranking of every single movie they know, as this would be too burdensome for them. These limitations underline the need for a so-called selective model where each sample is a permutation of some randomly selected subset of the full set of items or a set of pairwise comparisons between items of the full set.

Next, we are going to explain the importance of assuming mixtures of ranking models rather than single models for the whole population. Populations may be heterogeneous, which implies that more than one collective preferences should be estimated, one for each cluster. Women for example might have similar movie tastes with each other and men might have similar tastes as well but the taste of men might be significantly different than the taste of women. In this case, estimating a single collective preference for the whole population with some method like Kemeny's rule would fail to express the ground truth of the population and on a more technical level the estimated distribution that should model the behaviour of the population would be too simplistic and thus fail to fit the samples with adequate accuracy. The idea of mixtures has been widely used to other kind of data as well, for example mixtures of Gaussians for feature vectors. A number of simple base models are superimposed to construct a more complex distribution, with more free parameters (and thus greater expressivity) to be fitted into the sampled data. Assuming a mixture rather than a single generative model drastically increases the difficulty of our problem because for each sample we have to guess its 'label', the id of the underlying cluster it has come from in order to assign it to its correct group of similar samples. The closer the underlying centers are and the more variance the samples have around these centers, the more difficult it becomes to classify the samples into clusters. Incompleteness of the samples also plays a major role in separability making it even impossible to identify the latent centers if the samples are too short meaning that different clustering

solutions would be equivalent in fitting the observed incomplete data.

Probabilistic models on rank data have been widely studied in the last decades. The following surveys cover much of the progress in this field : [7], [8] and [9]. Many ranking generative models have been proposed, such as the Mallows model [10] and its generalisations and the parametric models of [11],[12], [13], [14] and [15]. In this work we focus on the Mallows model. The Mallows model has been studied extensively in the last decades and this research led to various theoretical results. Braverman and Mossel in [16] proposed an efficient algorithm for computing the MLE of the central ranking with small error and high probability. We will present their work in detail in chapter 3. Tang in [17] studies the statistical properties of the MLE for the classical Mallows' ϕ model, as well as the Infinite Generalised Mallows model. He proves the biasedness of the spread parameter for the Mallows' ϕ model and the IGM with a single parameter. He also provides an upper and a lower bound for the convergence rate of the MLE of the central ranking to the correct value, in the case of the classical Mallows' ϕ model. Both bounds concern the probability that the MLE is different than the central ranking and they are exponentially decreasing on the number of samples. Another direction is exactly recovering the central ranking with high probability using an adequately large sample collection ([2], [18]) and estimating the spread parameters [3]. The authors provide lower and upper bounds for the sample complexity of the reconstruction. We will see some of these results in chapter 3.

In this work we consider two ways of generalising the classical Kendall-Mallows' ϕ model. Firstly, we assume that samples are incomplete in the sense of [19]. In fact, we consider the random selection mechanism of [20]. Secondly, we assume that the latent model is a mixture rather than a single Mallows model. We will first review research in the direction of incomplete samples. Fotakis, Kalavasis and Stavropoulos in [19] generalise the results of [16] and [2] to the setting of incomplete samples. They show that the positional estimator, which effectively applies to incomplete samples, can replace the average position estimator that requires complete samples, sharing similar convergence identities. This way, the central ranking reconstruction problem can be solved and a good initialisation can be found for the local search for the MLE calculation, similarly to [16]. Moreover, the authors study the problem of learning the top-k alternatives of the central ranking using incomplete samples. The task breaks down into learning the identities and the relative order of these items. Asymptotically tight upper and lower bounds are provided for the sample complexity of this task. Hajek et al. in [21] study a selection mechanism that selects subsets of a fixed length uniformly at random. They analyze a rank-breaking scheme that decomposes partial rankings into pairwise comparisons. They show that even if one applies the mismatched maximum likelihood estimator that assumes independence (on pairwise comparisons that are now dependent due to rank-breaking), minimax optimal performance is still achieved up to a logarithmic factor. In this work as well as in [22] the estimator error depends on the spectral gap of the Laplacian of the comparison graph constructed by the samples (nodes correspond to items, edges to pairwise comparisons

and edges are appropriately computed from the samples). The authors in [22] study different comparison graph topologies and examine their optimality. A graph is considered optimal if for a given budget n on the number of samples the minimax risk is the smallest (up to constants) among all graphs. It is worth mentioning that in contrast to our setting, in the above setting the MLE problem is proved to be convex.

Now we will present related work from the field of Mallows mixtures. Awasthi, Blum et al. in [23] were the first to provide theoretical guarantees for the efficiency of learning the mallows mixture. They provide an algorithm with polynomial sample and time complexity that learns the parameters of a mixture of two Mallows models. Except for the number of components, no other significant assumption is made about the model. Chierichetti et al. in [24] study mixtures of Mallows with more than two centers, common spread parameters and arbitrary close distance between centers. They construct a $n! \times n!$ matrix where each row corresponds to the vectorisation of a Mallows model and each column to a different ranking on the domain of the Mallows model on n items. They show that the determinant of this matrix is non-zero thus establishing identifiability for an arbitrary number of components, but their proposed algorithm requires a sample complexity exponential on n. They also study separable Mallows Mixtures. Liu and Moitra in [5] establish the polynomial identifiability of the Mallows mixture making minimal assumptions. They prove that learning the centers exactly and estimating the weights and spread parameters up to some degree of precision can be done using a polynomial number of samples. The sample complexity is exponential only to the number of components, however this parameter is generally assumed to be a small constant. Mao et al. in [6] improve the dependency on the number of items, making it logarithmic and thus bridging the gap between learning a single Mallows and learning a Mallows mixture, it terms of the number of items. They also prove an optimal dependency on the spread parameter, however working on the special case when all spread parameters are equal.

There are also many heuristic approaches to the problem. Brendan Murphy and Donal Martin in [25] studied mixtures of Mallows models with various distance metrics (Kendall,Cayley and Spearman). They implemented an EM variant for the fitting problem and considered two criteria for choosing an appropriate model hypothesis class (e.g. the number of components and the distance metric). The model choice criteria were the Bayesian information criterion (BIC) and integrated complete likelihood (ICL). Experiments were conducted on synthetic data. Lu Tyler and Boutilier Craig in [26] applied the EM approach proposed by Neal and Hinton in [27] and exploited a novel Generalised Repeated Insertion Model approach for efficient sampling from Mallows posterior distributions. This allowed them to avoid working directly with the intractable posterior required in the E-step of the algorithm and perform Gibbs sampling instead. In [28] the Affinity Propagation clustering algorithm introduced in [29] is used to cluster the ranking samples of a Mallows Mixture. Once the clustering is performed, methods for single Mallows learning are applied inside each cluster. For the central ranking estimation, the Local Kemenization method, which was proposed in [30], is applied. For the spread parameter estimation the authors

propose several EM variants. The case of incomplete samples is also studied and the Local Kemenization method is adopted in this case as well. The authors in [31] assume a Dirichlet process mixture of Generalised Mallows Models. Samples are incomplete in the sense of top-k observations. The authors study two Gibbs sampling inference techniques for estimating posterior clusterings.

Learning the Mallows mixture without any assumptions is provably a difficult problem. Thus, it is common to consider special instances that are solved with much more efficient algorithms. In [32] Chierichetti et al. studied the problem of learning uniform mixtures of top-k Mallows models with a common spread parameter. The authors assume that centers are far from each other and single-linkage clustering succeeds with high probability for all samples. Thus, the problem is reduced to learning a single top-k Mallows model. Fabien et al. in [33] studied concentric mixtures of Mallows models, that is Mallows models with the same central ranking but different spread parameters. This models a heterogeneous population in terms of confidence about a ranking opinion (e.g. a population consisting of experts and non-experts). Interestingly, mixtures of concentric Gaussians are proved to be non-identifiable. The authors provide an algorithm for clustering the samples of a mixture of two concentric Mallows models under some separation condition of the spread parameters. They also extend the Borda algorithm of [2] for estimating the central ranking to the case of concentric Mixtures and top-k samples.

The problem of mixture learning has been studied in ranking models other than Kendall Mallows as well. Zhao et al. in [34] provide necessary conditions for the identifiability and of finite mixtures of Plackett-Luce models and sufficient conditions for generic identifiability. They also propose an efficient generalized method of moments (GMM) algorithm to learn the mixture of two Plackett-Luce models and show that the algorithm is consistent. Zhang et al. in [35] prove the generic identifiability of a range of ranking models with two components (Plackett-Luce, multinomial logistic model with slates of size 3 and BTL). They also provide a framework for verifying the number of solutions in a general family of polynomial systems using algebraic geometry. Anindya et al. in [36] consider a range of different noise models: the symmetric noise, the Heat kernel random walk under Cayley distance and the Cayley-Mallows model. They propose an algorithm that under certain mild assumptions applies to each of the above models and learns the unknown mixture to high accuracy, running in $O(n^{logk})$ time.

The (polynomial) identifiability of mixture models has been studied in other kinds of distributions as well. Teicher in [37] and [38] obtained sufficient conditions for the identifiability of a wide class of finite mixtures but these conditions do not apply in the setting of Mallows Mixtures. Another important direction of research is learning Gaussian mixtures. The Mallows model is closely related to the Gaussian Distribution as they both belong to the location-scale family. Thus, it is interesting to compare the methods and results in the field of learning Gaussian Mixtures to those in Mallows Mixtures.

Moitra and Valiant in [39] settled the polynomial learnability of Gaussian Mixtures, making minimal assumptions. The authors first solve the problem of learning mixtures of univariate Gaussian Mixtures and use this tool to tackle the multidimensional problem. A series of projections down to one dimension is considered. An important first step is to bring the multidimensional mixture in isotropic position, where the mean value is the all zero vector and the variance is 1 in every direction. This way, for each random direction, there exist two components whose projections have a polynomially large parameter distance, with high probability, and the univariate algorithm does not have to be executed with extreme precision in this direction to distinguish them. The univariate learning algorithm is applied in each projection and the estimates in different projections are used as constraints for the multidimensional parameters, making a linear system of equations that can be (robustly) backsolved to get estimations for the original multidimensional parameters. All projections are made on directions close to some random initial direction. In each projection the direction changes slightly and the parameters of the mixture change continuously. This way, one can match the components learned in one direction to those learned in another and the equations of the system are correctly aligned.
The univariate algorithm uses the method of moments and a brute-force gridsearch over candidate parameters. Each candidate model is compared with the samples in terms of the first 4k-2 moments. This number of moments is proved to be sufficient because it is connected to the number of the zero crossings of the mixture density. The univariate algorithm performs hierarchical clustering. Initially, some Gaussians may become very close when projected to the selected direction and they will appear as a single Gaussian. However, in this case the variance of the single Gaussian should be very small and thus this phenomenon can be detected. The solution is to isolate each Gaussian with small variance and bring it to isotropic position, revealing the subcomponents.
The authors conclude that given any n dimensional mixture of k Gaussians F that is ϵ-statistically learnable, we can output an ϵ-close (in parameter distance) estimate $\hat{F}$ and the running time and data requirements of the learning algorithm (for any fixed k) are polynomial in n, and $1/\epsilon$. They also prove that an exponential dependence on k is inevitable. Recent work (e.g. [40]) focuses on sufficient conditions to overcome this dependence.

Chapter 2

Permutations and Ranking Distributions

2.1 Permutations As Mathematical Objects

2.1.1 Permutations as Functions

Definition 2.1.1 (Permutation)**.** *A permutation is an arrangement of objects in a definite order. Technically, a permutation of a set S is defined as a bijection from S to itself. That is, it is a function from S to S for which every element occurs exactly once as an image value.*

For example consider the set S={1, 2, 3}. (3,1,2) is a permutation of S and it can be written as a function π where $\pi(1) = 3, \pi(2) = 1, \pi(3) = 2$. The inverse function π^{-1} gives the position of each element of S in the list representation of sequence π. We can also define partial permutations, which are ordered arrangements of k distinct elements selected from a set A, where $2 \leq k \leq |A|$. When k is equal to the size of the set, these are the (complete) permutations of the set. Let $n = |A|$. The number of (partial) permutations of S of length k is equal to $\frac{n!}{(n-k)!}$. Let A be some non empty set. $\mathbb{S}_A$ is the set of all (complete) permutations of A. If A is equal to $\{1, 2, 3, ..n\}$ we write $\mathbb{S}_A$ as $\mathbb{S}_n$.

2.1.2 Permutations as Groups

Definition 2.1.2 (Group)**.** *Let $G \neq \emptyset$ and $* : G \times G \rightarrow G$ be a binary operation. $(G, *)$ is a group if the following three requirements, known as group axioms, are satisfied:*

- *Operation $*$ is associative* $\iff \forall a, b, c \in G$ *it holds that* $a * (b * c) = (a * b) * c$.
- *G has an identity element* $\iff \exists e \in G : \forall a \in G \; a * e = e * a = a$
- *Every element of G has an inverse* $\iff \forall a \in G \; \exists \; a^{-1} \in G : a * a^{-1} = a^{-1} * a = e$

Three important properties, which can be derived from the above axioms are the uniqueness of the identity element, the uniqueness of the inverse of each element and the existence of a unique solution to the equation $a * x = b$ with respect to x, where $a, b, x \in G$.

Proposition 2.1.1. *Let A be a nonempty set, $\mathbb{S}_A$ the set of all its permutations and $\circ$ be the function composition operation. The structure $(\mathbb{S}_A, \circ)$ is a group.*

Proof. First we show that permutation composition is an internal operation in $\mathbb{S}_A$. We consider two permutations π, σ and we want to show that the composition $\pi \circ \sigma$ is injective

and onto. Let $x_1, x_2 \in \mathbb{S}_A$. If $(\pi \circ \sigma)(x_1) = (\pi \circ \sigma)(x_2) \Rightarrow \pi(\sigma(x_1)) = \pi(\sigma(x_2)) \Rightarrow \sigma(x_1)) = \sigma(x_2)(because\ \pi\ is\ 1-1) \Rightarrow x_1 = x_2(because\ \sigma\ is\ 1-1)$. This implies that permutation composition injective. We will now show that it is surjective as well. Given a permutation $y \in \mathbb{S}_A$ and the composition $\pi \circ \sigma$ it is possible to find $x \in \mathbb{S}_A$, such that $\pi(\sigma(x)) = y$, by setting $x = \sigma^{-1}\pi^{-1}(y)$. Associativity is a direct consequence of the fact that function composition is associative. The identity element of the group is the identity function $\pi(x) = x$. As for the existence of inverse permutations this follows from the definition of permutations as bijective functions.

2.2 Distance Metrics

In this section we are going to present distance metrics between permutations. These metrics are functions $\mathbb{S}_n \times \mathbb{S}_n \to \mathbb{R}$ that receive as argument a pair of permutations and output a value that measures the similarity between the two permutations. The higher the value the less similar the two permutations are. For example, consider permutations $\pi_1 = [1, 2, 3, 4]$ and $\pi_2 = [1, 3, 4, 2]$. We can easily see that the two permutations differ but how dissimilar are they? Also, does π_1 differ more from π_2 than it does from another permutation, for example $\pi_3 = [4, 3, 2, 1]$? There is no unique answer to these questions because different permutation distance metrics can be considered, all of which make sense intuitively, but are not equivalent. We are now going to discuss the most important ones of these metrics.

2.2.1 Kendall Tau Distance

This distance metric is equal to the number of pairs on which the two permutations discord. More formally this can be written as follows:
$d_{KT}(\pi_1, \pi_2) = \sum_{1 \le i < j \le n} \mathbb{1}\{(\pi_1(i) - \pi_1(j))(\pi_2(i) - \pi_2(j)) < 0\}$

KT distance satisfies the fundamental metric axioms:

1. It is a non-negative real-valued function : $d_{KT}(\pi_1, \pi_2) \ge 0$
2. The identity of indiscernibles holds: $d_{KT}(\pi_1, \pi_2) = 0 \Leftrightarrow \pi_1 = \pi_2$
3. It is symmetric: $d_{KT}(\pi_1, \pi_2) = d_{KT}(\pi_2, \pi_1)$.
4. The triangular inequality is satisfied: $d_{KT}(\pi_1, \pi_3) \le d_{KT}(\pi_1, \pi_2) + d_{KT}(\pi_2, \pi_3)$.

The KT distance is minimised at 0, when the two permutations are equal and it is minimised at $\frac{n(n-1)}{2}$, when the two permutations are reversals. It can be computed in $O(n^2)$ using a naive algorithm. By employing divide and conquer it can be sped up to $O(n \cdot log(n))$. Using the Van Emde Boas tree data structure the computation can be done in $O(n \cdot \sqrt{(log(n))})$.

• One important property of the KT distance is its independence of relabeling. In particular $d_{KT}(\pi_1, \pi_2) = d_{KT}(\pi_1\sigma, \pi_2\sigma)$, where $\pi_1, \pi_2, \sigma \in \mathbb{S}_n$ and $\pi\sigma(i) = \pi(\sigma(i)), \forall i \in [n]$.

• Another important property of the KT distance is swap increasingness.
That is $d_{KT}(\pi_{i\leftrightarrow j}, \sigma) \geq d_{KT}(\pi, \sigma) + 1$, where $i,j \in [n]$ are items , such that the pair(i,j) is the same order in π as in σ.
To prove this property we consider two cases. Firstly, if i and j are adjacent in π, then this pair becomes discordant and all the other pairs preserve their order. Thus, $d_{KT}(\pi_{i\leftrightarrow j}, \sigma) = d_{KT}(\pi, \sigma) + 1$. If i and j are not adjacent in π, then pair (i,j) becomes discordant. However, some pairs that involve either i or j and some alternative k that is ordered between them in π, might become concordant (with respect to σ) after the swap, while they were previously discordant. We will show that for each such pair another pair that was previously concordant becomes discordant, so in total these pairs do not decrease the distance. WLOG we assume that i>j in both π and σ. Consider an item k such that i>k>j in π and after the swap (i,k) becomes concordant. This means that k>i in σ and since i>j in σ then k>j in σ. Consequently, (j,k) becomes discordant after the swap. Similar arguments hold for the pairs (k,j) that become concordant.

One interesting question concerning the KT distance is how many permutations lie of the hypercircle of radius r. Consider a fixed permutation of reference $\pi \in \mathbb{S}_n$. We would like to know the number A(n,d) of permutations $\sigma \in \mathbb{S}_n$ that satisfy the equation $d_{KT}(\pi, \sigma) = d$. This is a combinatorial problem and the solution has been proved to be the following:

$$A(n,k) = \begin{cases} 1 & n = 1, k = 0 \\ 0 & n < 0, k < 0 \ \ or \ \ k > \frac{n(n-1)}{2} \\ \sum_{j=0}^{n-1} A(n-1, k-j) & otherwise \end{cases}$$

The recursion step can be done in an equivalent but more efficient way:
A(n,k) = A(n,k-1) + A(n-1,k) - A(n-1,k-n).
Unfortunately, no closed form expression can be derived for the two-dimensional sequence A(n,k), which is called the Triangle of Mahonian numbers.

In chapter 5 we discuss some useful properties of the Mahonian numbers, for example symmetry and we try to provide some convenient closed form bounds for these numbers. One property of the Mahonian numbers that is worth mentioning (although it falls outside the scope of our contribution) is the relation to the "Major Index". In particular A(n,k) is also equal to the number of permutations $\pi = (\pi(1), \ldots, \pi(n))$ of $\{1..n\}$ such that $\sum_{i:\pi(i)>\pi(i+1)} = k$. In this case k is called the Major index of π. For more information on the Mahonian numbers one can visit the The On-Line Encyclopedia of Integer Sequences (OEIS) and look up sequence number A008302.

2.2.2 Other distances

Hamming distance

This distance metric counts the number of positions at which the two permutations differ.
$d_{Ham}(\pi_1, \pi_2) = \sum_{i=1}^{n} \mathbb{1}\{{\pi_1}^{-1}(i) \neq {\pi_2}^{-1}(i)\}$
This metric fails to capture how great the displacement of each element is. It only con-

siders the existence of a displacement. For example, rotating a permutation by one step to the right, or swapping adjacent elements ($item_{2i-1} \leftrightarrow item_{2i}, \forall i \in [n/2]$) has the same effect as reversing the permutation. However, in most cases this is counter-intuitive because the former actions lead to rankings similar to the initial one, while the latter gives a permutation utterly different from the initial. We will now present a metric that takes the dislocation of each element into account.

Spearman's footrule

This metric sums the absolute dislocations of the items between their position in the first permutation and their position in the second permutation.

$d_{Sf}(\pi_1, \pi_2) = \sum_{i=1}^{n} |\pi_1(i) - \pi_2(i)|$

An important inequality holds for the KT distance and the Spearman's footrule as shown in [41]. The inequality is the following: $d_{KT}(\pi_1, \pi_2) \leq d_{Sf}(\pi_1, \pi_2) \leq 2d_{KT}(\pi_1, \pi_2), \forall \pi_1, \pi_2 \in \mathbb{S}_n$ and it is tight.

2.3 Ranking Distributions

2.3.1 The Mallows Model

This model resembles the normal distribution but instead of vectors it is defined on elements of $\mathbb{S}_n$. Like the normal distribution, the Mallows model is described by a central parameter and a spread parameter. The probability assigned to each element in the support set is inversely proportional to an exponential of the distance between the element and the central parameter and the base of this exponential depends on the spread parameter. More formally, if a random permutation $\pi \in \mathbb{S}_n$ follows the Mallows distribution $\mathcal{M}(\pi_0, \phi)$, then $\mathbb{P}[\pi = \sigma] = \frac{\phi^{d(\pi_0,\sigma)}}{Z(\phi,n)}$.

- $\pi_0 \in \mathbb{S}_n$ is the central permutation of the model. It expresses the underlying "ground truth" about the preferences of the population and it is the most probable permutation in the support set (the mode of the model).

- $\phi \in (0, 1)$ is the spread parameter. The higher its value, the more dispersed the samples are around the central permutation. In the extreme case where ϕ approaches zero, the only sample with a non zero probability of appearance is the central permutation, so a constant distribution is formed. In the opposite extreme case, where ϕ approaches one, all permutations in $\mathbb{S}_n$ have the same probability to appear, so the Mallows model degenerates into a uniform distribution over $\mathbb{S}_n$.

- $d : \mathbb{S}_n \times \mathbb{S}_n \rightarrow \mathbb{R}$ is some distance metric, for example the KT distance, the Spearman's footrule or the Hamming distance. In this work we focus exclusively on the KT distance.

- $Z(\phi, n)$ is the normalisation constant, which makes the density function sum to 1 so that it expresses probability. In our case, where d is the KT distance, $Z(\phi, n) =$

$$\prod_{i=1}^{n} Z_i(\phi) = \prod_{i=1}^{n} \left(\sum_{j=0}^{i-1} \phi^i \right) = \frac{1}{(1-\phi)^{n-1}} \prod_{i=2}^{n} (1-\phi^i).$$

We will now discuss two generating mechanisms, which produce permutations that follow the Mallows distribution. This concept is interesting for two reasons. Firstly, the Mallows Model is proved to be equivalent to some other models, which do not seem similar to it at first glance. Secondly, two sampling algorithms are provided, one of which is an efficient one that is used in practice in order to generate synthetic samples from a Mallows Model.

Nicolas de Condorcet studied probabilistic rankings two centuries earlier than Mallows and Kemeny, in the context of collective political decision making (Condorcet, 1785). According to Condorcet, members of society, or voters, express their opinion in the form of a ranking over choices. These choices (e.g policies) affect the society and one has to judge them by their benefits and consequences. Condorcet assumed that some (latent) objective ranking orders choices from most to least beneficial to society and that each voter is able to provide an independent, random comparison of any pair of choices: if a > b, in the objective ranking, a voter will (correctly) vote for a against b with probability 1-p, or (erroneously) vote for b against a with probability p, where p<1/2.

Pairwise Comparison Sampling of Mallows
(Condorcet noisy ranking process)

1. Let π_0 be the reference ranking and $0 \leq p \leq 1/2$.
2. Initialize $v \leftarrow \emptyset$.
3. For each pair of items x, y in A, such that x>y in π_0:
   (a) with probability 1-p add x>y to v,
   (b) otherwise add x<y to v.
4. If v is intransitive, go back to step 1 and start over.
5. v is transitive and corresponds to a ranking.

The distribution deriving from the above procedure is the following:

$$P(v \mid \pi_0, p) = \frac{1}{Z'} \prod_{\{x,y\} \subseteq A} \begin{cases} p & \text{if } v \text{ and } \pi_0 \text{ disagree on } x, y \\ 1-p & \text{otherwise} \end{cases}$$

$P(v \mid \pi_0, p) = \frac{1}{Z'} p^{d(v,\pi_0)} (1-p)^{s(v,\pi_0)} = \frac{1}{Z'} p^{d(v,\pi_0)} (1-p)^{\binom{n}{2} - d(v,\pi_0)} = \frac{1}{Z'} (1-p)^{\binom{n}{2}} \left(\frac{p}{1-p} \right)^{d(v,\pi_0)}$.
We set $\phi = \frac{p}{1-p}$ and notice that

$$Z' = (1-p)^{\binom{n}{2}} \left(1 + \frac{p}{1-p}\right) \left(1 + \frac{p}{1-p} + \left(\frac{p}{1-p}\right)^2\right) \ldots \left(1 + \ldots \left(\frac{p}{1-p}\right)^{n-1}\right) \tag{2.1}$$

$$= (1-p)^{\binom{n}{2}} Z\left(\frac{p}{1-p}, n\right) \tag{2.2}$$

Thus, $P(v \mid \pi_0, p) = \mathcal{M}(\pi_0, \phi)(v)$.

The Condorcet/Mallows sampling procedure did not originate from a demand of efficient sampling algorithms but an attempt to model voting procedures. As a result, it happens to be computationally inefficient, since it relies on rejection of partially constructed rankings as soon as a single circular triad (a > b > c > a) is drawn. The Repeated Insertion Sampling method provides an efficient alternatives and shows the relation between the Mallows and the RIM model, which we will discuss later.

RIM Sampling of Mallows

1. Let π_0 be the reference ranking and ϕ the spread parameter.
2. Start with an empty ranking r.
3. For i = 1..n:
   - Insert $\pi_0[i]$ into r at rank position $j \leq i$ with probability $\phi^{i-j}/(1 + \phi + ... + \phi^{i-1})$

The above algorithm produces a sample r that follows the Mallows distribution $\mathcal{M}(\pi_0, \phi)$. The complexity of the algorithm is equal to the total number of Bernoulli draws that take place. The worst case complexity is $O(n^2)$, the same as insertion sort). However, the average-case time complexity can be much smaller, since insertions at each stage of the algorithm are likely to occur near the bottom of the partial ranking. The expected time complexity of the algorithm is proportional to $\sum_{i=1}^{n}\left(\frac{\sum_{j=0}^{i-1}(j+1)\phi^j}{\sum_{j=0}^{i-1}\phi^j}\right) = \sum_{i=1}^{n}\left(\frac{1}{1-\phi} - i\phi^i\right) \leq \frac{n(1+\phi^{n+1})}{1-\phi} - \frac{\phi(1-\phi^n)}{(1-\phi)^2}$. Thus the average complexity is $O\left(min\left\{\frac{n(1+\phi^{n+1})}{1-\phi} - \frac{\phi(1-\phi^n)}{(1-\phi)^2}, n^2\right\}\right)$.

2.3.2 The Mallows Mixture Model

The Mallows Mixture Model $\mathcal{M}$ is parameterized by its set of central permutations π_i, the weights w_i and spread parameters ϕ_i corresponding the the centers π_i. The probability mass function of the Mallows mixture is:

$$M(\pi = \sigma) = \sum_{i=1}^{k} w_i \cdot \frac{\phi_i^{d_{KT}(\pi_i,\sigma)}}{Z(\phi_i, n)}$$

Each central permutation π_i is a distinct permutation of n items ($\pi_i \in \mathbb{S}_n$ and $\pi_i \neq \pi_j$ for $i \neq j$). The weights w_i are non negative and sum to one ($\sum_{i=1}^{k} w_i = 1$). The sampling process has two steps. Firstly, a center $i \in [n]$ is chosen with probability w_i. Then, a permutation is sampled from the single Mallows Model $\mathcal{M}(\pi_i, \phi_i)$ as analysed in previous chapters. In the scope of this work all spread parameters are supposed to be equal ($\phi_i = \phi \forall i \in [n]$).

2.3.3 The Selective Mallows Model

The probability mass function of this model is $\mathbb{P}[\pi = \sigma] = f(s) \cdot \frac{\phi^{d_{KT}(\pi_0,\sigma)}}{Z(\phi,|s|)}$, where s is the set of items found in σ. Each observation π is a permutation of the items appearing in its

corresponding selection set s. f(s) is the selection mechanism, a probability function that assigns a probability of selection to each subset of the full set of items $[n]$. π_0 is the central permutation of the model and it is complete (it contains all the items in [n]). $d_{KT}(\pi_0, \pi)$ is the Kendall's Tau distance between the central permutation and the sample. It needs to be redefined because the sample is possibly incomplete. A natural generalisation of the classical definition is the following:

$$d_{KT}(\pi_0, \pi) = \sum_{a,b \in s \wedge a<b} \mathbb{1}\{(\pi_0(a) - \pi_0(b)) \cdot (\pi(a) - \pi(b)) < 0\}$$

What differs is that the sum counts discordant pairs (a,b) where $a, b \in s$, rather than $a, b \in [n]$, where s is the selection set.

There is also another (less realistic) version of the selective Mallows Model, where the sampling process first draws a complete sample from the latent Mallows Model M and then projects it into some random selection set s. In this case the pmf is written as follows:

$$f_{M|s}(\pi) = \mathbb{P}_{\sigma \sim M}\{\sigma|_s = \pi\} \cdot f(s)$$

Definition 2.3.1. *A selection mechanism f(s) is said to be p-frequent with respect to l-wise comparisons for some order l, if for all sets $x \subseteq \{1, ..n\}$ with length less or equal to l $\mathbb{P}\{x \subseteq s\} \geq p \Leftrightarrow \forall x \ \sum_{x \in s} f(s) \geq p$.*

2.3.4 The Selective Mallows Mixture Model

This model combines the properties of the selective Mallows model and the Mixture Mallows model. It is a mixture model, because a collection of distinct centers $\{\pi_1, ...\pi_k\}$ rather than a single central ranking is assumed. It is also selective because samples generated by this model do not contain all possible alternatives but a random subset J of them, which is given by a selection mechanism $f(J)$ for each sample. The probability mass function of the model is the following:

$$M(\pi = \sigma) = f(J) \cdot \sum_{i=1}^{k} w_i \cdot \frac{\phi^{d_{KT}(\pi_i||_J, \sigma)}}{Z(\phi, |J|)}$$

The sample generating process consists of three steps. In the first step, selection mechanism $f(J)$ selects a random subset J of items in $[n]$ with probability $f(J)$. Then one of the k components of the mixture is activated with probability given by the mixing weights. Component i has probability w_i to be activated each time a sample is drawn. Finally, a random permutation π of the items in J is drawn from the Mallows Model $M_i(\pi) = \frac{\phi^{d_{KT}(\pi_i||_J, \pi)}}{Z(\phi, |J|)}$, where i is the index of the activated component. Notice that center π_i is restricted on J ($\pi_i||_J$) and the KT distance function counts discordant pairs only on items appearing in J.

We can also define a version of the Selective Mixture model that first draws a complete sample from a latent complete Mallows Mixture Model M and then projects it into some

random selection set s. The pmf is written as follows:

$$f_{M|s}(\pi) = \mathbb{P}_{\sigma\sim M}\{\sigma|_s = \pi\} \cdot f(s)$$

2.3.5 The RIM Model

In this model samples are generated by an iterative procedure that inserts alternatives into the constructed permutation one after another. The position at which each new element is placed follows Multinoulli distribution. The probabilities of these position distributions are parameters of the model, along with the central ranking.

We consider the model $RIM(\pi_0, \Pi)$. $\pi_0 \in \mathbb{S}_n$ is the latent "ground truth" permutation. Parameter Π is called insertion probability function and it assigns a probability $\Pi(i,j)$ to each pair of indices (i,j), $1 \leq j \leq i \leq n$, such that $\sum_{j=1}^{i} \Pi(i,j) = 1$ for all i in $\{1, 2, ...n\}$. A random ranking $r \sim RIM(\pi_0, \Pi)$ is generated by the following randomized process:

RIM Sampling

1. Let π_0 be the reference ranking and ϕ the spread parameter.
2. Start with an empty ranking r.
3. For i = 1..n:
   - Insert $\pi_0[i]$ into r at rank position $j \leq i$ with probability $\Pi(i,j)$

Note that the insertion position of each $\pi_0[i]$ is probabilistically independent of the positions of the previous items $\pi_0[1], ..., \pi_0[i-1]$. We also observe that every insertion sequence results to a unique ranking.

2.3.6 The Plackett-Luce Model

The Plackett-Luce model was introduced independently by the two scientists it is named after, [14],[15]. It is different from the aforementioned models in the sense that no reference permutation is assumed. Instead the alternatives are assumed to have different values w_i and the probability they are chosen is proportional to these values. The model is parameterized by the vector of weights $W = (w_1, ...w_n) \in [0,1]^n$, such that $\sum_{i=1}^{n} w_i = 1$.

The sample generation process is performed in n rounds. In i-th round the alternative that will be placed in position i is picked with probability that is proportional to its weight. The probability mass function of the model is:

$$\mathbb{P}[\pi = \sigma] = \prod_{i=1}^{n} \left(\frac{w_{\sigma^{-1}(i)}}{\sum_{j=i}^{n} w_{\sigma^{-1}(j)}} \right)$$

The mode of this distribution is the ranking that places the alternatives in decreasing order of weights. $\sigma^* = argsort_{i\in[n]}\{w_1, ..., w_n\}$. An interesting property is that $\mathbb{P}[\pi(i) < \pi(j)] = \frac{w_i}{w_i + w_j}$, for $\pi \sim \mathcal{PL}(w)$. In this model dispersion is related to the variance of the weights. The closer the weights are to a uniform vector $[\frac{1}{n}, ...\frac{1}{n}]$, the closer the ranking distribution is to a uniform over $\mathbb{S}_n$.

Chapter 3

Distribution Learning

3.1 Definition of Learnability and Parameter Estimation

A class of distributions C is called efficiently learnable if for every $\epsilon > 0$ and $0 < \delta \leq 1$ given access to an oracle $GEN(D)$ that returns samples from an unknown distribution $D \in C$, there exists a polynomial time algorithm A, called learning algorithm of C, that outputs a generator or an evaluator of a distribution D' such that $Pr[d(D, D') \leq \epsilon] \geq 1 - \delta$, where d is some distance metric between distributions D and D', for example the TV distance or the KL divergence, which we will discuss later in this chapter.If we know that $D' \in C$ then A is called a proper learning algorithm, otherwise it is called an improper learning algorithm.

In some cases each distribution $D \in C$ is uniquely identified by a set of parameters. For example, the class of univariate Gaussian distributions $N(\mu, \sigma^2)$ is parameterized by the pair (μ, σ). Different values of (μ, σ) give different distributions $D \in C$, covering the whole class C. In this case algorithm A should be able to estimate the parameters (μ, σ) and we would call it a parameter learning algorithm.

3.2 PAC Learning

PAC-learning is a theoretical framework introduced by Valiant in [42] to study learning problems. In learning problems one aims to find the way in which elements of one set X are mapped on another set Y, the label set. We assume that a function $f : X \to Y$ performs this mapping and we aim to approximate this function by using a finite number of samples as input to a learning algorithm. The input consists of pairs (x,y), where $x \in X$ and $y = f(x) \in Y$. If no assumption is made about f all we can infer is what the sample data directly suggest and no generalisation is possible. The assumption we have to make is that f has a particular form and is in a particular class H of functions, called a hypothesis class, $H \subseteq Y^X$. We also assume that there is an unknown distribution D over the domain X which generates the samples and the samples are iid, so for each independent pair (x,y) $x \sim D$ and y=f(x). In order to quantify how good an estimation h of the function f is, we define loss functions $L_{D,f}(h)$ which output a non negative real number. The smaller this number is, the better the estimation of f. For example a loss function could be $L_{D,f}(h) = Pr_{x \sim D}[h(x) \neq f(x)]$. We can also define the loss function l empirically

on a group S of samples ($l : H \times S \to \Re^+$). In this case the theoretical framework focuses on the expected value of the loss over the distribution D of samples S. We are interested in hypothesis classes H that allow us to approximate f with as small error ϵ we want with probability at least $1 - \delta$, given enough samples. Then H is called PAC-learnable.

Definition 3.2.1 (PAC Learnability)**.** *Let $H \subseteq \{0,1\}^X$ be a hypothesis class of functions $f : X \to \{0,1\}$. H is called PAC learnable with respect to a loss function L, if there exists a sample complexity $N = N(H, \epsilon, \delta)$, where $\epsilon, \delta \in (0,1)$ such that for any $\epsilon, \delta \in (0,1)$, every distribution D over X and every labeling function $f : X \to \{0,1\}$, there exists an algorithm that given an input of size at least $N(H, \epsilon, \delta)$ of i.i.d. samples generated by D and labeled by f, returns with probability at least $1 - \delta$ a hypothesis $h \in H$ with $L_{D,f}(h) \leq \epsilon$*

3.3 Information Theory

3.3.1 KL Divergence and TV Distance

The Total Variation Distance between two discrete distributions P an Q is defined as $d_{TV}(P,Q) = sup\{|P(A) - Q(A)| : A \in F\}$, where F is a sigma-algebra of subsets of the sample space Ω.
A sigma-algebra on a set Ω is a collection Σ of subsets of Ω satisfying the following conditions :
(1) it includes Ω itself,
(2) it is closed under complement,
(3) it is closed under countable unions and
(4) it is closed under countable intersections.

The supremum is achieved at either $A = \{x : P(x) \geq Q(x)\}$ or its complementary set A^c.
But $P(A) - Q(A) + P(A^c) - Q(A^c) = 0$.
Thus, $d_{TV}(P,Q) = \sum_{x \in A} (P(x) - Q(x)) = \frac{1}{2}\left(\sum_{x \in A} (P(x) - Q(x)) + \sum_{x \in A^c} (Q(x) - P(x))\right) =$
$\frac{1}{2}\left(\sum_{x \in A} |P(x) - Q(x)| + \sum_{x \in A^c} |P(x) - Q(x)|\right) = \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)|$

In conclusion, an alternative expression for the TV distance between two discrete distributions P and Q is $d_{TV}(P,Q) = \frac{1}{2} \sum_{x \in \Omega} |P(x) - Q(x)|$, which is more practical than the formal definition that states that the TV distance between two distributions is the largest possible difference between the probabilities that the two distributions can assign to the same event.

Another distribution distance metric is the Kullback–Leibler divergence. It is defined as the relative entropy from the one distribution to the other:

$$D_{KL}(P||Q) = \sum_{x \in \Omega} P(x) \frac{P(x)}{Q(x)}$$

We observe that KL divergence is not a proper distance metric, because it is not symmetric. $D_{KL}(P||Q)$ can be different from $D_{KL}(Q||P)$. We also observe that it is defined only if for all x in sample space Ω, $Q(x) = 0$ implies $P(x) = 0$. However $P(x)$ can be zero without $Q(x)$ being zero at the same time, because $\lim_{P(x)\to 0^+} P(x)log(P(x)) = 0$. So, whenever $P(x) = 0$ the corresponding term of distance is interpreted as zero. It can be proved that KL divergence takes non negative values and that it is equal to zero if and only if the two distributions it takes as input are equal.

Viewing the KL divergence from an information theoretical perspective, it is the expected number of extra bits required to code samples from distribution P using a code optimized for distribution Q rather than the code optimized for P.
In the context of Bayesian inference, it can be interpreted as the amount of information lost when Q is used to approximate P. P is considered the "ground truth" distribution of data, while Q represents a model approximating P.

Two important properties of the KL divergence are the following:
• The chain rule says that $D_{KL}(P(x,y)||Q(x,y)) = D_{KL}(P(x)||Q(x)) + D_{KL}(P(x|y)||Q(x|y))$. Consequently, if $P(x,y) = P_1(x)P_2(y)$, where P_1, P_2 are independent and similarly $Q(x,y) = Q_1(x)Q_2(y)$, where Q_1, Q_2 are independent, then the KL divergence is additive over the two variables x and y: $D_{KL}(P||Q) = D_{KL}(P_1||Q_1) + D_{KL}(P_2||Q_2)$.
•KL divergence is convex in the pair of probability mass functions (p, q), i.e. if (p_1, q_1) and (p_2, q_2) are two pairs of probability mass functions and λ is some constant in $[0, 1]$ then $D_{KL}(\lambda p_1 + (1-\lambda)p_2 \parallel \lambda q_1 + (1-\lambda)q_2) \leq \lambda D_{KL}(p_1||q_1) + (1-\lambda)D_{KL}(p_2||q_2)$.

TV distance and KL are connected by Pinsker's inequality [43]:

$$d_{TV}(P,Q) \leq \sqrt{\frac{1}{2} D_{KL}(P||Q)}$$

This inequality is tight up to constant factors. However, it is trivial when $D_{KL}(P||Q) > 2$. Bretagnolle and Huber in [44] proved a sharper inequality :

$$d_{TV}(P,Q) \leq \sqrt{1 - e^{-D_{KL}(P||Q)}}$$

3.3.2 Fano's Inequality

Fano's Inequality bounds the error of approximating some random variable Y using knowledge of the correlated random variable X. The inequality involves the conditional entropy entropy $H(X|Y)$. $H(X|Y) = -\sum_{i,j} p(x_i, y_j) log\left(\frac{p(x_i,y_j)}{p(y_j)}\right)$, where $p(x_i, y_j)$ is the probability that $X = x_i$ and $Y = y_i$ and $p(y_j)$ is the probability that $Y = y_i$. Conditional entropy expresses the amount of randomness in the random variable X given the random variable Y.For example, if $X = f(Y)$, then $H(X|Y) = 0$.

Let X be a random variable following distribution $p(x)$ and let Y be a random variable

related to X through conditional distribution $p(y|x)$. We make an estimation $\hat{X}$ of X using function g and Y: $\hat{X} = g(Y)$. We observe that $\{X, \hat{X}, Y\}$ form a Markov chain $X \rightarrow Y \rightarrow \hat{X}$. We define error probability $P_e = P\{X \neq \hat{X}\}$. Fano's Inequality states that:

$$H(P_e) + P_e log|\mathcal{X}| \geq H(X|\hat{X}) \geq H(X|Y)$$

where $\mathcal{X}$ denotes the support domain of X and $H(P_e)$ is the error binary entropy: $H(P_e) = -P_e \cdot log(P_e) - (1 - P_e) \cdot log(1 - P_e)$.

A weaker version of the inequality is the following:

$$1 + P_e log|\mathcal{X}| \geq H(X|Y)$$

More material on the topic can be found in Fano's textbook [45].

3.4 Concentration Inequalities

3.4.1 Markov's Inequality

Let X be a non negative random variable and $a > 0$ a positive constant. It holds that:

$$\mathbb{P}\{X > a \cdot \mathbb{E}[X]\} \leq \frac{1}{a}$$

This inequality gives a measure of the concentration of a random variable around its expected value without making any assumption about the distribution family. The probability of drawing samples that are multiples of the expected value is inversely proportional to the multiplication factor applied to the expected value. These bounds do not require knowledge of any of the parameters of the distribution, except the expected value, however this may lead to relatively loose bounds.

3.4.2 Chebyshev's Inequality

Chebyshev's inequality bounds the probability that a random variable deviates far from its mean value. While Markov's inequality only required knowledge of the expected value, Chebyshev's inequality also requires knowledge of the variance. However, it does not make the assumption that the random variable is non negative, as Markov did.

Chebyshev's inequality can be derived from Markov's inequality by considering the random variable $Y = (X - \mathbb{E}[X])^2$. $\mathbb{E}[Y] = Var(X)$, by definition of variance.

We apply Markov's inequality on Y with a scaling of a^2 on the expected value, $a > 0$:

$$\mathbb{P}\{Y > a^2 \cdot \mathbb{E}[Y]\} \leq \frac{1}{a^2} \Leftrightarrow$$
$$\mathbb{P}\{(X - \mathbb{E}[X])^2 > a^2 \cdot Var(X)\} \leq \frac{1}{a^2} \Leftrightarrow$$
$$\mathbb{P}\{|X - \mathbb{E}[X]| > a \cdot \sigma_X\} \leq \frac{1}{a^2}$$

The final inequality is Chebyshev's inequality for random variable X. Note that $\mathbb{E}[X]$ must be finite and σ_X non zero.

The inequality can be used to construct confidence intervals. For example, to ensure that an interval centered on the mean value includes at least 95% of the total probability mass, the interval must have length at least 10 times the standard deviation ($a = 5$).
Of course, if we had more information about the distribution of the random variable better bounds might be provided. For example, on a univariate Gaussian it suffices to take an interval of length 4 times the standard deviation centered on the mean to achieve a confidence of 95% , not 10 as the Chebyshev inequality would imply. This is due to the strong concentration property of the Gaussian. However, all random variables have some concentration tendency according to Chebyshev's inequality.

3.4.3 Chernoff Bounds

Other bounds derived by the the Markov inequality are the Chernoff Bounds (first appeared in [46]). In particular, we consider random variable e^{tX} and apply the Markov inequality on it. Thinking of this variable as a Taylor expansion series, it captures all orders of moments of the distribution. The more moments used the better the tail bounds derived, because we use more information about the distribution. Under certain conditions the sequence of moments can uniquely determine the distribution, through the characteristic function $\phi_X(a) = \mathbb{E}[exp(i \cdot aX)]$, as long as the characteristic function has an infinite radius of convergence. We observe that the characteristic equation is very similar to the function used to derive the Chernoff bounds. The only difference is the introduction of the imaginary unit.

The inequalities we discussed earlier used lower moments of the distribution of X so the bounds were less tight. Markov's inequality that only used the first moment (the mean) provided bounds with a linear dependency on the error. Chebyshev's inequality used the first two moments(mean and variance) and guaranteed a tail decay inversely proportional to the squared deviation from the mean value. The Chernoff bounds provide exponentially decreasing bounds but require knowledge of the expected value of an exponential function of the random distribution.

We will now present the way in which generic Chernoff bounds are derived.
For every $t > 0$ we have that $\mathbb{P}\{X \geq a\} = \mathbb{P}\{e^{tX} \geq e^{t \cdot a}\} \leq \frac{\mathbb{E}[e^{tX}]}{e^{t \cdot a}}$. Calculating the quantity

$\mathbb{E}[e^{tX}]$ we can minimize the RHS of the inequality with respect to t to achieve tight bounds.

An interesting case is when X is a sum of n i.i.d. random variables X_i.
Then $\mathbb{E}[e^{tX}] = \prod_{i=1}^{n} \mathbb{E}[e^{tX_i}]$, so $\mathbb{P}\{X \geq a\} \leq min_{t>0}\left\{e^{-t\cdot a}\prod_{i=1}^{n}\mathbb{E}[e^{tX_i}]\right\}$.

For left tail bounds we work with variable e^{-tX} and yield
$\mathbb{P}\{X \leq a\} \leq min_{t>0}\left\{e^{t\cdot a}\prod_{i=1}^{n}\mathbb{E}[e^{-tX_i}]\right\}$.

3.4.4 Hoeffding Bounds

In [1] the Chernoff–Hoeffding theorem is proposed, which provides exponentially decreasing tail bounds for sums of independent Bernoulli variables.
Suppose $X_1, ...X_n$ are i.i.d. Bernoulli variables and $X = \sum_{i=1}^{n} X_i$ with $\mathbb{E}[X] = p$. Then for all $a \in (0, n-p)$:

$$\mathbb{P}[X \geq p + a] \leq e^{-n\cdot D_{KL}(\frac{p+a}{n}\|\frac{p}{n})}$$

and for all $a \in (0, p)$:

$$\mathbb{P}[X \leq p - a] \leq e^{-n\cdot D_{KL}(1-\frac{p-a}{n}\|1-\frac{p}{n})}$$

where $D_{KL}(x\|y) = x\cdot ln(\frac{x}{y}) + (1-x)ln\left(\frac{1-x}{1-y}\right)$ is the Kullback–Leibler divergence between two Bernoulli distributions with parameters x and y respectively.

Theorem 2 of [1] provides exponential tail bounds for sums of independent bounded variables.
Let $X_1, ...X_n$ be independent variables and each X_i is bounded by interval $[a_i, b_i]$. Let $\bar{X}$ be the empirical mean of these variables, $\bar{X} = \frac{1}{n}(X_1 + ... + X_n)$. Then for $t > 0$:

$$\mathbb{P}\{\bar{X} - \mathbb{E}[\bar{X}] \geq t\} \leq exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

$$\mathbb{P}\left\{\left|\bar{X} - \mathbb{E}[\bar{X}]\right| \geq t\right\} \leq 2exp\left(-\frac{2n^2t^2}{\sum_{i=1}^{n}(b_i - a_i)^2}\right)$$

In this work we make extensive use of Hoeffding bounds for binomial distributions.
Let $X \sim Bin(n, p)$. Then:

$$\mathbb{P}[X \leq k] \leq exp\left(-2n\left(p - \frac{k}{n}\right)^2\right)$$

Proof

$\mathbb{P}\{X \leq k\} = \mathbb{P}\{Y \geq n-k\}$, where Y is the complementary binomial variable of X ($Y = n - X$). Hoeffding inequality can be applied for binomial variable Y because it is a sum of inde-

pendent Bernoulli variables Y_i. Bernoulli variables are bounded in $[0, 1]$.

$$\mathbb{P}\{X \leq k\} = \mathbb{P}\{Y \geq n-k\} = \mathbb{P}\{Y - \mathbb{E}[Y] \geq n-k-(1-p)n\} =$$
$$\mathbb{P}\left\{\sum_{i=1}^{n} Y_i - \mathbb{E}\left[\sum_{i=1}^{n} Y_i\right] \geq pn-k\right\} =$$
$$\mathbb{P}\left\{\frac{1}{n}\sum_{i=1}^{n} Y_i - \mathbb{E}\left[\frac{1}{n}\sum_{i=1}^{n} Y_i\right] \geq p-\frac{k}{n}\right\} \leq exp\left(-\frac{2n^2\left(p-\frac{k}{n}\right)^2}{n}\right) = exp\left(-2n\left(p-\frac{k}{n}\right)^2\right)$$

3.5 Learning the Mallows Model

3.5.1 Reconstructing the Central Ranking

In this section we will present some important results of [2]. Suppose we are given a set of N samples $\sigma_1, \ldots \sigma_N$, drawn from a Mallows Model. The samples are possibly incomplete. We want to use these samples to estimate the latent central permutation with high probability. For this purpose we will use a kind of pairwise majority consistent (PM-c) estimator, called the Positional Estimator, $\hat{\pi}$.

$$\hat{\pi}[i] = 1 + \sum_{j \in [n] \setminus \{i\}} \mathbb{1}\left\{\sum_{k=1}^{N} \mathbb{1}\{j > i \text{ in } \sigma_k\} > \sum_{k=1}^{N} \mathbb{1}\{i > j \text{ in } \sigma_k\}\right\}, \ \forall i \in [n]$$

$\hat{\pi}$ estimates the position of each item in the latent central permutation. Ties may arise, which are broken uniformly from left to right. If N is sufficiently large, then the positional estimator retrieves the correct latent central permutation π_0 with high probability, as we will see in the next theorem.

Theorem 3.5.1. *Let $\mathcal{M}(\pi_0, \phi)$ be a Mallows distribution with central ranking $\pi_0 \in \mathbb{S}_n$ and spread parameter $\phi \in (0, 1)$. For any $\epsilon > 0$, given a sample profile drawn from $\mathcal{M}(\pi_0, \phi)^N$ for any N at least equal to some value $O\left(\frac{log(n/\epsilon)}{(1-\phi)^2}\right)$, the positional estimator retrieves the central ranking π_0 with probability at least $1-\epsilon$.*

Proof.
Samples are assumed to be complete. However a similar analysis can be made in the selective setting. The difference is that each pairwise comparison has its own sample complexity rather than a common complexity N. Introducing the notion of p-frequency, N can be replaced by $p \cdot N$ in the analysis, where p is the frequency of the least frequent pair in the samples.

For each pair of alternatives i, j, we let $q(i > j)$ be the number of rankings in the sample set, which place item i before item j. Let $\hat{\pi}$ be the estimation of the central ranking returned by the positional estimator. WLOG we assume that the latent central ranking π_0 is the identity ranking and we bound the probability of the event $\hat{\pi} \neq \pi_0$ from above.

$$\Pr_r[\hat{\pi}(\mathrm{I}) \neq \pi_0] \leq \Pr[\exists i < j : q(i > j) \leq q(j > i)] \leq \sum_{i<j} \Pr[q(i > j) \leq q(j > i)]$$

The value of $\Pr[q(i \succ j) \leq q(j \succ i)]$ depends on the distance of items i and j in the central ranking (or the restriction of the central ranking on the selection sets in the selective setting). Items that are closer in the central ranking have a greater probability to swap. The probability of swapping i and j also increases as the selection mechanism drops items that are placed between i and j in the central ranking. To bound the probability of swap from above we consider the worst case, where i and j are adjacent. In this case the event of swap follows the Bernoulli distribution with parameter $\frac{\phi}{1+\phi}$. We consider N variables $X_\ell \sim Be\left(\frac{\phi}{1+\phi}\right)$, one for each sample. Since samples are iid, X_l are also iid. We also consider the complementary variables $Y_l = 1 - X_l$. Then we have:

$$\Pr[q(i \succ j) \leq q(j \succ i)] \leq \Pr\left[\sum_{l\in[N]} (X_l - Y_l) \geq 0\right] =$$
$$\Pr\left[\frac{1}{N}\sum_{l\in[N]} (X_l - Y_l) - \frac{\phi-1}{1+\phi} \geq \frac{1-\phi}{1+\phi}\right] \leq \exp\left(-2N\left(\frac{1-\phi}{1+\phi}\right)^2\right)$$

where the last step follows from Hoeffding's inequality. We set $\zeta := \left(\frac{1-\phi}{1+\phi}\right)^2$. Returning to the overall probability of error we have:

$$\Pr[\hat{\pi} \neq \pi_0] \leq n^2 \exp(-2N\zeta)$$

We set $n^2 \exp(-2N\zeta)$ equal to the tolerance of error probability ϵ and solve for N. This way, the desired result is obtained.

The bound for the sample complexity is in fact tight, as the following theorem states.

Theorem 3.5.2. *For any $\epsilon \in (0, 1/2]$ and any central ranking estimator, there exists a central ranking $\pi_0 \in \mathbb{S}_n$ such that, for any $\phi \in (0, 1)$, the estimator, given a sample profile drawn from $\mathcal{M}(\pi_0, \phi)^N$, retrieves π_0 with probability at least $1 - \epsilon$, only if $N = \Omega\left(\frac{\log(n/\epsilon)}{\log(1/\phi)}\right)$.*

Proof.
The proof is based on the idea that any estimator could mistake the latent central ranking for some other ranking close to it, with non negligible probability.
Let $\tilde{\pi}$ be any (possibly randomized) estimator of the. central ranking. Assume that:

$$\Pr_{\Pi\sim(M_{\pi_0,\phi})^N}[\tilde{\pi}(\Pi) = \pi_0] \geq 1 - \epsilon, \forall \pi_0 \in \mathcal{S}_n$$

Let $\pi_0 \in S_n$. We define the neighbourhood of π_0 as the set $\mathcal{N}(\pi_0) = \{\sigma_0 \in S_n : d_{KT}(\sigma_0, \pi_0) = 1\}$. The cardinality of this set is $|\mathcal{N}(\pi_0)| = n - 1$. In this proof π_0 plays the role of the latent central ranking and set $\mathcal{N}(\pi_0)$ is a set of hard instances, that is instances that the estimator has high to output instead of the correct one.
Moreover, for each observation π in Π_{ob} and $\sigma_0 \in \mathcal{N}(\pi_0)$ we have from triangle inequality that $d_{KT}(\pi, \sigma_0) \leq d_{KT}(\pi, \pi_0) + d_{KT}(\sigma_0, \pi_0) = d_{KT}(\pi, \pi_0) + 1$. Thus, for any $\sigma_0 \in \mathcal{N}(\pi_0)$ it holds that $\Pr[\Pi_{ob} \mid \sigma_0] \geq \phi^N \Pr[\Pi_{ob} \mid \pi_0]$.
We start from the assumption about the high accuracy of the estimator and use the above observations to obtain a lower bound about the sample complexity.

$$\Pr_{\Pi_{ob}\sim(M_{\pi_0,\phi})^N}\left[\tilde{\pi}(\Pi_{ob}) = \pi_0\right] \geq 1-\epsilon \Leftrightarrow$$

$$\sum_{\Pi\in S_n^N} \Pr_{\Pi_{ob}\sim(M_{\pi_0,\phi})^N}\left[\Pi = \Pi_{ob}\right] \cdot \Pr\left[\tilde{\pi}(\Pi) = \pi_0\right] \geq 1-\epsilon \Leftrightarrow$$

$$1 - \sum_{\Pi\in S_n^N} \Pr_{\Pi_{ob}\sim(M_{\pi_0,\phi})^N}\left[\Pi = \Pi_{ob}\right] \cdot \Pr\left[\tilde{\pi}(\Pi) \neq \pi_0\right] \geq 1-\epsilon \Leftrightarrow$$

$$1 - \sum_{\Pi\in S_n^N} \Pr_{\Pi_{ob}\sim(M_{\pi_0,\phi})^N}\left[\Pi = \Pi_{ob}\right] \cdot \left(\sum_{\sigma_0\in\mathcal{N}(\pi_0)} \Pr\left[\tilde{\pi}(\Pi) = \sigma_0\right] + \sum_{\sigma_0\in(S_n-\mathcal{N}(\pi_0)-\{\pi_0\})} \Pr\left[\tilde{\pi}(\Pi) = \sigma_0\right]\right) \geq 1-\epsilon \Rightarrow$$

$$1 - \sum_{\Pi\in S_n^N} \Pr_{\Pi_{ob}\sim(M_{\pi_0,\phi})^N}\left[\Pi = \Pi_{ob}\right] \cdot \sum_{\sigma_0\in\mathcal{N}(\pi_0)} \Pr\left[\tilde{\pi}(\Pi) = \sigma_0\right] \geq 1-\epsilon \Rightarrow$$

$$1 - \sum_{\sigma_0\in\mathcal{N}(\pi_0)} \sum_{\Pi\in S_n^N} \phi^N \Pr_{\Pi_{ob}\sim(M_{\sigma_0,\phi})^N}\left[\Pi = \Pi_{ob}\right] \cdot \Pr\left[\tilde{\pi}(\Pi) = \sigma_0\right] \geq 1-\epsilon \Leftrightarrow$$

$$1 - \phi^N \sum_{\sigma_0\in\mathcal{N}(\pi_0)} \Pr_{\Pi_{ob}\sim(M_{\sigma_0,\phi})^N}\left[\tilde{\pi}(\Pi_{ob}) = \sigma_0\right] \geq 1-\epsilon \Rightarrow$$

$$1 - \phi^N(n-1)(1-\epsilon) \geq 1-\epsilon$$

The final inequality implies that N is $\Omega\left(log\left(\frac{n}{\epsilon}\right)\right)$ for all estimators.

This lower bound is tight with respect to the upper bound of the sample complexity, because for $\phi \to 1$ $\frac{1}{(1-\phi)^2} = O(\frac{1}{log(1/\phi)^2})$.

3.5.2 MLE of the Central Ranking in the Mallows Model

Braverman and Mossel in [47] give an efficient algorithm for computing a maximum likelihood estimation for the Mallows Model. The goal is to find a permutation π^* that best fits a sample set of r independent observations $\pi_1, \ldots, \pi_r$ drawn from the Mallows Model.

Definition 3.5.1. *The Mallow Reconstruction Problem (MRP) is the problem of finding a permutation π^* maximizing the quantity*

$$\prod_{k=1}^{r} \Pi\left[\pi_k \mid \pi^*\right] = \frac{1}{Z(\phi)^r} \phi^{\sum_{k=1}^{r} d_{KT}(\pi_k,\pi^*)}$$

or equivalently minimizing

$$d\left(\pi^*\right) := \sum_{k=1}^{r} d_{KT}\left(\pi_k, \pi^*\right).$$

The optimization problem without any assumptions on the generating process is NP hard. However, leveraging the concentration properties of the Mallows Model, we can reduce the search space. The general idea is the following:

Firstly, apply a simple estimator that ranks the items according to their average index in

the samples. In the produced estimation, with high probability, all elements lie close to their true positions according to the latent central ranking. That is they lie O(log(n) places away from their correct position, with constants depending on the spread parameter and the sample complexity. We also know that the MLE solution indices are close to the indices of the latent central ranking (again the displacement is O(log(n), with constants depending on the spread parameter and the sample complexity). Thus, we expect the MLE solution to be close to the average index estimation $\overline{\pi}$, in the sense that all items in $\overline{\pi}$ are placed at most L places away from their position in the MLE, where L is O(log(n) with constants depending on the spread parameter and the sample complexity. The final step is to use a dynamic programming algorithm that finds an MLE searching locally in the space defined by the constraint that all items lie at most L places away from their average position in the samples. Typically, the time complexity is an increasing function of the input size, however in this case the contrary holds. As the input size grows (more samples are used) the time complexity decreases (because L decreases and the search space shrinks).

Now we are going to formulate the above ideas more strictly, presenting the results of [47].
We begin with a basic lemma that guarantees a geometric concentration of the location of each item around the "correct" location, that is the location of the item in the latent central ranking. The proof is based on the Mallows RIM sampling and it is omitted for brevity.

Lemma 3.5.1. *Let a be an element that is ranked k-th by π^*. In other words, $\pi^*(a) = k$. Then for $\pi \sim M(\pi^*, \phi)$ holds that $\mathrm{P}[|\pi(a) - k| \geq i] < 2 \cdot \phi^i / (1 - \phi)$, for all i.*

The next lemma analyses the behaviour of the average index estimator using the geometric concentration of indices in the samples.

Lemma 3.5.2. *Suppose that the permutations $\pi_1, \ldots, \pi_r$ are drawn from $M(\pi^*, \phi)$. Let $a = k$ be the element ranked k-th by π^*. Let $\overline{\pi(a)}$ be the average index of a under the permutations $\pi_1, \ldots, \pi_r$:*

$$\overline{\pi(a)} = \frac{1}{r} \sum_{i=1}^{r} \pi_i(a)$$

Then

$$\mathrm{P}[|\overline{\pi(a)} - k| \geq i] \leq 2 \cdot \left(\frac{(5i + 1) \cdot \phi^i}{1 - \phi} \right)^r$$

for all i.

Proof

For a vector $b = (b_1, \ldots, b_r)$ of non-negative integers let A_b denote the event that $\pi_j(a) \leq k - b_j$ for $j = 1, \ldots, r$ for which $b_j > 0$. By Lemma 3.5.1 we have

$$\mathrm{P}[A_b] < \frac{\phi^{\sum_{j=1}^{r} b_j}}{(1 - \phi)^r}.$$

Next, we note that the event $[\overline{\pi(a)} \leq k - i]$ is covered by

$$\bigcup_{\sum_{j=1}^{r} b_j = r - i} [A_b] .$$

Hence

$$P[\overline{\pi(a)} \leq k - i] < \#\left\{b : \sum_{j=1}^{r} b_j = ri\right\} \cdot \frac{\varphi^{ri}}{(1-\varphi)^r} = \binom{ri + r - 1}{r - 1} \cdot \frac{\varphi^{ri}}{(1-\varphi)^r} < \frac{(5i+1)^r \cdot \varphi^{ri}}{(1-\varphi)^r}$$

Taking the symmetric bound for $P[\overline{\pi(a)} \geq k + i]$ completes the proof.

Next we consider the error probability tolerance and we express it in terms of n as $\epsilon = n^{-a}$, where a is some positive constant. Also, set $e^{-\beta} = \varphi$. Lemma 3.5.1 can be directly applied to derive a bound for the displacement of the items that holds with high probability $(1 - n^{-a})$.

Proposition 3.5.1. *Let $a > 0$. Then for sufficiently large n,*

$$\mathrm{P}\left[|\overline{\pi(k)} - k| \geq \frac{a+2}{\beta \cdot r} \log n \text{ for some } k\right] < n^{-a}$$

The margin of error for each element is inversely proportionally to the sample complexity r. The above proposition guarantees that the average index estimator is pointwise close to the central ranking. Moreover, it can be proved (the proof is omitted) that the MLE solution π^m is close to the central ranking π^* with high probability.
We consider quantity $L = \max\left(6 \cdot \frac{a+2}{\beta \cdot r} \log n, 6 \cdot \frac{a+2+1/\beta}{\beta}\right)$, which is a measure of item displacement that appears in the following lemma.

Lemma 3.5.3. *For any optimal π^m the probability that there is some item k, such that $|\pi^m(k) - \pi^*(k)| > 32L$, is less than $2n^{-a}$.*

Combining the above results we get that with high probability the pointwise distance between the MLE solution π^m and the average index estimator $\bar{\pi}$ is less than $33L$.

This way the search space for the MLE solution is restricted to a zone around $\bar{\pi}$, where the pointwise distance from $\bar{\pi}$ is less than $k = 33L$. A brute force search would require time $k^{\vartheta(n)}$. Instead we use dynamic programming to reduce the running time.

Lemma 3.5.4. *Let $[n]$ be n elements together with a scoring function q. Suppose that we are given that there is an optimal ordering $\sigma(1), \sigma(2), \ldots, \sigma(n)$, that maximizes the score*

$$s(\sigma) = \sum_{\sigma(i) < \sigma(j)} q(i < j),$$

such that $|\sigma(i) - i| \leq k$ for all i. Then we can find such an optimal σ in time $O\left(n \cdot k^2 \cdot 2^{6k}\right)$.

In our setting $k = 33L$ is $O(\log n)$. When k is small ($o(\log n)$), the algorithm tends to linear.

We can use Lemma 3.5.4 to give an efficient algorithm that finds the maximum likelihood permutation π^m given $\pi_1, \ldots, \pi_r$. Recall that such a π^m minimizes

$$\sum_{k=1}^{r} d_K(\pi_k, \pi^m) = \sum_{k=1}^{r} \sum_{\pi^m(i)<\pi^m(j)} 1_{\pi_k(i)>\pi_k(j)} = \sum_{\pi^m(i)<\pi^m(j)} \#\{k : \pi_k(i) > \pi_k(j)\}.$$

Considering the score function $q(i<j) := \#\{k : \pi_k(i) < \pi_k(j)\}$ we have that minimizing the above cost is equivalent to maximizing

$$s(\pi^m) = \sum_{\pi^m(i)<\pi^m(j)} q(i<j).$$

and thus Lemma 3.5.4 can be employed for the MLE calculation. This leads us to the final theorem:

Theorem 3.5.3. *Let $\pi_1, \ldots, \pi_r$ be rankings on n elements independently generated by a Mallows model with spread parameter $\beta = \log(1/\phi)$, and let $a > 0$. Then a maximum probability order π^m can be computed in time*

$$T(n) = O\left(n^{1+O\left(\frac{a}{\beta r}\right)} \cdot 2^{O\left(\frac{a}{\beta}+\frac{1}{\beta^2}\right)} \cdot \log^2 n\right).$$

except with probability $< n^{-a}$.

Note that the algorithm tends to almost linear as r grows.

3.5.3 Spread Parameter Estimation

Busa-Fekete et al. in [3] study the sample complexity of the estimation of the Mallows spread parameter as well as the maximum likelihood estimation of the spread parameter. To tackle these problems they consider a more general model, the Generalized Kendall-Mallows model. In this model the KT-distance is decomposed into terms corresponding to single items, with the i-th term being equal to the number of discordant pairs that contain item e_i. Each item has its own spread parameter. The term corresponding to the i-th item is $V_i(\pi, \pi_0) = \sum_{j=0}^{i-1} \mathbb{1}\{(\pi(i) - \pi(j))(\pi_0(i) - \pi_0(j)) < 0\}$ and the total KT-distance is written as $d_{KT}(\pi, \pi_0) = \sum_{i=1}^{n} V_i(\pi, \pi_0)$ The pmf of the Generalized Mallows model is the following:

$$\mathbb{P}_{\pi\sim\mathcal{M}(\phi,\pi_0)}[\pi = \sigma] = \prod_{i=1}^{m} \frac{\phi_i^{V_i(\sigma,\pi_0)}}{Z_i(\phi_i)}$$

We see that the random variables Y_i are independent, since their joint probability distribution is written as the product of their pmfs. For the estimation of the spread parameters we focus on the marginals of the random variables $Y_i = V_i(\pi, \pi_0)$, that follow the truncated geometric distribution:

$$\mathbb{P}_{\pi\sim\mathcal{M}(\phi,\pi_0)}[Y_i = k_i] = \frac{\phi_i^{k_i}}{Z_i(\phi_i)}, k_i \in \{0, 1, \ldots i-1\}$$

The truncated geometric distribution is parameterised by its probability of failure ϕ_i and its truncation parameter $i-1$ and it is denoted as $\mathcal{TG}(\phi_i, i-1))$. It belongs to the exponential family of distributions, which is a very important class of distributions, with algebraic properties that help derive useful results. It includes many famous distributions such as the normal, exponential, gamma, chi-squared, beta, Dirichlet, Bernoulli, categorical, Poisson and the (truncated) geometric distribution.

We will briefly describe the general form of these distributions, before studying the special case of the truncated geometric distribution. The pdf (or pmf) of these distributions is written in the form $f_X(x|\eta) = h(x) \cdot exp[\eta \cdot T(x) - A(\eta)]$.
•T(x) is a sufficient statistic of the distribution. For exponential families, the sufficient statistic is a function of the data that holds all information the data x provides with regard to the unknown parameter values. This means that, for any data sets x and y, the likelihood ratio is the same: $\frac{f(x;\eta_1)}{f(x;\eta_2)} = \frac{f(y;\eta_1)}{f(y;\eta_2)}$, if $T(x) = T(y)$.
• η is called the natural parameter. The set of values of η for which the function $f_X(x;\eta)$ is finite is called the natural parameter space.
• $A(\eta)$ is called the log-partition function, because it is the logarithm of a normalization factor, without which $f_X(x;\eta)$ would not be a probability distribution: $A(\eta) = log\left(\int_X h(x) \cdot exp[\eta \cdot T(x)]dx\right)$. The function A is also important, because the mean, variance and other moments of the sufficient statistic T(x) can be derived simply by differentiating $A(\eta)$. For example, $\mathbb{E}[T(x)] = \nabla A(\eta)$ and $Var(T(x)) = \nabla^2 A(\eta)$

In the case of the truncated geometric distribution $\mathcal{TG}(\phi_i, i-1))$ we have:

$$p_{\eta_i}(x) = \exp\left(\eta_i T(x) - A(\eta_i)\right), x \in \{0, 1, \ldots i-1\}$$
$$\eta_i = \ln(\phi_i)$$
$$T(x) = x$$
$$A(\eta_i) = \ln(Z_i(e^{\eta_i}))$$

The Generalized Kendall-Mallows model also belongs to the exponential family:

$$p_\eta(\pi) = \exp\left(\partial^T T(x) - A(\eta)\right), \pi \in S_n$$
$$\eta = (\ln(\phi_1), \ldots, \ln(\phi_n))$$
$$T(\pi) = (V_1(\pi, \pi_0), \ldots, V_n(\pi, \pi_0))$$
$$A(\eta) = \sum_{i=1}^{n} A_i(\eta_i) = \sum_{i=1}^{n} \ln(Z_i(e^{\eta_i}))$$

The maximum likelihood estimation of the spread parameters of the (Generalized) Mallows Model is equivalent to the maximum likelihood estimation of the ϕ parameter of each truncated geometric derived by marginalising the Mallows Model. The solution to the MLE problem of $\mathcal{TG}(\hat{\phi}, i-1)$ given the truncation parameter $i-1$ and a collection of N iid samples $\mathbf{X} = [X_1, X_2, \ldots X_N]$ of the unknown distribution is described below. For simplicity, since i is known, we denote ϕ_i by ϕ and $Z_i(.)$ by $Z(.)$.

$$\hat{\phi} = \underset{\phi}{\operatorname{argmax}}\{L(\phi \mid \mathbf{X})\} = \underset{\phi}{\operatorname{argmax}} \left\{ \prod_{j=1}^{N} \frac{\phi^{X_j}}{Z(\phi)} \right\} = \underset{\phi}{\operatorname{argmax}} \left\{ \frac{\phi^{\sum_{j=1}^{N} X_i}}{Z(\phi)^N} \right\} =$$
$$= \underset{\phi}{\operatorname{argmax}} \left\{ \left(\sum_{j=1}^{N} X_i \right) \ln(\phi) - N \ln Z(\phi) \right\}$$

We set the derivative with respect to ϕ equal to zero:

$$\left(\sum_{j=1}^{N} X_j \right) \frac{1}{\hat{\phi}} - N \frac{Z'(\hat{\phi})}{Z(\hat{\phi})} = 0 \Leftrightarrow \hat{\phi} \frac{Z'(\hat{\phi})}{Z(\hat{\phi})} = \frac{1}{N} \left(\sum_{j=1}^{N} X_j \right)$$

Quantity $\hat{\phi}\frac{Z'(\hat{\phi})}{Z(\hat{\phi})}$ is equal to $\frac{d}{d\hat{\eta}}(A(\hat{\eta}))$, thus it is the expected value of $\mathcal{TG}(\hat{\phi}, i-1)$. We want to find a value for $\hat{\phi}$ (or equivalently for $\hat{\eta} = ln(\hat{\phi})$), such that the (theoretical) expected value of $\mathcal{TG}(\hat{\phi}, i-1)$ is equal to the empirical mean. The theoretical expected value is an increasing function of $\hat{\eta}$, since its derivative with respect to $\hat{\eta}$ is equal to $\frac{d^2}{d\hat{\eta}^2}(A(\hat{\eta})) = \underset{X \sim \mathcal{TG}(\hat{\phi}, i-1)}{\operatorname{Var}}(X) > 0$. The fact that it is increasing allows us to perform binary search and find an approximation of $\hat{\phi}$ in logarithmic steps with respect to the reciprocal of the absolute error.

We complete this section with two theorems on the sample complexity of the spread parameter estimation given in [3].

Theorem 3.5.4. *For any $\pi_0 \in S_n$, $\phi^* \in [0, 1-\gamma]$, $\epsilon, \delta > 0$, given $N = \Omega\left(\frac{log(1/\delta)}{n\epsilon^2} + \frac{log(n/\delta)}{\gamma}\right)$ iid samples from $\mathcal{M}_{\phi^*,\pi_0}$, we can compute in polynomial time estimates $\hat{\pi}$ and $\hat{\phi}$ such that:*

$$\underset{\Pi \sim \mathcal{M}^N_{\phi^*,\pi_0}}{\mathbb{P}} \left[(\hat{\pi}(\Pi) = \pi_0) \wedge \left(\left| \hat{\phi}(\Pi) - \phi^* \right| \leq \epsilon \right) \right] \geq 1 - \delta$$

If π_0 is known, then with $N = \Omega\left(\frac{log(1/\delta)}{n\epsilon^2}\right)$ we have:

$$\underset{\Pi \sim \mathcal{M}^N_{\phi^*,\pi_0}}{\mathbb{P}} \left[\left| \hat{\phi}(\Pi) - \phi^* \right| \leq \epsilon \right] \geq 1 - \delta$$

Theorem 3.5.5. *Given a single sample from Kendall-Mallows distribution $\mathcal{M}_{\phi^*,\pi_0}$ with known central ranking π_0 and unknown spread parameters ϕ^*, we can estimate $\hat{\phi}$ so that:*

$$\underset{\pi \sim \mathcal{P}_{\phi^*,\pi_0}}{\mathbb{P}} \left[\left| \hat{\phi}(\pi) - \phi^* \right| \leq O\left(\sqrt{\frac{log(1/\delta)}{n}} \right) \right] \geq 1 - \delta$$

Note that, as n goes to infinity, a single sample is enough for the estimation.

Chapter 4

Related Work

In this chapter we will present three papers that achieved breakthroughs in learning Mallows Mixtures. The presentation is organised in chronological order.

4.1 The Work of Awasthi, Blum et al.

This paper was the first to provide theoretical guarantees for the efficiency of learning the mallows mixture learning. In [23] Awasthi, Blum et al. worked on the case of two components. They used a method of moments, where the order k moment is defined as the vector that contains for all sets of k items the probabilities that these items are the top k. A rank 2 decomposition is possible in the third moment tensor. This tensor can be decomposed into two rank-1 terms, so that each term provides information for one of the base models. This information includes estimations of the mixing weights, the spread parameters and the prefixes of the centers. A first step is to construct an empirical estimation of the third moment of the mixture using the available samples. Then, tensor decomposition of the empirical third moment provides estimations of the weights, spread parameters and prefixes for each of the two base models.

Having those estimations we proceed to the second phase of the learning algorithm that uses this information to cluster the samples, assigning each sample to the correct base model that generated it. If this clustering is possible, then the task of obtaining the rest of the base permutations is an easy one, studied extensively in previous work. If the prefixes are not distinct enough to provide a pivot element, that is an element that is ranked in significantly higher positions in the one base ranking than the other, then decomposition is again required, in form of a linear system of equations that give the probabilities of assigning each of the items in each of the positions of each of the two centers. Then for each center and each item the most probable position is chosen as an estimation of its true position.

The proposed algorithm runs in polynomial time with respect to the parameters of the mixture and the accuracy parameter. It is worth mentioning that the proposed algorithm has practical value as well, as it outperforms the EM both in accuracy and speed, as shown in the experiments conducted in [23].

4.1.1 Notation and Important Properties

Tensors are a key tool in the techniques of this paper. Tensors are derived from a set of vectors. They are multidimensional objects. The number of dimensions is equal to the number of vectors combined. In this paper vectors of dimension at most 3 are used. In subsequent papers, that we will discuss later, higher dimensional tensors are used to learn the mixture of k Mallows models. Given two vectors $u \in \mathbb{R}^{n_1}$ and $v \in \mathbb{R}^{n_2}$. Tensor $u \otimes v \in \mathbb{R}^{n_1 \times n_2}$ is equal to uv^T. Given three vectors $u \in \mathbb{R}^{n_1}$, $v \in \mathbb{R}^{n_2}$ and $z \in \mathbb{R}^{n_3}$ tensor $u \otimes v \otimes z \in \mathbb{R}^{n_1 \times n_2 \times n_3}$ is a matrix P with $P_{ijk} = u_i \cdot v_j \cdot z_k$.

The notion of tensors is used to define the Moments of the Mallows Mixture Model. There is no obvious way to define the moments of a probabilistic ranking model. Awasthi, Blum et al. defined the first three moments of the Mallows Mixture Model on items $\{e_1, ...e_n\}$as follows:

- The first moment is a 1-tensor P such that $P_i = \mathbb{P}[pos(e_i) = 1]$. It contains the probabilities of ranking each element at the first position.
- The second moment is a 2-tensor P such that $P_{ij} = \mathbb{P}[\{pos(e_i), pos(e_j)\} = \{1, 2\}]$. It contains the probabilities of ranking each pair of element at the first two positions (in any order).
- The third moment is a 3-tensor P such that $P_{ijk} = \mathbb{P}[\{pos(e_i), pos(e_j), pos(e_k)\} = \{1, 2, 3\}]$. It contains the probabilities of ranking each triplet of element at the first three positions (in any order).

The first moment of a single Mallows $M(\phi, \pi)$ model is called representative vector of the model because it holds information that uniquely determine the parameters of the model. It can be proved that the formula for the i-th coordinate of this vector is equal to $\phi^{pos_\pi(e_i)-1}/Z_n$. Returning to the case of a mixture of two Mallows models $\mathcal{M}_1 = M(\phi_1, \pi_1)$ and $\mathcal{M}_2 = M(\phi_2, \pi_2)$ we denote x and y as their representative vectors (first moments).

The mathematical expressions for the first three moments of the mixture in terms of the representative vectors x and y of the base models are the following:

* First moment: $P_i = w_1 x_i + w_2 y_i$
* Second moment: $P_{ij} = w_1 \cdot c_2(\phi_1) \cdot x_i \cdot x_j + w_2 \cdot c_2(\phi_2) \cdot y_i \cdot y_j$, where $c_2(\phi) = \frac{Z(n,\phi)}{Z(n-1,\phi)} \frac{\phi+1}{\phi}$
* Third moment: $P_{ijk} = w_1 \cdot c_3(\phi_1) \cdot x_i \cdot x_j \cdot x_k + w_2 \cdot c_3(\phi_2) \cdot y_i \cdot y_j \cdot y_k$, where $c_3(\phi) = \frac{Z^2(n,\phi)}{Z(n-1,\phi)Z(n-2,\phi)} \frac{1+2\phi+2\phi^2+\phi^3}{\phi^3}$.

The third moment is non trivial only if the three coordinates i, j, k are all distinct. Thus a partition is made on the set of items into three groups S_a, S_b, S_c. We consider the third

moment of the mixture on this partition: $T^{abc} = (P_{ijk})_{i\in S_a, j\in S_b, k\in S_c}$.
Tensor T^{abc} has a rank-2 decomposition into two rank-1 terms, each one corresponding to a base model:

$$T^{abc} = w_1 \cdot c_3(\phi_1) \cdot x^{(a)} \otimes x^{(b)} \otimes x^{(c)} + w_2 \cdot c_3(\phi_2) \cdot y^{(a)} \otimes y^{(b)} \otimes y^{(c)}$$

where $x^{(a)}, x^{(b)}, x^{(c)}$ are the restrictions of representative vector x of base model $\mathcal{M}_1 = M(\phi_1, \pi_1)$ into subsets S_a, S_b, S_c of items. The same goes for y.

4.1.2 Algorithms

We present the main algorithm for learning Mixtures of two Mallows Models. The algorithm invokes several subroutines, which we present separately.

Learning Algorithm For Mixtures of two Mallows Models

Repeat $O(log(n))$ times:

* Make a random partition of [n], the full set of options into three subsets S_a, S_b, S_c.
* Compute $\hat{P}$, the empirical estimation of the third moment on the partition set.

This yields a tensor $T^{abc} = (\hat{P}_{ijk})_{i\in S_a, j\in S_b, k\in S_c}$.

* Perform TENSOR DECOMPOSITION to express T^{abc} as $u^{(a)}\otimes u^{(b)}\otimes u^{(c)} + v^{(a)}\otimes v^{(b)}\otimes v^{(c)}$

Next we apply a decomposition success criterion

Let $\sigma_2(A)$ denote the second largest singular value of matrix A.

If $min\{\sigma_2(u^{(a)}; v^{(a)}), \sigma_2(u^{(b)}; v^{(b)}), \sigma_2(u^{(c)}; v^{(c)})\} > \epsilon_2 = poly(\frac{1}{n}, \epsilon, \phi_{min}, w_{min})$

* Obtain parameter estimations for the weights, spread parameters and top k prefixes of the centers of the mixture by invoking routine INFER-TOP-K($\hat{P}, (u^{(a)}; v^{(a)}), (u^{(b)}; v^{(b)}), (u^{(c)}; v^{(c)})$).
* Invoke RECOVER-REST routine to reconstruct the rest of the centers.
* Return Success message and output the mixture parameter estimations.

Handle Degenerate cases

The guarantee for the successful execution of the main algorithm is stated in the next theorem.

Theorem 4.1.1. *Let $w_1\mathcal{M}(\phi_1, \pi_1) \oplus w_2\mathcal{M}(\phi_2, \pi_2)$ be a mixture of two Mallows models and let $w_{\min} = \min\{w_1, w_2\}$ and $\phi_{\max} = \max\{\phi_1, \phi_2\}$ and similarly $\phi_{\min} = \min\{\phi_1, \phi_2\}$. Denote $\epsilon_0 = \frac{w_{\min}^2(1-\phi_{\max})^{10}}{16n^{22}\phi_{\max}^2}$. Then, given any $0 < \epsilon < \epsilon_0$, suitably small $\epsilon_2 = poly\left(\frac{1}{n}, \epsilon, \phi_{\min}, w_{\min}\right)$ and $N = poly\left(n, \frac{1}{\min\{\epsilon,\epsilon_0\}}, \frac{1}{\phi_1(1-\phi_1)}, \frac{1}{\phi_2(1-\phi_2)}, \frac{1}{w_1}, \frac{1}{w_2}\right)$ i.i.d samples from the mixture model, Algorithm 1 recovers, in poly-time and with probability $\geq 1 - n^{-3}$, the model's parameters with w_1, w_2, ϕ_1, ϕ_2 recovered up to ϵ-accuracy.*

The authors use the algorithm of [48] for tensor decomposition. The algorithm works when the factor matrices M_a, M_b, M_c have polynomially bounded condition number (that is their second largest singular values $\sigma_2(\cdot)$ is lower bounded by a polynomial in the input parameters). If this condition is satisfied the tensor $T^{(abc)}$ has a unique rank-2 decomposition and the algorithm achieves to find it. The factor matrices are passed to the INFER-TOP-K procedure and this way the top few elements of both π_1 and π_2 are estimated correctly and we can also infer the parameters w' s and ϕ' s to good accuracy ϵ. If all log(n) random partition S_a, S_b, S_c fail to produce a tensor $T^{(abc)}$ with well-conditioned factor matrices, then we are in a special case and it can be shown that in this case, the scaling parameters $\phi_1 \approx \phi_2$ with high probability.

The second part of the algorithm is implemented in the RECOVER-REST procedure. It is based on the observation that the probability of an element e_i going to position j can be written as a weighted combination of the corresponding probabilities under π_1 and π_2. In addition, the reduced distribution obtained by conditioning on a particular element e_j going to position 1 is again a mixture of two Mallows models with the same parameters. Hence, by conditioning on a particular element which appears in the initial learned prefix, we get a system of linear equations. We use estimates so the linear system is correct up to some small error δ (inversely polynomial). Solving the system robustly we can infer good estimates for the probability of every other element e_i going to position j in both π_1 and π_2. This allows us to infer the entire rankings by choosing for each element the most probable position.

RECOVER-REST Procedure

Input: Sample Set S drawn from the latent mixture,
parameter estimations $\hat{w}_1, \hat{w}_2, \hat{\phi}_1, \hat{\phi}_2$ and prefixes $\hat{\pi}_1, \hat{\pi}_2$

∗Compute representative vectors (probabilities of appearing in the first position) $\hat{x}$ and $\hat{y}$ for elements in $\hat{\pi}_1$ and $\hat{\pi}_2$ respectively.
We consider the lengths r_1, r_2 of prefixes $\hat{\pi}_1$ and $\hat{\pi}_2$ respectively. Wlog suppose $r_1 \geq r_2$.
If there exists an element e_i, such that $pos_{\hat{\pi}_1}(e_i) > r_1$ and $pos_{\hat{\pi}_2}(e_i) > r_2/2$
(or in the symmetric case):

Let S_1 be the subset of samples having e_i ranked in the first position.

∗ Learn a single Mallows model on S_1 to complete the estimation $\hat{\pi}_1$.

∗ Use dynamic programming to find a complete estimation $\hat{\pi}_2$ for the other center.

∗ Return the complete estimations.

The above was the simple case. Now we handle the difficult one, where no such pivot e_i exists.

Let e_{i^*} be the first element in $\hat{\pi}_1$ having its probabilities of appearing in first place in π_1 and π_2 differ by at least ϵ. Let $\hat{w_1}' = \left(1 + \frac{\hat{w_2}\hat{y}(e_{i^*})}{\hat{w_1}\hat{x}(e_{i^*})}\right)^{-1}$, $\hat{w_2}' = 1 - \hat{w_1}'$ and S_1 be the subset of samples with e_{i^*} ranked at the first position.

For each e_i that does not appear in $\hat{\pi}_1$ nor $\hat{\pi}_2$ and any possible ranking position j:

∗ Use sample set S to estimate $\hat{f}(i \to j) = \mathcal{P}[e_i \textit{ goes to position } j]$ and
S_1 to estimate $\hat{f}(i \to j | e_{i^*} \to 1) = \mathcal{P}[e_i \textit{ goes to position } j \textit{ given that } e_{i^*} \textit{ goes to position } 1]$.

∗ Solve the system

$$\hat{f}(i \to j) = \hat{w_1} f^{(1)}(i \to j) + \hat{w_2} f^{(2)}(i \to j)$$
$$\hat{f}(i \to j | e_{i^*} \to 1) = \hat{w_1}' f^{(1)}(i \to j) + \hat{w_2}' f^{(2)}(i \to j)$$

This yields probabilities $f^{(1)}(i \to j)$ and $f^{(1)}(i \to j)$ of e_i going to position j in base models $\mathcal{M}_1$ and $\mathcal{M}_2$ respectively.

∗ Complete $\hat{\pi}_1$ by assigning each e_i to position $argmax_j\{f^{(1)}(i \to j)\}$.

∗ Complete $\hat{\pi}_2$ by assigning each e_i to position $argmax_j\{f^{(2)}(i \to j)\}$.

∗ Return $\hat{\pi}_1$, $\hat{\pi}_2$

INFER TOP-K Procedure

Input: $\hat{P}, M'_a = (u^{(a)}; v^{(a)}), M'_b = (u^{(b)}; v^{(b)}), M'_c = (u^{(c)}; v^{(c)})$.
$\hat{P}$ is the empirical estimation of the third moment.
$u^{(\tau)}$ is close to proportional (from tensor decomposition guarantees) to the restriction of the empirical first moment of base model $\mathcal{M}_1$ on partition S_τ, where $\tau \in \{a, b, c\}$.
The same holds for $v^{(\tau)}$ and $\mathcal{M}_2$.
Let $\hat{P}_\tau = \hat{P}(i \in \tau), \tau \in \{a, b, c\}$.
$*$ Set $(a_\tau, \beta_\tau)^T = M'^{\dagger}_\tau \hat{P}_\tau$ for all $\tau \in \{a, b, c\}$.
$*$ Set $\hat{w}_1 = \|a_a u^{(a)}\|_1 + \|a_b u^{(b)}\|_1 + \|a_c u^{(c)}\|_1$ and $\hat{w}_2 = 1 - \hat{w}_1$
Let $u = \left(\frac{a_a}{\hat{w}_1} u^{(a)}, \frac{a_b}{\hat{w}_1} u^{(b)}, \frac{a_c}{\hat{w}_1} u^{(c)}\right)$ and
$v = \left(\frac{\beta_a}{\hat{w}_2} v^{(a)}, \frac{\beta_b}{\hat{w}_2} v^{(b)}, \frac{\beta_c}{\hat{w}_2} v^{(c)}\right)$
$*$ Sort vectors u and v in decreasing order. Let $U = sort(u)$, $V = sort(v)$.
$*$ Set $\hat{\phi}_1 = \frac{U_2}{U_1}$ and $\hat{\phi}_2 = \frac{V_2}{V_1}$
Let $\gamma = \frac{(1-\hat{\phi}_{max})^2}{4n\hat{\phi}_{max}}$.
$*$ Set $r_1 = log_{1/\hat{\phi}_1}\left(\frac{n^{10}}{w^2_{min}\gamma^2}\right)$ and $r_2 = log_{1/\hat{\phi}_2}\left(\frac{n^{10}}{w^2_{min}\gamma^2}\right)$
$*$ Return prefixes $\hat{\pi}_1 = U[:r_1]$, $\hat{\pi}_2 = V[:r_2]$

4.2 The Work of Liu and Moitra

Liu and Moitra in [5] were the first to solve the problem in the general case of k centers. They leveraged the results of Zagier et al. in [4] and established the polynomial identifiability of the Mallows Mixture model. Firstly, they worked on the general setting $M = w_1 M\left(\phi_1, \pi^*_1\right) + \cdots + w_k M\left(\phi_k, \pi^*_k\right)$, with minimal assumptions (no components coincide with each other in TV distance and no component is completely uniform). They provide upper bounds for the sample complexity as well as information-theoretic lower bounds and lower bounds against restricted families of algorithms that make only local queries. Moreover they make a beyond worst case analysis of the Mallows Mixture learning problem.

On a technical level, they define distribution moments as groups of pairwise comparisons, define the block structure based on this notion of moments and prove that two models satisfying the same block structure do not differ much from each other. They provide upper bounds for the TV distance between an empirical model and its corresponding latent generative model and bounds that translate TV-distance closeness to parameter closeness. Their learning algorithm uses these results and constructs test functions to peel off one component at a time.

4.2.1 Block Structures and Tensors

First, we define three different kinds of structures, the block structure, the order structure and the ordered block structure.

Definition 4.2.1. *A block structure $\mathcal{B} = S_1, S_2, \cdots, S_j$ is an ordered collection of disjoint subsets of $[n]$. We say that a permutation π satisfies $\mathcal{B}$ as a block structure if for each i, the elements of S_i occur consecutively (i.e. in positions $a_i, a_i + 1, \ldots, a_i + |S_i| - 1$ for some a_i) in π and moreover the blocks occur in the order $S_1, S_2, \cdots, S_j$. Finally we let $\mathcal{S}_{\mathcal{B}}$ denote the set of permutations satisfying $\mathcal{B}$ as a block structure.*

Definition 4.2.2. *An order structure $\mathcal{O} = S_1, S_2, \cdots, S_j$ is a collection of ordered subsets of $[n]$. We say a permutation π satisfies $\mathcal{O}$ as an order structure if for each i, the elements of S_i occur in π in the same relative order as they do in S_i.*

Definition 4.2.3. *An ordered block structure $\mathcal{A} = S_1, S_2, \cdots, S_j$ is an ordered collection of ordered disjoint subsets of $[n]$. We say a permutation π satisfies $\mathcal{A}$ as an ordered block structure if it satisfies $S_1, S_2, \cdots, S_j$ both as a block structure and as an order structure.*

From the above definitions we see that we forget the order within each S_i when we treat the structure as a block structure and we forget the order among the S_i 's when we treat it as an order structure. For example, let $n = 7$ and consider $\mathcal{B} = (1,2),(4,5,6)$. The permutation $(1,2,3,7,6,5,4)$ satisfies $\mathcal{B}$ as a block structure. The permutation $(1,3,4,2,5,6,7)$ satisfies $\mathcal{B}$ as an order structure and the permutation $(1,2,3,4,5,6,7)$ satisfies $\mathcal{B}$ as an ordered block structure.

Next we define a special tensor that helps us express the conditional distribution on permutations that satisfy a given block structure.

Definition 4.2.4. *Given a Mallows model $M(\phi, \pi^*)$ and a block structure $\mathcal{B} = S_1, S_2, \cdots, S_j$, we define a $|S_1|! \times |S_2|! \times \cdots \times |S_j|!$ dimensional $T_{M,\mathcal{B}}$ as follows: Each entry corresponds to orderings $\pi_1, \pi_2, \cdots, \pi_j$ of $S_1, S_2, \cdots, S_j$ respectively and in it, we put the probability that a ranking drawn from M satisfies $\mathcal{B}$ and for each i, the elements in S_i occur in the order specified by π_i*

Tensor $T_{M,\mathcal{B}}$ has rank one and it can be written as the following product:
$T_{M,\mathcal{B}} = \Pr_M[\pi \in \mathcal{S}_{\mathcal{B}}] \cdot v\left(M\left(\phi, \pi|_{S_1}\right)\right) \otimes \cdots \otimes v\left(M\left(\phi, \pi|_{S_j}\right)\right)$, where $v(M(\phi, \pi^*))$ denotes the vectorisation of the Mallows distribution $M(\phi, \pi^*)$.

In the expression of $T_{M,\mathcal{B}}$ the factor $\Pr_M[\pi \in \mathcal{S}_{\mathcal{B}}]$ is the least convenient because it has no explicit formula. A convenient lower bound can be derived, considering the Mallows repeated insertion sampling process and the fact that $\frac{1}{1+\phi+\cdots+\phi^i} \geq \frac{1}{n}$. This yields that $\Pr_M[\pi \in \mathcal{S}_{\mathcal{B}}] \geq \frac{1}{n^{2j}}$. To better understand the definition of $T_{M,\mathcal{B}}$ we give a simple example. Consider $M(\phi, \pi)$ for $\pi = (1,2,3,5,4)$ and let $\mathcal{B} = (1,2),(4,5)$. Then $T_{M,\mathcal{B}} \sim \left(\frac{1}{1+\phi}, \frac{\phi}{1+\phi}\right) \otimes \left(\frac{\phi}{1+\phi}, \frac{1}{1+\phi}\right)$

4.2.2 Robust Linear Independence

The authors consider a determinant calculated in [4]. Let $A_n(\phi)$ be the $n! \times n!$ matrix whose rows and columns are indexed by permutations π, σ on $[n]$ and whose entries $A_{\pi\sigma}$ are $\phi^{d_{KT}(\pi,\sigma)}$. It holds that $\det(A_n(\phi)) = \prod_{i=1}^{n-1} \left(1 - \phi^{i^2+i}\right)^{\frac{n!(n-i)}{i^2+i}}$. This result is important because it implies that all rows c_i of this matrix (that is all vectorisations of Mallows models with spread parameter ϕ defined on [n]) are linearly independent. Thus, if two mixtures with spread parameter ϕ have the same pmf, then their centers and weights are identical (up to a relabeling). This is guaranteed because the determinant is non zero. Knowing the exact formula of the determinant we can prove "robust" linear independence, by bounding the ℓ^1 norm of any linear combination of distinct rows $\|z_1 c_1 + \cdots + z_k c_k\|_1$ from below.

A first step is using the determinant evaluation to bound the projection of any column onto the orthogonal complement of the span of any $k-1$ other columns.

Lemma 4.2.1. *Suppose $\phi < 1 - \epsilon$ and consider k columns of $A_n(\phi)$. The projection of one column onto the orthogonal complement of the other $k-1$ has euclidean length at least $\left(\frac{\epsilon^n}{\sqrt{n!}}\right)^k$.*

This identity is useful, but it is not strong enough to establish polynomial identifiability, as there is an exponential dependence on n. Apart from that, the rows in $A_n(\phi)$ correspond to non-normalised Mallows mass functions. To overcome the exponential dependence on n, the authors make use of block structures and choose the sets appropriately so that their total length depends on k rather than n. They prove the following lemma:

Lemma 4.2.2. *Let $B_n(\phi)$ be obtained from $A_n(\phi)$ by normalizing its columns to sum to one. Suppose $\phi < 1 - \epsilon$ and consider any k columns $c_1, c_2, \ldots, c_k$ of $B_n(\phi)$. Then*

$$\|z_1 c_1 + \cdots + z_k c_k\|_1 \geq \frac{1}{n^{4k}} \frac{\epsilon^{2k^2}}{(k+1)^{k^2+2k}}$$

provided that $\max(|z_1|, |z_2|, \ldots, |z_k|) \geq 1$.

The results for the case of equal spread parameters are generalised to the case of spread non equal but close to each other. The general case, where spread parameters might be very different from each other, is reduced to the case of similar spread parameters, by using test functions to peel off components with $|\phi_1 - \phi_i|$ non negligible. The remaining ones are all close to each other. The final identifiability result is the following:

Lemma 4.2.3. *Consider any k (not necessarily distinct) permutations $\pi_1, \pi_2, \cdots, \pi_k$ and scaling parameters $\phi_1, \phi_2, \cdots, \phi_k$. Set $M_i = M(\phi_i, \pi_i)$ and suppose that the collection of Mallows models is μ-non degenerate (that is $\forall i,j \in [k]$ $i \neq j \Rightarrow TV(M_i, M_j) > \mu$ and $\forall i \in [k] TV(M_i, Uniform) > \mu$). Then for any coefficients z_i with $\max(|z_1|, |z_2|, \ldots, |z_k|) \geq 1$ we have*

$$\|z_1 v(M_1) + \cdots + z_k v(M_k)\|_1 \geq \left(\frac{\mu^2}{10n^4 k}\right)^{20k^3}$$

In the above lemma we should think of real coefficients z_is as weight differences $w_i - w_i'$. This way, the bounded quantity is equal to the TV distance between two Mallows Mixture models with different parameters.

4.2.3 Test Functions and Learning Algorithm

The learning algorithm performs a brute force search over the spread parameters and mixing weights. It also considers different test functions with the goal to find appropriate test functions that isolate some component. It suffices to look at $O(k)$ positions to distinguish k permutations (we will see this in more detail in the next chapter, where we discuss the work of Mao et al.). Thus, to form candidate test functions we exhaustively consider groups of $O(k)$ items. The number of different test functions is $poly_k(n)$. The algorithm first peels off components with small spread parameters, based on the intuition that these components frequently generate their central permutation. Then, the algorithm tries to isolate a single component, peel it off and continue iteratively.

Suppose for example that we have a mixture of three Mallows Models. The components of the mixture are the following:
$\phi_1 < \phi_2 < \phi_3$, $\pi_1 = (1,2,3,4,5)$, $\pi_2 = (2,4,1,3,5)$, $\pi_3 = (5,1,2,3,4)$.
We consider the block structure $\mathcal{B} = \{\{1,2\},\{3,4\}\}$.
It holds that

$$T_{M_2,\mathcal{B}} \sim \left(\frac{\phi_2}{1+\phi_2}, \frac{1}{1+\phi_2}\right) \otimes \left(\frac{\phi_2}{1+\phi_2}, \frac{1}{1+\phi_2}\right)$$

$$T_{M_3,\mathcal{B}} \sim \left(\frac{1}{1+\phi_3}, \frac{\phi_3}{1+\phi_3}\right) \otimes \left(\frac{1}{1+\phi_3}, \frac{\phi_3}{1+\phi_3}\right)$$

We construct the test function

$$X = \left(\frac{1}{1+\phi_2}, \frac{-\phi_2}{1+\phi_2}\right) \otimes \left(\frac{\phi_3}{1+\phi_3}, \frac{-1}{1+\phi_3}\right)$$

We have that $\langle X, T_{M_2,\mathcal{B}}\rangle = 0$, $\langle X, T_{M_3,\mathcal{B}}\rangle = 0$ *but* $\langle X, T_{M_1,\mathcal{B}}\rangle \neq 0$. This way the test function isolates information from the first component.
The method used in the example is generalised to mixtures of k Mallows models.

4.2.4 Lower Bounds and Beyond Worst Case Analysis

Fist the authors prove that any algorithm for learning the components of a mixture of k Mallows models within μ in total variation distance must take at least $(1/m\mu)^{2k-1}$ samples.

Lemma 4.2.4. *For any $\mu \leq \frac{1}{40k^2}$ and $n \geq 40k^2$ there are two mixture of at most k Mallows models M and M' with the following properties:*

1. *Each mixture is $\left(\mu, \frac{1}{10\cdot 2^{2k}}\right)$-non degenerate*
2. *$d_{TV}(M, M') \leq 4(8\mu k)^{2k-1}$*

3. *M and M′ are not component-wise μ-close*

The proof considers a concentric mixture with spread parameters constructed by an arithmetic sequence.

Then, the authors study a restricted model of learning, the local query model and bound its cost.

Definition 4.2.5. *In the local query model, the learner queries a subset of elements $x_1, x_2, \cdots, x_c$ and locations $i_1, i_2, \cdots, i_c$ with a tolerance parameter τ and is answered with the probability, up to an additive τ, that a sample from the mixture has x_j in position i_j for all $1 \leq j \leq c$. The cost of the query is $\frac{1}{\tau^2}$ and the total cost of an algorithm is the sum of its query costs.*

Summarizing the results into an informal theorem we have that any algorithm for learning a mixture of k Mallows models through local queries must incur cost at least $n^{\log k}$.

Finally we present the beyond worst case results. Essentially, the only assumption is that spread parameters differ from each other, so the analysis and algorithm design provide an alternative expression for the complexity, in terms of the minimal spread parameter difference and the minimal distance of a spread parameter from 1.

Theorem 4.2.1. *Given samples from a mixture of k Mallows models with all spread parameters γ separated from each other and from 1 and $n \geq 10k$, there is an algorithm whose running time and sample complexity are*

$$poly(1/\gamma^{k^2}, 1/\partial^{k^2}, 1/w_m in^{k^2}) \cdot poly(n, \log(1/\delta))$$

for learning each center π_i exactly and the mixing weights and spread parameters to within an additive ∂, with probability at least $1 - \delta$.

The algorithm, similarly to [23], first tries to estimate the prefixes and spread parameters. Then, it recovers the rest of the centers and finally it estimates the mixing weights, conditioning on the event that centers have been recovered correctly. In this case, similarly to the general one, candidate parameters from a polynomial size list are tested. The spread parameters and weights candidates are produced with simple gridsearch. The prefix candidates are taken from the observations using a frequency threshold (we examine the most common prefixes) and prefix length is set to $10k$. The prefixes are used as signatures of the corresponding centers. Then, the center reconstruction process is similar to that in [6] and a detailed implementation of this process can be found in Algorithm for learning the Mallows mixture performing noiseless queries. Candidate models are tested with the tensor test function criteria we described earlier.

4.3 The Work of Mao et al.

Mao et al. in [6] study the problem of learning mixtures of Mallows models. The general case of k centers is considered as in the paper of Moitra et al, that was published two

years earlier, in 2018. However in contrast to the paper of Moitra et al. in this paper the spread parameters of the mixed models are assumed to be equal and known. Most techniques used are similar to those of Moitra et al. but some improvements are made. In particular, the sample complexity depends logarithmically on the number of items n while in previous work it scaled polynomially on n. Another contribution was proving an optimal dependency of the sample complexity on ϕ, the scaling parameter of the models, in the high noise regime.

4.3.1 Noiseless Oracles And Noiseless Learning Algorithms

In this paper, similar to previous work both on Mallows and Gaussian Mixtures, moments of the latent distribution are considered. The authors define the moments of order m as groups of m pairwise comparisons, that are simultaneously submitted to the model via some oracle. "Simultaneously" means that all answers correspond to the same center, the one that is activated (randomly following Multinoulli on the mixing weights). There are two kinds of oracles, the "strong" and the "weak". Both oracles are noiseless, that is they are not empirical depending on a collection of samples. Instead, they provide accurate information about the restriction of the model on a group of pairwise comparisons. The weak oracle reveals the restriction of each distinct central permutation on the group of pairwise comparisons and the strong oracle returns the distribution of the group of pairwise comparisons, viewed as a random vector.We will now give the formal definitions of the two oracles.

Definition 4.3.1. *Consider a distribution* M *on* $\mathcal{S}_n$ *and a random permutation* $\pi \sim \mathrm{M}$. *For* $m \in \mathbb{N}$, *let* $\mathcal{I}$ *be the tuple of* m *pairs of distinct indices* $(i_1, j_1), \ldots, (i_m, j_m) \in [n]^2$. *Upon a query on* $\mathcal{I}$, *the (strong) oracle of group of* m *pairwise comparisons returns the distribution of the random vector* $\chi(\pi, \mathcal{I})$ *in* $\{0,1\}^m$, *whose* r*th coordinate is defined by*

$$\chi(\pi, \mathcal{I})_r := 1\{\pi(i_r) < \pi(j_r)\} \quad \text{for } r \in [m]$$

Definition 4.3.2. *Consider a set* $\{\pi_1, \ldots, \pi_k\}$ *of* k *permutations in* $\mathcal{S}_n$. *For* $m \in \mathbb{N}$, *let* $\mathcal{I}$ *be a tuple of* m *pairs of distinct indices in* $[n]$. *Upon a query on* $\mathcal{I}$, *the weak oracle of group of* m *pairwise comparisons returns the set of binary vectors* $\{\chi(\pi_i, \mathcal{I}) : i \in [k]\}$, *where*

$$\chi(\pi_i, \mathcal{I})_r := 1\{\pi_i(i_r) < \pi_i(j_r)\} \quad \text{for } r \in [m]$$

We will now demonstrate why groups of m pairwise comparisons are similar to the order-m moments of distributions defined on vectors of real numbers. Let $X^{\pi}_{i,j} := \mathbb{1}\{\pi(i) < \pi(j)\}$ be a pairwise comparison. Also let

$$f_{\chi(\pi,\mathcal{I})}(v) := \mathbb{P}\{\chi(\pi, \mathcal{I}) = v\} \quad \text{for each } v \in \{0,1\}^m$$

This is the distribution returned by the strong oracle. This pmf has similar form with the

classical moments that we define for example in Gaussians.

$$f_{\chi(\pi,I)}(v) = \mathbb{E}[\mathbb{1}\{\chi(\pi,\mathcal{I}) = v\}] = \mathbb{E}\left[\prod_{r=1}^{m} \mathbb{1}\left\{X^{\pi}_{i_r,j_r} = v_r\right\}\right]$$
$$= \mathbb{E}\left[\prod_{r=1}^{m} \left(X^{\pi}_{i_r,j_r}\right)^{v_r}\left(1 - X^{\pi}_{i_r,j_r}\right)^{1-v_r}\right] = \mathfrak{m}(\pi,\mathcal{I})_v.$$

Thus, the strong oracle in fact returns moments of the latent distribution and a learning algorithm that uses this oracle can be viewed as a combinatorial method of moments. Another notion of moment for ranking distributions are the marginals on subsets of items (referred to as l-wise comparisons). In this work we consider two different ways of marginalising a ranking $\pi \in S_n$ an a subset J of $[n]$. Firstly, let $\pi|J$ denote the restriction of π on J, which is an injection from J to $[n]$. This marginalisation keeps the information about the positions of selected items in the complete ranking. Moreover, let $\pi\|J$ denote the bijection from J to $[|J|]$ induced by $\pi|J$. This marginalisation is equivalent to an object in $S_{|J|}$, which is achieved by reindexing the selecting items, assigning them ids in $|J|$. The second way of marginalising is less informative but more natural and corresponds to the selection mechanism that we gave on selective Mallows Models. For example, in the case of $\pi = (3,2,4,6,1,5)$ and $J = \{1,4,5\}$ we have for the injection $\pi|J(1) = 5$, $\pi|J(4) = 3$ and $\pi|J(5) = 6$ and for the bijection $\pi\|J = (4,1,5)$.
The weak and the strong oracle can be defined on l-wise comparisons as well and are more informative than oracles on groups of $l/2$ comparisons.Below we give the formal definitions of these oracles.

Definition 4.3.3. *Consider a distribution* M *on* $\mathcal{S}_n$ *and a random permutation* $\pi \sim$ M. *For* $\ell \in \mathbb{N}$, *let* J *be a subset of* $[n]$ *of cardinality* $|J| = \ell$. *Upon a query on* J, *the (strong) oracle of* ℓ*-wise comparison returns the distribution of the relative order* $\pi\|_J$.

Definition 4.3.4. *Consider a set* $\{\pi_1, \ldots, \pi_k\}$ *of* k *permutations in* $\mathcal{S}_n$. *For* $\ell \in \mathbb{N}$, *let* J *be a subset of* $[n]$ *of cardinality* $|J| = \ell$. *Upon a query on* J, *the weak oracle of* ℓ*-wise comparison returns the set of relative orders* $\{\pi_i\|_J : i \in [k]\}$.

The noiseless oracles we defined above are used by learning algorithms that aim to learn the parameters of the latent Mallows Mixture. The authors aim to minimise the order of the moments they use, that is they try to use small selection sets. In their setting this reduces the complexity but in our selective setting it also allows learning in the regime of strict selectivity.
• The algorithm that uses the weak oracle has the advantage of being independent of the estimation of the mixing weights and the spread parameters. However, it requires bigger selection sets (higher order moments). It uses moments of order k, where k is the number of central permutations. The optimality in terms of query length has not been proved.
• The algorithm that uses the strong oracle depends on the knowledge of the mixing weights and the assumption of common spread parameters. However, once it has this extra information it can perform more effective queries which require smaller selection sets of size logarithmic on k, where k is the number of central permutations. The optimality

in terms of query length is proved and it matches the bounds of identifiability.

Theorem 4.3.1. *(Learning algorithm using minimal length queries to the strong oracle)*
Let $m_k^* := \lfloor \log_2 k \rfloor + 1$.
For any mixture $\mathrm{M} = \sum_{i=1}^{k} w_i \delta_{\pi_i}$ *of permutations in* $\mathcal{S}_n$*, there is a poly* (n, k)*-time algorithm that recovers* M *from groups of* m_k^* *pairwise comparisons, with at most* $1 + \frac{k}{2}(n-2)(n+1)$ *adaptive queries to the weak oracle.*

Conversely, for $n \geq 2m_k^*$ *and* $\ell \leq 2m_k^* - 1$*, there exist distinct mixtures* $M = \frac{1}{k}\sum_{i=1}^{k} \delta_{\pi_i}$ *and* $\mathrm{M}' = \frac{1}{k}\sum_{i=1}^{k} \delta_{\pi_i'}$ *of permutations in* $\mathcal{S}_n$*, which cannot be distinguished even if all* $\binom{n}{\ell}$ *ℓ-wise comparisons are queried from the strong oracle.*

Since the oracle of ℓ-wise comparison is stronger than the oracle of group of $\ell/2$ pairwise comparisons, the above theorem implies that the oracle of ℓ-wise comparison is sufficient for identifying the k mixture if and only if $\ell \geq 2m_k^*$. We will prove this later more formally. Next we present the results on learning from the weak oracle.

Theorem 4.3.2. *(Learning algorithm using queries to the weak oracle)*
Consider a set $\{\pi_1, \ldots, \pi_k\}$ *of* k *permutations in* $\mathcal{S}_n$*. There is a poly* (n, k)*-time algorithm that learns the set* $\{\pi_1, \ldots, \pi_k\}$ *from queries on groups of* $k+1$ *pairwise comparisons to the weak oracle, using at most* $1 + \frac{k}{2}(n-2)(n+3)$ *adaptive queries.*

As for the implementation, both algorithms try to build the central permutations inductively on the number n of items. Each query consists of a "signature" set of pairwise comparisons, that aims to isolate a particular center, and one or two more comparisons that aim to detect the position of $n-th$ item in the isolated central ranking.

In the weak oracle a single signature set has to be able to isolate each distinct marginalised center. The signature contains pairwise comparisons that create a decision tree and each leaf of the tree corresponds to a single distinct center. The height of this tree is at most k, so the length of the signature is at most k pairwise comparisons.

In the strong oracle weights are also returned, apart from marginalised centers. Because of this extra information the signature set does not have to isolate each and every center but it suffices to isolate only one of them at a time. In this case the signature set length is logarithmic on k.

We present the details of the algorithm that learns from the strong oracle in the next chapter.

4.3.2 Using Noisy Samples to Simulate Noiseless Oracles-The Subroutine

Now we will see what happens when noise is introduced. Noisy samples are collected and they are used by the "subroutine" that simulates the noiseless oracle. In each query samples are projected on the queried subset of items and aggregated into an empirical marginal distribution. A cover is made on the space of all possible marginals and the one closest to the empirical one is selected. The parameters of the selected marginal provide the information needed by the oracle.

Given i.i.d. observations $\sigma_1, \ldots, \sigma_N$ from $\mathcal{M} = \sum_{i=1}^{k} w_i M(\pi_i)$, the goal of SubOrder($J$) is to learn the set of relative orders $\pi_1 \|_J, \ldots, \pi_k \|_J$ for a given subset $J \subset [n]$.
To study the SubOrder we have to define the marginalization of the Mallows mixture, as well as the observations, as follows. Note that the authors use the injective version of marginalisation in the models and samples, that is they require that information about the position in the complete samples is preserved. For any distribution $\mathcal{M}$ on $\mathcal{S}_n$ and a set of indices $J \subset [n]$, we let $\mathcal{M}|_J$ denote the marginal distribution of $\sigma|_J$ where $\sigma \sim \mathcal{M}$. That is, the PMF of $\mathcal{M}|_J$ is given by

$$f_{\mathcal{M}|_J}(\rho) = \mathbb{P}_{\sigma \sim \mathcal{M}} \{ \sigma|_J = \rho \}$$

Moreover, given N i.i.d. observations $\sigma_1, \ldots, \sigma_N$ from $\mathcal{M}$, the empirical version of $\mathcal{M}|_J$ is given by

$$f_{\mathcal{M}_N|_J}(\rho) = \frac{1}{N} \sum_{m=1}^{N} \mathbb{1} \{ \sigma_m|_J = \rho \} .$$

In the bijective definition of marginalisation, classical identifiability results guarantee that if the central permutations (and weights) are equal as sets in two mixtures, then the mixture have the same pmf on the projection set (because they form a Mallows Mixture distribution on this set). However, the distribution on injective marginals is not a Mallows Mixture, so classical identifiability results are not applicable. The following lemma guarantees that identribiability holds in the injective marginalisation similarly to the bijective one.

Lemma 4.3.1. *For any subset $J \subset [n]$, if the central permutations $\pi, \pi' \in \mathcal{S}_n$ satisfy $\pi|_J = \pi'|_J$, then the marginalized Mallows models $M(\pi, \phi)|_J$ and $M(\pi', \phi)|_J$ coincide for all $\phi \in (0, 1)$.*

Next, the authors provide a guarantee about the identifiability of the marginalised Mallows Mixture with respect to the central rankings. The result guarantees that two marginal Mallows Mixtures can not be too close in TV distance if the corresponding sets of marginalised central rankings are not equal.

Proposition 4.3.1. *Consider Mallows mixtures $\mathcal{M} = \sum_{i=1}^{k} w_i M(\pi_i)$ and $\mathcal{M}' = \sum_{i=1}^{k} w'_i M(\pi'_i)$ on $\mathcal{S}_n$ with a common noise parameter $\phi \in (0, 1)$. Let $\gamma := \min_{i \in [k]} (w_i \wedge w'_i) > 0$. Fix a set of indices $J \subset [n]$ and let $\ell := |J|$. Suppose that the two sets of central permutations $\{\pi_1 \|_J, \ldots, \pi_k \|_J\}$ and $\left\{\pi'_1 \right\|_J, \ldots, \left. \pi'_k \right\|_J\}$ are not equal (as sets). Then*

$$\mathrm{TV}\left(\mathcal{M}|_J\,,\,\mathcal{M}'\big|_J\right) \geq \eta(k,\ell,\varphi,\gamma) := \left(\frac{\gamma}{6k}\right)^{(3\ell)^{\ell+1}} \left(\frac{1-\varphi}{\ell}\right)^{(4\ell)^{\ell}+2k\ell^2} \tag{4.1}$$

The proof uses the notion of the "block structure" and bounds the probability of different models satisfying the same block structure.

The final tool needed to analyse the "Subroutine" is the following proposition, that bounds the TV-distance between the latent marginalised Mallows mixture model and its empirical version constructed by N iid samples. The upper bound decreases exponentially on N and will be used to derive a polynomial sample complexity sufficient to approximate the PMF with an empirical histogram.

Proposition 4.3.2. *For $J \subset [n]$, let $\mathcal{M}|_J$ and $\mathcal{M}_N|_J$ be the marginalized Mallows mixture and the marginalized empirical distribution. Then for any $s \in (0,1)$, we have:*

$$\mathbb{P}\{\mathrm{TV}\,(\mathcal{M}|_J\,,\,\mathcal{M}_N|_J) > s\} \leq \exp\left(-N\frac{3s}{10}\right) + 2(2kq)^{\ell}\exp\left(-N\frac{s^2}{(2kq)^{2\ell}}\right)$$

where $\ell := |J|$ and $q := 1 + \frac{1}{1-\varphi}\log\frac{8\ell}{s(1-\varphi)}$.

The proof is based on the observation that the TV-distance is defined on a domain of size n^{ℓ}, which can be divided into two parts. The first part contains rankings that are close to all central permutations. Samples in this set have a relatively high probability of appearance, but their cardinality is small, due to the constraint of being close to all central rankings. The other (complementary) part of the domain has a relatively big cardinality but its elements have small probabilities of appearance. In both cases, the empirical frequency of a sample is connected with the theoretical one with strong, exponentially decreasing bounds (applying the Hoeffding and Bernstein inequalities).

An important identity of the TV-distance bounds provided in the last two propositions is that the bounds do not depend on the total number n of items, only on ℓ, the number of selected items.

The above propositions are used to develop the SubOrder(.) function that simulates the weak oracle. SubOrder performs a brute-force search over candidate marginals and outputs the parameters of the one that better fits the available samples in terms of TV-distance. The brute-force is made on all possible marginalised central ranking combinations and on the corresponding mixing weights. The weights are continuous parameters so a discretization is performed with a grid step equal to $1/L$. However, the weak oracle only expects the marginalised central rankings, so in this step estimated weights are not returned.

A key step in the search procedure is that candidate models are not directly compared to the empirical on the sample set in terms of TV-distance. This is due to the fact that the marginalized distribution $\mathcal{M}'|_J$ does not have an explicit formula (it is defined on injections rather than bijections). To overcome this problem, fake samples are generated from

each candidate model (in $O(n^2)$ time per sample using RIMf) and the empirical distribution of these samples is compared to the empirical of the real samples in TV-distance. Due to the triangle inequality, if the two empiricals are close, then the candidate model is close to the latent model, given that the two models are close to their empirical versions as stated in Proposition 4.3.2. The closeness in TV-distance between the optimal candidate model and the latent model means that their sets of central permutations are equal due to Proposition 4.3.1. Thus, the central permutations of the latent model are correctly estimated with high probability, given enough samples and the weak oracle is successfully simulated. We will now give the formal definition of SubOrder.

We define a set of polynomially many candidate models ($poly_k(n)$). Let $\mathcal{S}_{n,J}$ denote the set of injections $\rho : J \to [n]$, which has cardinality at most n^ℓ where $\ell = |J|$. For each $\rho \in \mathcal{S}_{n,J}$, fix an arbitrary permutation π_ρ in $\mathcal{S}_n$ such that $\pi_\rho|_J = \rho$. Let L be a positive integer to be determined later. For $\phi \in (0,1)$ and $\gamma \in (0, 1/k]$, we define a set of Mallows mixtures by discretizing the weights:

$$\mathscr{M} \equiv \mathscr{M}(n,k,\phi,\gamma,J,L) := \left\{ \sum_{i=1}^{k} \frac{r_i}{L} M\left(\pi_{\rho_i}, \phi\right) : \rho_i \in \mathcal{S}_{n,J}, r_i \in [L], r_i \geq \gamma L, \sum_{i=1}^{k} r_i = L \right\}. \quad (4.2)$$

Parameter γ corresponds to the minimal mixing weight and $1/L$ to the step of the grid search over mixing weights. The weights r_i/L sum to 1 and each weight is at least γ. Since there are at most L choices for each weight and at most $\left|\mathcal{S}_{n,J}\right| \leq n^\ell$ choices for each ρ_i, we have $|\mathscr{M}| \leq L^k n^{k\ell}$. We remind that ℓ is $O(k)$.

SubOrder Function

Input: observations $\sigma_1, \ldots, \sigma_N \in \mathcal{S}_n$, a subset $J \subset [n]$, $\ell := |J|$, and parameters $k \in \mathbb{N}$, $\phi \in (0,1)$, $\gamma \in (0, 1/k]$, $N' \in \mathbb{N}$, and $L = \lceil 3k/\eta \rceil$ where $\eta = \eta(k, \ell, \phi, \gamma)$ is defined in 4.1

* For each Mallows mixture $\mathcal{M}' \in \mathscr{M}$, where $\mathscr{M}$ is defined in 4.2, generate N' i.i.d. random permutations $\sigma'_1, \ldots, \sigma'_{N'}$ from $\mathcal{M}'$. Compute the marginalized empirical distribution $\mathcal{M}'_{N'}\big|_J = \frac{1}{N'} \sum_{m=1}^{N'} \delta_{\sigma'_m|_J}$.

* If for some $\mathcal{M}' = \sum_{i=1}^{k} \frac{r_i}{L} M\left(\pi_{\rho_i}, \phi\right) \in \mathscr{M}$ it holds that $\mathrm{TV}\left(\mathcal{M}'_{N'}\big|_J, \mathcal{M}_N|_J\right) \leq \eta/2$, return the set of relative orders $\left\{\pi_{\rho_i}\|_J : i \in [k]\right\}$. If there are multiple models $\mathcal{M}'$ in $\mathscr{M}$ satisfying the condition, an arbitrary $\mathcal{M}'$ is chosen. If no models in $\mathcal{M}$ satisfy this condition, then return "error".

The following theorem states that a polynomial sample complexity suffices to guarantee that *SubOrder*(J) successfully simulates the weak oracle on query set J with high probability.

Theorem 4.3.3. *Suppose we are given N i.i.d. observations $\sigma_1, \ldots, \sigma_N$ from the Mallows mixture $\mathcal{M} = \sum_{i=1}^{k} w_i M(\pi_i)$ on $\mathcal{S}_n$ with a noise parameter $\phi \in (0,1)$. Fix a set of indices $J \subset [n]$ and let $\ell := |J|$. Fix $\gamma > 0$ such that $\gamma \leq \min_{i\in[k]} w_i$. Fix a probability of error $\delta \in (0,1)$. If the sample size satisfies $N \geq \text{poly}_{k,\ell}\left(\frac{1}{1-\phi}, \frac{1}{\gamma}\right) \cdot \log \frac{1}{\delta}$ and we choose an integer $N' \geq \text{poly}_{k,\ell}\left(\frac{1}{1-\phi}, \frac{1}{\gamma}\right) \cdot \log \frac{n}{\delta}$, then* SubOrder$(J)$ *returns the set of relative orders $\{\pi_i\|_J : i \in [k]\}$ with probability at least $1-\delta$.*

To analyse the time complexity of SubOrder we first observe that the set of candidate models is of polynomial size and the sample complexity of both original and fake samples is polynomial as well. Moreover, sampling from each candidate model is performed efficiently in polynomial time (e.g. using the RIM sampling method). Calculating the TV-distance between the empirical models can be performed in time linear to the number of samples. Thus, the time complexity of SubOrder is polynomial on the spread parameter, minimal mixing weight, number of items and error probability tolerance. The complexity is exponential on the number k of central rankings but we assume that this parameter is constant.

4.3.3 Recovering the Central Rankings and the Corresponding Weights

To recover the central rankings of a latent mixture, using noisy samples drawn from the mixture, we could use the algorithm presented in 4.3.2, simulating the weak oracle with the SubOrder. The time complexity of the noiseless algorithm is $O(n^2 \cdot k)$ and the time complexity of each SubOrder call is polynomial on all parameters except k. SubOrder(J) will be called on $O(n^2 \cdot k)$ sets J, where $|J|$ is at most $2k+2$. Since there are less than n^{2k+2} possible subsets of $[n]$ that have cardinality $2k+2$, we can set $\delta = n^{-2k-12}$ in Theorem 4.3.3 and take a union bound to ensure that with high probability (n^-10) all SubOrder calls will be successful and thus the central permutations will be exactly recovered.This yields the following result:

Theorem 4.3.4. *Given N i.i.d. observations from the Mallows mixture $\mathcal{M} = \sum_{i=1}^{k} w_i M(\pi_i)$ on $\mathcal{S}_n$ with a known noise parameter $\phi \in (0,1)$. Suppose we are given $\gamma > 0$ such that $\gamma \leq \min_{i\in[k]} w_i$. Then there exists a poly $_k\left(n, \frac{1}{1-\delta}, \frac{1}{\gamma}\right)$-time algorithm that exactly recovers the set of central permutations $\{\pi_1, \ldots, \pi_k\}$ with probability at least $1 - n^{-10}$, provided that $N \geq \text{poly}_k\left(\frac{1}{1-\phi}, \frac{1}{\gamma}\right) \cdot \log n$.*

Having recovered the central rankings of the mixture we will try to approximate the corresponding weights with small absolute error. The tool that will be used to this end is a proposition that bounds from below the TV-distance between two marginalised mixtures that have the same centers but different corresponding weights.

Proposition 4.3.3. *Consider Mallows mixtures $\mathcal{M} = \sum_{i=1}^{k} w_i M(\pi_i)$ and $\mathcal{M}' = \sum_{i=1}^{k} w'_i M(\pi_i)$ on $\mathcal{S}_n$ with a common noise parameter $\phi \in (0,1)$. Suppose that $\xi \triangleq \max_{i\in[k]} \left|w_i - w'_i\right| > 0$.*

Let J be a subset of $[n]$ such that $\pi_i\big|_J \neq \pi_j\big|_J$ for any distinct $i,j \in [k]$. Define $\ell \triangleq |J|$ and define $\eta(k/2, \ell, \phi, 1)$ as in 4.3.1. Then we have:

$$\mathrm{TV}\left(\mathcal{M}|_J\,,\, \mathcal{M}'\big|_J\right) \geq \xi \cdot \eta(k/2, \ell, \phi, 1)$$

The authors propose the following algorithm for learning the weights assuming that the centers have been correctly estimated. The main idea is to perform marginalisation on a set J such that all k central permutations have distinct projections on J. This ensures that each one of the k mixing weights will appear individually as the weight of a single component and no merging will be made. Such a J can be easily computed using a decision tree that performs a split in each node, according to some pairwise comparison, partitioning the set of permutations of the node into two non-empty subsets. The root node contains the full set of k central permutations. Each leaf contains a single center and the total number of splits is $k-1$. Each split is a pairwise comparison and J is the set of all distinct elements appearing in these comparisons. A brute force search is performed on candidate combinations of mixing weights, using a precision (step size) equal to $1/L$ per weight. The choosing criterion is TV-distance minimization between the candidate model and the empirical of the latent model. Similarly with the algorithm that estimates the central rankings, the weight retrieval algorithm computes the TV-distance between the empirical of the latent model and the empirical of each candidate model computed on fake samples.

Weights Retrieval

Input: $\hat{\pi}_1, \ldots, \hat{\pi}_k$, which are the central permutations returned by the algorithm in Theorem 4.3.4, L, N' and a set of N i.i.d. observations $\sigma_1, \ldots, \sigma_N$.

* Find in polynomial time a tuple $\mathcal{I}$ of $k-1$ pairs of distinct indices in $[n]$ such that $\chi(\hat{\pi}_i, \mathcal{I}) \neq \chi\left(\hat{\pi}_j, \mathcal{I}\right)$ for any distinct $i,j \in [k]$.

* Set J equal to the set of all indices appearing in the pairs in $\mathcal{I}$.

* Define a set of integer-valued vectors $\mathcal{R}(L) := \left\{r \in [L]^k : r_i \geq \gamma L, \sum_{i=1}^k r_i = L\right\}$.

* For each $r \in \mathcal{R}(L)$:
  * Generate N' i.i.d. random permutations $\sigma'_1, \ldots, \sigma'_{N'}$ from the Mallows mixture $\mathcal{M}'(r) = \sum_{i=1}^k \frac{r_i}{L} M(\hat{\pi}_i, \phi)$.
  * Compute the marginalized empirical distribution $\mathcal{M}'_{N'}(r)\big|_J$ of the generated sample set.

* Compute the marginalized empirical distribution $\mathcal{M}_N|_J$ of samples $\sigma_1, \ldots, \sigma_N$.

* Return the estimator $\hat{w} \in \mathbb{R}^k$ given by: $\hat{w} = \frac{1}{L}\underset{r \in \mathcal{R}(L)}{\operatorname{argmin}} \mathrm{TV}\left(\mathcal{M}'_{N'}(r)\big|_J\,,\, \mathcal{M}_N|_J\right)$.

The above algorithm is guaranteed to estimate the weights with high probability up to some desired degree of precision, as long as the number N of observations and N' of fake samples per candidate model are big enough and the step size $1/L$ is small enough.

Let $\xi > 0$ denote the aimed accuracy of estimating each weight w_i. We apply Proposition 4.3.2 with $s = \xi\eta/6$ where $\eta = \eta(k/2, \ell, \phi, 1)$ and we have that:

$$\mathrm{TV}\left(\mathcal{M}|_J\,,\,\mathcal{M}_N|_J\right) \leq \xi\eta/6 \tag{4.3}$$

with probability at least $1 - n^{-11}$, if $N \geq \frac{1}{\xi^2}\left(\log\frac{1}{\xi}\right)^{2\ell+1}\ \mathrm{poly}_k\left(\frac{1}{1-\phi}\right)\cdot\log n$. Similarly, if we choose $N' \geq N \cdot k\log L$, then Proposition 4.3.2 together with a union bound over all $r \in \mathcal{R}(L)$ implies that with probability at least $1 - n^{-11}$. it holds for all $r \in \mathcal{R}(L)$ that:

$$\mathrm{TV}\left(\mathcal{M}'(r)\big|_J\,,\,\mathcal{M}'_{N'}(r)\big|_J\right) \leq \xi\eta/6 \tag{4.4}$$

In the sequel, we condition on the event $\mathcal{E}$ of probability at least $1 - n^{-10}$ that both of the above bounds hold.

Moreover, if we choose $L \geq \frac{3k}{\xi\eta}$, then there exists $r \in \mathcal{R}(L)$ for which $\left|\frac{r_i}{L} - w_i\right| \leq \frac{\xi\eta}{3k}$ for any $i \in [k]$. For this r it holds that:

$$\mathrm{TV}\left(\mathcal{M}|_J\,,\,\mathcal{M}'(r)\big|_J\right) \leq \xi\eta/6 \tag{4.5}$$

Applying the triangle inequality on the distances in relations 4.3, 4.4 and 4.5 we obtain:

$$\mathrm{TV}\left(\mathcal{M}'_{N'}(r)\big|_J\,,\,\mathcal{M}_N|_J\right) \leq \xi\eta/2$$

On the other hand, for any $r' \in \mathcal{R}(L)$, for which there exists $i \in [k]$ $\left|\frac{r_i}{L} - w_i\right| \geq \xi$, we obtain from Proposition 4.3.3 that

$$\mathrm{TV}\left(\mathcal{M}'_{N'}(r)\big|_J\,,\,\mathcal{M}_N|_J\right) \geq 2\xi\eta/3$$

on the event $\mathcal{E}$. r' cannot be equal to $L\hat{w}$ because weight vector r defined earlier exists and achieves a better TV-distance score. We conclude that $\hat{w}$ must satisfy that $|\hat{w}_i - w_i| \leq \xi$ for each $i \in [k]$.

To satisfy the bound in 4.3 with high probability, we demand that the sample complexity is at least $\frac{1}{\xi^2}\left(\log\frac{1}{\xi}\right)^{2\ell+1}\ \mathrm{poly}_k\left(\frac{1}{1-\phi}\right)\cdot\log n$. Thus, for the weight accuracy ξ we have: $\xi \leq \frac{(\log N)^{\ell+1}}{N^{1/2}}\left(\ \mathrm{poly}_k\left(\frac{1}{1-\phi}\right)\cdot\log n\right)^{1/2} \leq \frac{(\log N)^{2k-1}}{N^{1/2}}\left(\ \mathrm{poly}_k\left(\frac{1}{1-\phi}\right)\cdot\log n\right)^{1/2}$.

The results of this paper on central ranking and weight estimation are summed up in the following theorem:

Theorem 4.3.5. *Given N i.i.d. observations from the Mallows mixture $\mathcal{M} = \sum_{i=1}^{k} w_i M(\pi_i)$*

on $\mathcal{S}_n$ with distinct central permutations $\pi_1, \ldots, \pi_k$ and a known noise parameter $\phi \in (0, 1)$. Suppose we are given $\gamma > 0$ such that $\gamma \leq \min_{i \in [k]} wi$. If $N \geq$ poly $_k\left(\frac{1}{1-\phi}, \frac{1}{\gamma}\right) \cdot \log n$, then there exists a poly $_k\left(n, \frac{1}{1-\phi}, \frac{1}{\gamma}\right)$-time algorithm which returns a mixture $\widehat{M} = \sum_{i=1}^{k} \hat{w}_i M(\hat{\pi}_i)$ such that the following holds with probability at least $1 - 2n^{-10}$: Up to a relabeling, we have $\hat{\pi}_i = \pi_i$ and $|\hat{w}_i - w_i| \leq N^{-1/2}(\log N)^{2k-1}(\log n)^{1/2}$ poly $_k\left(\frac{1}{1-\phi}\right)$ for each $i \in [k]$.

Note that the authors achieve a logarithmic dependency of the sample complexity on n, generalising the result of the single Mallows case.

Chapter 5

Learning Selective Mallows Mixture Models

5.1 Identifiability of the Selective Mallows Mixture Model

In this work our goal is to estimate the parameters of the Mallows Mixture using incomplete samples. We would like small selection sets (ideally pairwise comparisons) to be sufficient for this purpose. Practically, a selection mechanism implies identifiablity if given enough incomplete samples supported by this mechanism, that is samples that have non zero probability to be drawn, the latent parameters of the mixture can be uniquely estimated. We will now provide some formal definition of the identifiability.

Suppose models M and M' are identical. This means that they are supported on the same set of (possibly incomplete) permutations and $M(\pi) = M'(\pi)\ \forall \pi$ in the support set.

$$M(\pi) = f(s) \cdot \sum_{i=1}^{k} w_i \cdot \frac{\phi^{d_{KT}(\pi_i,\pi)}}{Z(\phi,|s|)}$$

$$M'(\pi) = f'(s) \cdot \sum_{i=1}^{k} w'_i \cdot \frac{\phi^{d_{KT}(\pi'_i,\pi)}}{Z(\phi,|s|)}$$

We suppose that the base models of the mixture have all the same spread parameter ϕ. The rankings observed are incomplete and f is the selection mechanism. The full set of items is [n]. The probability that some set s of items, $s \subseteq [n]$, is selected is equal to f(s).

Definition 5.1.1. *We say that the mixture of k distinct Mallows models is identifiable on a support set S if 2 mallows mixtures being identical implies that the two sets of central permutations must coincide, and so do the corresponding weights.*

Note that the selection mechanism is assumed to be the same among all candidate models. However even if this assumption is not made it can be easily derived that two identical models have the same selection mechanism.
If models M and M' are identical then $\forall s \subseteq [n] : \forall \pi$ supported on s : $M(\pi) = M'(\pi)$.

For each selection set s we sum over all permutations supported on s and we obtain

$$\sum_{\pi \text{ supported on } s} \sum_{i=1}^{k} w_i \cdot \frac{\phi^{d_{KT}(\pi_i,\pi)}}{Z(\phi,|s|)} = \sum_{\pi \text{ supported on } s} \sum_{i=1}^{k} w'_i \cdot \frac{\phi^{d_{KT}(\pi'_i,\pi)}}{Z(\phi,|s|)} = 1$$

$$\sum_{\pi \text{ supported on } s} M(\pi) = \sum_{\pi \text{ supported on } s} M'(\pi) \Rightarrow f(s) = f'(s)$$

We remind that [5] used a determinant calculated in [4] to show that if two complete mallows mixtures are equal on every permutation in $\mathbb{S}_n$, then they have the same (distinct) centers and the same corresponding weights. We can generalize this concept to selective mixtures. In particular we can construct a range I containing all permutations of n items and a range J containing all partial permutations of n items supported by selection mechanism f(s). We then construct an $N \times M$ matrix A, where the i-th row (c_i) corresponds to the i-th permutation in I and the j-th column (π_j) corresponds to the j-th selective permutation in J. Rows play the role of the central permutation of the mixture. Columns correspond to supported inputs to the mixture. The element A[i][j] is set equal to $f(s) \cdot \frac{\phi^{d_{KT}(c_i,\pi_j)}}{Z(\phi,|s|)}$, where s is the set of items found in π_j and is equal to the density of the selective mallows distribution with center c_i and selection mechanism f(s) calculated at point π_j. The i-th row of A is the vectorization of the selective mallows distribution with center c_i and selection mechanism f(s). Any linear combination of k rows of matrix A is the vectorization of a selective mallows mixture. So the problem of identifiability can be reduced to an algebraic problem of linear independence. The k-mixture subject to a selection mechanism f(s) is identifiable iff any set of k rows of matrix A is linearly independent.

In the case of complete mixtures f(s)=0 for all incomplete sets s and f(s)=1 for the full set of n items. In this case the normalisation constant $Z(\phi,|s|)$ can be factored out of A and does not affect the rank of A so it can be completely skipped. Moreover, if samples are complete the determinant of A can be calculated using the results of [4] and is non zero. This implies that any number of k rows of A is linearly independent so any complete mallows mixture is identifiable.

The size of A grows superexponentially with n, so brute forcing over it is impractical. We provide the example of A on 3 items. Firstly, we suppose that samples are complete. In this case normalisation constant $Z(\phi,|s|)$ can be factored out so we omit it.

$$A = \begin{bmatrix} 1 & \phi & \phi & \phi^2 & \phi^2 & \phi^3 \\ \phi & 1 & \phi^2 & \phi^3 & \phi & \phi^2 \\ \phi & \phi^2 & 1 & \phi & \phi^3 & \phi^2 \\ \phi^2 & \phi^3 & \phi & 1 & \phi^2 & \phi \\ \phi^2 & \phi & \phi^3 & \phi^2 & 1 & \phi \\ \phi^3 & \phi^2 & \phi^2 & \phi & \phi & 1 \end{bmatrix} \quad det(A) = -(\phi^2-1)^7(\phi^2-\phi+1)(\phi^2+\phi+1) \neq 0 \ \forall \phi \in (0,1)$$

Then we suppose that samples are incomplete. Since we only have 3 items, incomplete samples can only be pairwise comparisons. We construct A with rows corresponding to permutations in $\mathbb{S}_n$ and columns to pairwise comparisons of items 1,2,3.

$$A = \begin{bmatrix} 1 & 1 & \phi & \phi & 1 & \phi \\ 1 & 1 & 1 & \phi & \phi & \phi \\ \phi & \phi & 1 & 1 & \phi & 1 \\ 1 & \phi & 1 & 1 & \phi & \phi \\ \phi & \phi & \phi & 1 & 1 & 1 \\ \phi & 1 & \phi & \phi & 1 & 1 \end{bmatrix} \quad det(A) = 0 \ \forall \phi \in (0,1) \ \Rightarrow \text{no identifiability.}$$

We will now see what the identifiability of the complete mixture implies about the selective mixture.For each selection set s supported by f(s) models M and M' are projected on s and these projections are complete mixtures so the identifiability theorem of [5] holds.

$$\sum_{i=1}^{k} w_i \cdot \frac{\phi^{d_{KT}(\pi_i,\pi)}}{Z(\phi,|s|)} = \sum_{i=1}^{k} w'_i \cdot \frac{\phi^{d_{KT}(\pi'_i,\pi)}}{Z(\phi,|s|)} \quad \forall \pi \text{ supported on } s$$

This implies that projected models $M|_s$ and $M'|_s$ have the same distinct projected permutations $\pi_j, j \in [k']$, $k' \leq k$ and

$$\sum_{\pi_i \| s = \pi_j} w_i = \sum_{\pi'_i \| s = \pi_j} w'_i \quad \forall j \in [k']$$

These equations do not always ensure identifiability of the latent complete centers and even when they do, subtle manipulation of the equations obtained by different selection sets s is needed.

Let's examine how strict a selection mechanism can be without stopping to preserve identifiability.

5.1.1 Pairwise Comparisons and k=2, the General Case:

Unfortunately, even in the simple case when the mixture consists of two centers pairwise comparisons may fail to preserve identifiability. In particular, when we have a mixture of two equally weighted reversals, then the observed density on all pairwise comparisons is $\frac{1}{2}$, irrespective of what exactly those reversal permutations are.
Suppose π_1, π_2 are and reversals, $w_1 = w_2 = 1/2$. Then for every pairwise comparison (i,j) $\mathbb{P}\{i < j\} = \frac{1}{2}\frac{\phi}{1+\phi} + \frac{1}{2}\frac{1}{1+\phi} = \frac{1}{2}$ because either $i < j$ in π_1 and $i > j$ in π_2 or $i > j$ in π_1 and $i < j$ in π_2 . Thus all equally weighted mixtures of 2 reversals have the same distribution over pairwise comparisons and can not be distinguished from each other only using pairwise comparisons.

5.1.2 Pairwise Comparisons, k=2, Non Equal Weights

The case when weights are equal is degenerate. If the weights are not equal then pairwise comparisons preserve identifiability. The reason for this is that non equal weights can be used as a "signature" of each one of the two centers. That is, in pairwise comparisons on which the two centers disagree with each other, we can conclude what the result of the pairwise comparison is on each of the centers. In the case when mixing weights are equal, we could only know that the two centers disagreed but we could not match the two different answers to the query to the correct components. We will now analyse this situation more formally.

Selection mechanism f only selects pairs of elements and we have a mixture of $\pi 1, \pi 2$.
$M(\pi) = f(s) \cdot [w_1 \cdot \frac{\varphi^{d_{KT}(\pi_1,\pi)}}{Z(\varphi,|s|)} + w_2 \cdot \frac{\varphi^{d_{KT}(\pi_2,\pi)}}{Z(\varphi,|s|)}]$
$M'(\pi) = f(s) \cdot [w'_1 \cdot \frac{\varphi^{d_{KT}(\pi'_1,\pi)}}{Z(\varphi,|s|)} + w'_2 \cdot \frac{\varphi^{d_{KT}(\pi'_2,\pi)}}{Z(\varphi,|s|)}]$

The marginalised model on each pair s=(i,j) is the following:
$M(i<j) = w_1 \cdot \frac{\varphi^{\mathbb{1}\{\pi_1(i)<\pi_1(j)\}}}{\varphi+1} + w_2 \cdot \frac{\varphi^{\mathbb{1}\{\pi_2(i)<\pi_2(j)\}}}{\varphi+1}$
We suppose $M(i<j) = M'(i<j)$ for all available pairs (i,j).

$\begin{vmatrix} 1/(1+\varphi) & \varphi/(1+\varphi) \\ \varphi/(1+\varphi) & 1/(1+\varphi) \end{vmatrix} = (1-\varphi)/(1+\varphi) > 0 \Rightarrow$the marginalised model is identifiable.

There are 4 cases for s=(i,j).

1) $\pi_1(i) < \pi_1(j)$ and $\pi_2(i) < \pi_2(j) \Rightarrow$ M(i<j)=$w_1 \cdot \frac{\varphi}{\varphi+1} + w_2 \cdot \frac{\varphi}{\varphi+1} = \frac{\varphi}{\varphi+1}$
2) $\pi_1(i) > \pi_1(j)$ and $\pi_2(i) > \pi_2(j) \Rightarrow$ M(i<j)=$w_1 \cdot \frac{1}{\varphi+1} + w_2 \cdot \frac{1}{\varphi+1} = \frac{1}{\varphi+1}$
3) $\pi_1(i) < \pi_1(j)$ and $\pi_2(i) > \pi_2(j) \Rightarrow$ M(i<j)=$w_1 \cdot \frac{\varphi}{\varphi+1} + w_2 \cdot \frac{1}{\varphi+1}$
4) $\pi_1(i) > \pi_1(j)$ and $\pi_2(i) < \pi_2(j) \Rightarrow$ M(i<j)=$w_1 \cdot \frac{1}{\varphi+1} + w_2 \cdot \frac{\varphi}{\varphi+1}$

If there is a pair s_0 with $f(s_0) \neq 0$ and $\pi_1||s_0 \neq \pi_2||s_0$ (cases 3, 4) then for this pair $k' = 2 \Rightarrow \{w_1, w_2\}$ is equal to $\{w'_1, w'_2\}$ as sets. Either $(w_1, w_2) = (w'_1, w'_2)$ or $(w_1, w_2) = (w'_2, w'_1)$ as tuples.
WLOG Suppose we were in case 3 for the pair $s_0 = (i_0, j_0)$. Then $M(i_0 < j_0) = w_1 \cdot \frac{\varphi}{\varphi+1} + w_2 \cdot \frac{1}{\varphi+1} = M'(i_0 < j_0)$.
If $(w_1, w_2) = (w'_1, w'_2)$ then $\pi'_1(i_0) < \pi'_1(j_0)$ (CASE I)
If $(w_1, w_2) = (w'_2, w'_1)$ then $\pi'_2(i_0) < \pi'_2(j_0)$ (CASE II)
WLOG we can suppose that $w_1 = w'_1, w_2 = w'_2$ because a relabeling of the components does not change a mixture.

Now we can use every available pair (i,j) to construct comparison graphs for π_1 and π_2. Since $\{w_1, w_2\} = \{w'_1, w'_2\}$ both $M|_{(i,j)}$ and $M'|_{(i,j)}$ take values in the set $\{\frac{\varphi}{\varphi+1}, \frac{1}{\varphi+1}, w_1 \cdot \frac{\varphi}{\varphi+1} + w_2 \cdot \frac{1}{\varphi+1}, w_1 \cdot \frac{1}{\varphi+1} + w_2 \cdot \frac{\varphi}{\varphi+1}\}$. Since $w_1 \neq w_2$ all these four values are different from each other.

We have supposed M and M' are identical, so M(i<j)=M'(i<j).
• If models M and M' both assign probability $\frac{\varphi}{\varphi+1}$ to i<j, then we know that $\pi_1(i) < \pi_1(j)$, $\pi_2(i) < \pi_2(j)$, $\pi'_1(i) < \pi'_1(j)$ and $\pi'_2(i) < \pi'_2(j)$.
• If models M and M' both assign probability $\frac{1}{\varphi+1}$ to i<j, then we know that $\pi_1(i) > \pi_1(j)$, $\pi_2(i) > \pi_2(j)$, $\pi'_1(i) > \pi'_1(j)$ and $\pi'_2(i) > \pi'_2(j)$.
• If models M and M' both assign probability $w_1 \cdot \frac{\varphi}{\varphi+1} + w_2 \cdot \frac{1}{\varphi+1}$ to i<j, then we know that $\pi_1(i) < \pi_1(j)$, $\pi_2(i) > \pi_2(j)$, $\pi'_1(i) < \pi'_1(j)$ and $\pi'_2(i) > \pi'_2(j)$ because if the relation between i and j in the central permutations was different than the relation between i_0 and j_0 a different probability would have been assigned to $i < j$ than the probability of $i_0 < j_0$.
• If models M and M' both assign probability $w_1 \cdot \frac{1}{\varphi+1} + w_2 \cdot \frac{\varphi}{\varphi+1}$ to i<j, then we know that $\pi_1(i) > \pi_1(j)$, $\pi_2(i) < \pi_2(j)$, $\pi'_1(i) > \pi'_1(j)$ and $\pi'_2(i) < \pi'_2(j)$ The centers in each model are in a discord with each other because the probability is neither $\frac{\varphi}{\varphi+1}$ nor $\frac{1}{\varphi+1}$. If the relation between i and j in the central permutations was the same as the relation between i_0 and j_0 the same probability would have been assigned to $i < j$ as the probability of $i_0 < j_0$.

Thus the only possible case is $\pi_1(i) > \pi_1(j)$, $\pi_2(i) < \pi_2(j)$, $\pi'_1(i) > \pi'_1(j)$ and $\pi'_2(i) < \pi'_2(j)$.

In each case, the pairwise comparisons in π_1 agree with those in π'_1 and the comparisons in π_2 agree with those in π'_2. As a result the comparison graph for π_1 is the same as π'_1 and the comparison graph for π_2 is the same as π'_2. If the support set contains enough pairs, the constructed comparison graphs give total order and π'_1, π'_2 are unique and equal to π_1, π_2 respectively and identifiability is preserved. On the contrary, if some graph gives only partial order then multiple possible central permutations can be derived from this graph and identifiability does not hold.

5.1.3 Sufficient Conditions for Identifiability

We consider two selective Mallows mixture models M_1 and M_2 :
$M_1(\pi) = f(J) \cdot \sum_{i=1}^{k} w_{1,i} \cdot \frac{\varphi^{d_{KT}(\pi_{1,i}|_J, \pi)}}{Z(\varphi, |J|)}$,
$M_2(\pi) = f(J) \cdot \sum_{i=1}^{k} w_{2,i} \cdot \frac{\varphi^{d_{KT}(\pi_{2,i}|_J, \pi)}}{Z(\varphi, |J|)}$,
where the argument π is some incomplete ranking and J is the set of items found in π. The central permutations of M_1 $\{\pi_{1,1}, \pi_{1,2}, \ldots \pi_{1,k}\}$ are all distinct. The same holds for the centers of M_2. Each central permutation $\pi_{1,i}$ is a complete permutation of n items ($\pi_{1,i} \in \mathbb{S}_n$). The same holds for each center $\pi_{2,i}$.

A key ingredient for the analysis of identifiability is the following lemma found in [6]. This lemma guarantees that a small (logarithmic on k) "signature" set I of pairwise comparisons can always be found for one of the distinct permutations of a permutation set Σ. Set I is characterised as the "signature" of the corresponding permutation π^*, because the way in which the items found in I compare with each other is unique in π^* and differs from all the other permutations in Σ. For the analysis of identifiability only the existence

of the pair (I, π^*) matters, however the proof of the lemma provides a construction algorithm for I, which will be utilised by an algorithm discussed in the next chapter, which reconstructs a Mallows Mixture using small (logarithmic on k) noiseless queries. Note that this lemma only provides a way to isolate one permutation of the set Σ through its signature. It does not find a signature for each one of the elements of Σ. This limitation will need careful manipulation as we will see later.

Lemma 5.1.1 (Mao et al. 2020)**.** *Let $\chi(\pi, I)$ be a vector st $\chi(\pi, I)_r = \mathbb{1}\{\pi(i_r) < \pi(j_r)\}$ and $I = [(i_1, j_1), ..., (i_l, j_l)]$. $\forall$ set Σ of k distinct permutations in $\mathbb{S}_n$, $n \geq 2$, there exist $\pi^* \in \Sigma$ and $I = [(i_1, j_1), ..., (i_l, j_l)]$, st $l \leq \lfloor log_2(k) \rfloor$ and $(\pi \neq \pi^* \Rightarrow \chi(\pi, I) \neq \chi(\pi^*, I))$, $\forall \pi \in \Sigma$.*

Proof (construction procedure):

$\Sigma_0 := \Sigma$, r=1
while $|\Sigma_{r-1}| > 1$:
find (i_r, j_r) st:
$\Sigma_r^+ = \{\pi \in \Sigma_{r-1} : \pi(i_r) > \pi(j_r)\} \neq \{\}$
$\Sigma_r^- = \{\pi \in \Sigma_{r-1} : \pi(i_r) < \pi(j_r)\} \neq \{\}$
Σ_r := the smallest between Σ_r^+ and Σ_r^-
r+=1

$|\Sigma_r| \leq k/2^k \Rightarrow l \leq \lfloor log_2(k) \rfloor$

The above procedure is similar to a binary search. The initial set Σ is bisected in each step, based on the result of a pairwise comparison. The resulting set contains a single element $\pi*$, the permutation we succeed to isolate, and the sequence of pairwise comparisons that lead as to this element are its signature.

Theorem 5.1.1. *If for all incomplete permutations π of length l, where l in $\{2, 3, ..., 2 \cdot \lfloor log_2(k) \rfloor + 3\}$ it holds that*

1. $f(J) \neq 0$, *where J is the set of items in π and*
2. $M_1(\pi) = M_2(\pi) \Leftrightarrow \sum_{i=1}^{k} w_{1,i} \cdot \frac{\phi^{d_{KT}(\pi_{1,i}|_J, \pi)}}{Z(\phi, |J|)} = \sum_{i=1}^{k} w_{2,i} \cdot \frac{\phi^{d_{KT}(\pi_{2,i}|_J, \pi)}}{Z(\phi, |J|)}$

then $\{(w_{1,1}, \pi_{1,1}), (w_{1,2}, \pi_{1,2}), ...(w_{1,k}, \pi_{1,k})\}$ and $\{(w_{2,1}, \pi_{2,1}), (w_{2,2}, \pi_{2,2}), ...(w_{2,k}, \pi_{2,k})\}$ are equal as sets.

Proof.

The theorem can be proved by induction on n, the number of items.

Base case: for $n = l = 2 \cdot \lfloor log_2(k) \rfloor + 3$, J is the full set of items, so no selection is made and we can apply the identifiability theorem for Mallows mixture models on complete rankings that has been proved in previous work.

Induction Hypothesis:
Set C^1_{n-1} contains the distinct elements of the set $\{\pi_{1,1}|_{[n-1]}, ..., \pi_{1,k}|_{[n-1]}\}$, that is the projections of the centers of M_1 on items $\{1, 2...n-1\}$. Some centers might have the same projections. However C^1_{n-1} is not a multiset. We consider the distinct elements of this set. $C^1_{n-1} = \{\pi^c_{1,1}, ..., \pi^c_{1,k_c}\}$, where $\pi^c_{1,i}$ are all distinct and their cardinality k_c is at most k ($k_c < k$ if some centers have the same projections).

We also consider the set WC^1_{n-1} that contains the tuples $(\pi^c_{1,j}, \sum_{i:\pi_{1,i}|[n-1]=\pi^c_{1,j}} w_{1,i})$ of distinct centers paired with their cumulative weight.
$WC^1_{n-1} := \{(\pi^c_{1,1}, \sum_{i:\pi_{1,i}|[n-1]=\pi^c_{1,1}} w_{1,i}), ..., (\pi^c_{1,k_c}, \sum_{i:\pi_{1,i}|[n-1]=\pi^c_{1,k_c}} w_{1,i})\}$

We now consider the same quantities for mixture M_2.

Set C^2_{n-1} contains the distinct elements of the set $\{\pi_{2,1}|_{[n-1]}, ..., \pi_{2,k}|_{[n-1]}\}$
$C^2_{n-1} = \{\pi^c_{2,1}, ..., \pi^c_{2,k_c}\}$
$WC^2_{n-1} := \{(\pi^c_{2,1}, \sum_{i:\pi_{2,i}|[n-1]=\pi^c_{2,1}} w_{2,i}), ..., (\pi^c_{2,k_c}, \sum_{i:\pi_{2,i}|[n-1]=\pi^c_{2,k_c}} w_{2,i})\}$

We suppose $WC^1_{n-1} = WC^2_{n-1}$.

Induction Step:
$\{(c_{1,1}, wt_{1,1}), (c_{1,2}, wt_{1,2}), ..., (c_{1,k_1}, wt_{1,k_1})\} := WC^1_n$
$\{(c_{2,1}, wt_{2,1}), (c_{2,2}, wt_{2,2}), ..., (c_{2,k_2}, wt_{2,k_2})\} := WC^2_n$

We will show that $WC^1_n = WC^2_n$.

From induction hypothesis it holds that $WC^1_{n-1} = WC^2_{n-1}$ but we only need the fact that

$C^1_{n-1} = C^2_{n-1}$.

Wlog we suppose that the order of the elements $\pi^c_{2,1}, ..., \pi^c_{2,k_c}$ is such that $\pi^c_{1,i} = \pi^c_{2,i}$, $\forall\ i \in [k_c]$.
Also, wlog, we suppose that the order of the elements $\pi^c_{1,1}, ..., \pi^c_{1,k_c}$ is such that if we apply Lemma 5.1.1 on them, permutation $\pi^c_{1,1}$ and its 'signature' set of comparisons will be returned, if we delete it and apply the lemma on the remaining ones $\pi^c_{1,2}$ (and its signature) will be returned and generally if we apply the lemma on the subset $\{\pi^c_{1,i}, \pi^c_{1,i+1}, ..., \pi^c_{1,k_c}\}$, $i \in [k_c]$, $\pi^c_{1,i}$ will be returned.

We apply Lemma 5.1.1 on the set C^1_{n-1}. There exists I, an l-tuple of pairwise comparisons st:
$l \leq \lfloor log_2(k) \rfloor$ and $(j \neq 1 \Rightarrow \chi(\pi^c_{1,j}, I) \neq \chi(\pi^c_{1,1}, I))$
We consider the centers $\boldsymbol{c_{1,l}}$ in WC^1_n that satisfy the restriction $c_{1,l}|[n-1] = \pi^c_{1,1} = [e_1, e_2, ...e_{n-1}]$.
There are 3 cases for $c_{1,l}$:
$\boldsymbol{c_{1,l}} = [e_1, e_2, ..., e_{r-1}, n, e_r, ..., e_{n-1}]$, for some r in $[2, n-1]$ (I) or
$\boldsymbol{c_{1,l}} = [e_1, e_2, ..., e_{n-1}, n]$ (II) or
$\boldsymbol{c_{1,l}} = [n, e_1, e_2, ..., e_{n-1}]$ (III)

In each of the 3 cases it suffices to know the relative order among e_{r-1}, e_r and n to fully determine $c_{1,l}$ (given the fact that $c_{1,l}|[n-1] = \pi^c_{1,1} = [e_1, e_2, ...e_{n-1}]$).
We take the set of indices J := set of elements appearing in I union $\{e_{r-1}, e_r, n\}$. $|J| \leq 2 \cdot \lfloor log_2(k) \rfloor + 3$
$M_1|_J(\pi) = \sum_{i=1}^{k} w_{1,i} \cdot \frac{\phi^{d_{KT}(\pi_{1,i}|_J, \pi)}}{Z(\phi, |J|)} = \sum_{j=1}^{k'} w'_j \cdot \frac{\phi^{d_{KT}(\pi'_j, \pi)}}{Z(\phi, |J|)}$,
The centers π'_j are distinct permutations of the elements in J .
$w'_j = \sum_{\pi_{1,i}|J = \pi'_j} w_{1,i}$.
$M_2|_J(\pi) = \sum_{i=1}^{k} w_{2,i} \cdot \frac{\phi^{d_{KT}(\pi_{2,i}|_J, \pi)}}{Z(\phi, |J|)}$
$M_1|_J(\pi) = M_2|_J(\pi)$ $\forall$ permutation π of the items of J. The identifiability theorem for complete mallows mixtures implies that $M_1|_J, M_2|_J$ have the same distinct centers and the same corresponding weights. So $M_2|_J(\pi)$ is also equal to $\sum_{j=1}^{k'} w'_j \cdot \frac{\phi^{d_{KT}(\pi'_j, \pi)}}{Z(\phi, |J|)}$.

For $M_1|_J$ we have:
One of the distinct centers π'_j is equal to $\boldsymbol{c_{1,l}|_J}$. We name it π'_{j^*}.
Since J contains all the elements of I, the set $S_1 = \{\pi_{1,i} : \pi_{1,i}|_J = \pi'_{j^*}\}$ is a subset of $\{\pi_{1,i} : \pi_{1,i}|_{[n-1]} = \pi^c_{1,1}\}$. In fact $S_1 = \{\pi_{1,i} : \pi_{1,i}|_{[n-1]} = \pi^c_{1,1}$ and the relative order among e_{r-1}, e_r and n is the same in $\pi_{1,i}$ and $\pi'_{j^*}\}$.
This yields that $S_1 = \{\pi_{1,i} : \pi_{1,i}|_{[n]} = \boldsymbol{c_{1,l}}\}$.
For the weights we have: $S_1 = \{\pi_{1,i} : \pi_{1,i}|_J = \pi'_{j^*}\} = \{\pi_{1,i} : \pi_{1,i}|_{[n]} = \boldsymbol{c_{1,l}}\} \Rightarrow w'_{j^*} = \sum_{\pi_{1,i}|J = \pi'_{j^*}} w_{1,i} = \sum_{i:\pi_{1,i}|[n] = c_{1,l}} w_{1,i} = wt_{1,l}$.

We make a similar analysis for $M_2|_J$:

$\pi'_{j^*} = c_{1,l}|_J$.
I is such that ($j \neq 1 \Rightarrow \chi(\pi^c_{1,j}, I) \neq \chi(\pi^c_{1,1}, I)$). But $\pi^c_{1,j} = \pi^c_{2,j}$, so it holds that ($j \neq 1 \Rightarrow \chi(\pi^c_{2,j}, I) \neq \chi(\pi^c_{2,1}, I)$) $S_2 = \{\pi_{2,i} : \pi_{2,i}|_J = \pi'_{j^*}\} = \{\pi_{2,i} : \pi_{2,i}|_{[n-1]} = \pi^c_{1,1}$ and the relative order among e_{r-1}, e_r and n is the same in $\pi_{2,i}$ and $\pi'_{j^*}\} = \{\pi_{2,i} : \pi_{2,i}|_{[n]} = c_{1,l}\}$
$w'_{j^*} = \sum_{\pi_{2,i}|J=\pi'_{j^*}} w_{2,i} = \sum_{i:\pi_{2,i}|[n]=c_{1,l}} w_{2,i} = wt_{1,l}$
The set $\{i : \pi_{2,i}|[n] = c_{1,l}\}$ is non empty, so for some l' in $[k_2]$ it holds that $c_{2,l'} = c_{1,l}$. We also have that $wt_{1,l} = \sum_{i:\pi_{2,i}|[n]=c_{1,l}} w_{2,i} = \sum_{i:\pi_{2,i}|[n]=c_{2,l'}} w_{2,i} = wt_{2,l'}$.

We have shown that $\forall (c_{1,l}, wt_{1,l}) \in WC^1_n$ *s.t.* $c_{1,l}|[n-1] = \pi^c_{1,1}$ there exists $(c_{2,l'}, wt_{2,l'}) \in WC^2_n$ *s.t.* $(c_{2,l'}, wt_{2,l'}) = (c_{1,l}, wt_{1,l})$.
Working symmetrically, we can prove that $\forall (c_{2,l}, wt_{2,l}) \in WC^2_n$ *s.t.* $c_{2,l}|[n-1] = \pi^c_{1,1}$ there exists $(c_{1,l'}, wt_{1,l'}) \in WC^1_n$ *s.t.* $(c_{1,l'}, wt_{1,l'}) = (c_{2,l}, wt_{2,l})$.

We now have to show the same for the centers $c_{1,l}$ in WC^1_n that satisfy the restriction $c_{1,l}|[n-1] = \pi^c_{1,i}$, where $i \geq 2$. We remind that if we apply Lemma 5.1.1 on the subset $\{\pi^c_{1,i}, \pi^c_{1,i+1}, ..., \pi^c_{1,k_c}\}$, $i \in [k_c]$, $\pi^c_{1,i}$ and its signature I will be returned. In each case I contains l pairwise comparisons, where $l \leq \lfloor log_2(k) \rfloor$, but I does only work as a signature on the subset $\{\pi^c_{1,i}, \pi^c_{1,i+1}, ..., \pi^c_{1,k_c}\}$. On the full set C^1_{n-1} there may be $\pi^c_{1,j}$ s.t. $j \neq i$ but $\chi(\pi^c_{1,j}, I) = \chi(\pi^c_{1,i}, I)$). In this case $wt_{1,l}$=w'_{j^*} - $\sum_{m:c_{1,m}|J=c_{1,l}|J} wt_{1,m}$. The set $c_{1,m} : c_{1,m}|J = c_{1,l}|J$ contains centers $c_{1,m}$ in $\mathbb{S}_n$ s.t. $c_{1,m}|[n-1] = \pi^c_{1,j}, j < i$. Inductively we yield that the sets WC^1_n and WC^2_n are equal constrained on$\{c_{1,m} : c_{1,m}|[n-1] = \pi^c_{1,j}, j < i\}$ and $\{c_{2,m} : c_{2,m}|[n-1] = \pi^c_{2,j}, j < i\}$. So $wt_{1,l}$=w'_{j^*} - $\sum_{m:c_{1,m}|J=c_{1,l}|J} wt_{1,m}$= w'_{j^*} - $\sum_{m:c_{2,m}|J=c_{1,l}|J} wt_{2,m}$=$wt_{2,l'}$ for some l' . This way we show again like the case of $\pi^c_{1,1}$, that $\forall (c_{1,l}, wt_{1,l}) \in WC^1_n$ *s.t.* $c_{1,l}|[n-1] = \pi^c_{1,i}, i \in [k_c]$ there exists $(c_{2,l'}, wt_{2,l'}) \in WC^2_n$ *s.t.* $(c_{2,l'}, wt_{2,l'}) = (c_{1,l}, wt_{1,l})$ and reversely $\forall (c_{2,l}, wt_{2,l}) \in WC^2_n$ *s.t.* $c_{2,l}|[n-1] = \pi^c_{1,i}, i \in [k_c]$ there exists $(c_{1,l'}, wt_{1,l'}) \in WC^1_n$ *s.t.* $(c_{1,l'}, wt_{1,l'}) = (c_{2,l}, wt_{2,l})$.

$WC^1_{n_{max}} = \{(w_{1,1}, \pi_{1,1}), (w_{1,2}, \pi_{1,2}), ...(w_{1,k}, \pi_{1,k})\}$ and $WC^1_{n_{max}} = WC^2_{n_{max}}$. So we have shown that $\{(w_{1,1}, \pi_{1,1}), (w_{1,2}, \pi_{1,2}), ...(w_{1,k}, \pi_{1,k})\}$ and
$\{(w_{2,1}, \pi_{2,1}), (w_{2,2}, \pi_{2,2}), ...(w_{2,k}, \pi_{2,k})\}$ are equal as sets.

In fact, one comparison can be saved from the query length, when we try to place the i-th item in the correct position on the centers restricted on items 1,2,..i-1. This can be achieved by checking the possible positions in a specific order, starting from item i-1 and continuing up to item 1. We make use of the fact that the total weight of the centers that place the i-th item between the consecutive items a and a+1 in the projected center is equal to the total weight of the centers that place it before a+1 minus the total weight of the centers that place it before a. Therefore, we only need to add pairwise comparisons (a,i), (b,i) to the signature rather than a 3-wise comparison (a,b,i). This way the queries to the strong oracle have length at most $2 \cdot \lfloor log_2(k) \rfloor + 2$.

The above analysis assumes that spread parameters are equal in all components. This

in fact is the most difficult case in terms of identifiability, due to [5]. In this paper the authors show that given enough samples and assuming $n \geq 10k$ we can learn the central rankings exactly and the corresponding weights and spread parameters with an absolute error arbitrarily small (thus approaching zero given infinite samples). Thus, the identifiability result for complete mixtures of Mallows models with equal spread parameters, following from the determinant of Zagier et al., extends to the case of non equal spread parameters, with the condition $n \geq 10k$. For the selective Mallows mixture, in the case of non equal spread parameters, the proof of identifiability would be very similar to the case of equal spread parameters, with the different spread parameters functioning as part of the "signature" of each component, making its isolation easier. In conclusion, based on the current literature, the bottleneck for the selectivity in the case of non equal spread parameters is the minimal number of items m_k^* required to learn the complete mixture of k components with non equal spread parameters. Currently, $m_k^* = 10k$. Generally, the identifiability condition in the case where spread parameters are not necessarily equal is that selection sets J should have length $|J| \geq max(m_k^*, 2 \cdot \lfloor log_2(k) \rfloor + 2$
In the case of equal spread parameters, the conditions for the length of the selections are actually tight. If the selection sets contain less than $2 \cdot \lfloor log_2(k) \rfloor + 2$ items, then certain k-mixtures are not identifiable.

5.1.4 Tight Examples for the sufficient Conditions for Identifiability

Theorem 5.1.2. *If $l < 2(\lfloor log_2(k) \rfloor + 1)$, then there exist two mixtures M_1, M_2 with different sets of central permutations and $M_1(\pi) = M_2(\pi), \forall \pi$ with length less or equal to l.*

Proof.
For $n = 2m, k = 2^{m-1}$ we can always construct two sets S_1, S_2 of distinct permutations, st. $|S_1| = |S_2| = k = 2^{m-1}$ and the projections of the two sets on any selection of n-1 items are equal. For each $v \in \{0,1\}^m$ we define a unique permutation $\pi_v \in S_{2m}$ as follows:

$$\pi_v(2j-1) = 2j-1, \quad \pi_v(2j) = 2j, \quad if \ v_j = 0$$

$$\pi_v(2j-1) = 2j, \quad \pi_v(2j) = 2j-1, \quad if \ v_j = 1$$

$\forall j \in [m]$

We define

$$S_1 = \{\pi_v : \ v \in \{0,1\}^m, \ sum(v) \ is \ odd\}$$

$$S_2 = \{\pi_v : \ v \in \{0,1\}^m, \ sum(v) \ is \ even\}$$

$\forall J \subset [n], |J| = n-1 : \{\pi_v||_J : \ v \in \{0,1\}^m, \ sum(v) \ is \ odd\} = \{\pi_v||_J : \ v \in \{0,1\}^m, \ sum(v) \ is \ even\}$
Proof: Sps the missing element is j_1. For every permutation π_1 in S_1 we find j_1's pair, which might be (j_1, j_2) or (j_2, j_1), j_2 adjacent to j_1, and we change the order between j_1, j_2. The resulting permutation π_2 is in S_2. j_1 is not in J so the pair simplifies to the single

element j_2 and $\pi_1\|_J = \pi_2\|_J$.

$$M_1(\pi) = f(J) \cdot \sum_{i=1}^{k} \frac{1}{k} \cdot \frac{\phi^{d_{KT}(\pi_{1,i}\|_J, \pi\|_J)}}{Z(\phi, |J|)}$$

$$M_2(\pi) = f(J) \cdot \sum_{i=1}^{k} \frac{1}{k} \cdot \frac{\phi^{d_{KT}(\pi_{2,i}\|_J, \pi\|_J)}}{Z(\phi, |J|)}$$

$\{\pi_{1,i}\} \neq \{\pi_{2,i}\}$ but $M_1(\pi) = M_2(\pi)$ for every π supported on a strict subset of $[n]$.
For $n > 2m$ we can extend the permutations if S_1,S_2 with identity permutations over the extra elements and the theorem for S_1, S_2 continues to hold.

Example for n=8

v	π_v
(0, 0, 0, 0)	[1, 2, 3, 4, 5, 6, 7, 8]
(0, 0, 0, 1)	[1, 2, 3, 4, 5, 6, 8, 7]
(0, 0, 1, 0)	[1, 2, 3, 4, 6, 5, 7, 8]
(0, 0, 1, 1)	[1, 2, 3, 4, 6, 5, 8, 7]
(0, 1, 0, 0)	[1, 2, 4, 3, 5, 6, 7, 8]
(0, 1, 0, 1)	[1, 2, 4, 3, 5, 6, 8, 7]
(0, 1, 1, 0)	[1, 2, 4, 3, 6, 5, 7, 8]
(0, 1, 1, 1)	[1, 2, 4, 3, 6, 5, 8, 7]
(1, 0, 0, 0)	[2, 1, 3, 4, 5, 6, 7, 8]
(1, 0, 0, 1)	[2, 1, 3, 4, 5, 6, 8, 7]
(1, 0, 1, 0)	[2, 1, 3, 4, 6, 5, 7, 8]
(1, 0, 1, 1)	[2, 1, 3, 4, 6, 5, 8, 7]
(1, 1, 0, 0)	[2, 1, 4, 3, 5, 6, 7, 8]
(1, 1, 0, 1)	[2, 1, 4, 3, 5, 6, 8, 7]
(1, 1, 1, 0)	[2, 1, 4, 3, 6, 5, 7, 8]
(1, 1, 1, 1)	[2, 1, 4, 3, 6, 5, 8, 7]

Suppose we exclude element 3.
S1=[1, 2, 4, 5, 6, 8, 7], [1, 2, 4, 6, 5, 7, 8], [1, 2, 4, 5, 6, 7, 8], [1, 2, 4, 6, 5, 8, 7], [2, 1, 4, 5, 6, 7, 8], [2, 1, 4, 6, 5, 8, 7], [2, 1, 4, 5, 6, 8, 7], [2, 1, 4, 6, 5, 7, 8]

S2=[1, 2, 4, 5, 6, 7, 8], [1, 2, 4, 6, 5, 8, 7], [1, 2, 4, 5, 6, 8, 7], [1, 2, 4, 6, 5, 7, 8], [2, 1, 4, 5, 6, 8, 7], [2, 1, 4, 6, 5, 7, 8], [2, 1, 4, 5, 6, 7, 8], [2, 1, 4, 6, 5, 8, 7]

S1=S2 as sets

5.2 Algorithm for learning the Mallows mixture performing noiseless queries.

Here we present an algorithm for learning the parameters of a mallows mixture given all its marginals on $2 \cdot \lfloor log_2(k) \rfloor + 3$ items.
For n=2 we can learn the restricted mixture with one query to the strong oracle.
for n in $[3, n_{max}]$:
Consider the set C containing the distinct elements of the set $\{\pi_1|_{[n-1]}, ..., \pi_k|_{[n-1]}\}$.
l=0
do{
Apply Lemma 5.1.1 on C.
There exist $s^* \in [k]$ and I l-tuple of pairwise comparisons st:
$l \leq \lfloor log_2(k) \rfloor$ and $(\pi_s|_{[n-1]} \neq \pi_{s^*}|_{[n-1]} \Rightarrow \chi(\pi_s, I) \neq \chi(\pi_{s^*}, I))$
$\pi_{s^*}|_{[n-1]}$ is a permutation of the items [1,2, .. n-1]. We express it as a sequence: $\pi_{s^*}|_{[n-1]} = [e_1, e_2, ...e_{n-1}]$
for r in $[2, n-1]$:
J := set of elements appearing in I union $\{e_{r-1}, e_r, n\}$
We obtain the distribution $M(\pi) = \sum_{i=1}^{k} w_i \cdot \frac{\phi^{d_{KT}(\pi_i|_J, \pi)}}{Z(\phi, |J|)} =$
$\sum_{j=1}^{k'} w'_j \cdot \frac{\phi^{d_{KT}(\pi'_j, \pi)}}{Z(\phi, |J|)}$, where π is in the set that contains all permutations of the items of J.
The centers π'_j are distinct permutations of the elements in J .

From the identifiability theorem for complete ranking mixtures we can identify each distinct center π'_j and its total weight $w'_j = \sum_{\pi_i|J=\pi'_j} w_i$. One of the distinct centers π'_{j^*} is equal to $\pi_{s^*|J}$.
Since J contains all the elements of I, the set S=$\{\pi_i : \pi_i|_J = \pi'_{j^*}\}$ is a subset of $\{\pi_i : \pi_i|_{[n-1]} = \pi_{s^*}|_{[n-1]}\}$.
In fact S = $\{\pi_i : \pi_i|_{[n-1]} = \pi_{s^*}|_{[n-1]}$ and the relative order among e_{r-1}, e_r and n is the same in π_i and $\pi_{s^*}\}$.

If π'_{j^*} contains the ordered triplet (e_{r-1}, n, e_r), then we have learned a new center c_l in $\mathbb{S}_n$, which is equal to $[e_1, e_2, ..., e_{r-1}, n, e_r, ..., e_{n-1}]$, and its corresponding weight $wt_l = w'_{j^*} - \sum_{m:c_m|J=c_l|J} wt_m$.
Else if π'_{j^*} ends in (e_{n-1}, n), then we have learned a new center c_l in $\mathbb{S}_n$, which is equal to $[e_1, e_2, ..., e_{n-1}, n]$, and its corresponding weight $wt_l = w'_{j^*} - \sum_{m:c_m|J=c_l|J} wt_m$.
Else if π'_{j^*} starts with (n, e_1), then we have learned a new center c_l in $\mathbb{S}_n$, which is equal to $[n, e_1, e_2, ..., e_{n-1}]$, and its corresponding weight $wt_l = w'_{j^*} - \sum_{m:c_m|J=c_l|J} wt_m$.
Remove $\pi_{s^*}|_{[n-1]}$ from C.
l+=1
}**while** C not empty

The above algorithm is based on the work of [6], filling in some important details missing from their description. It has been tested successfully on various synthetic mixtures. We will now present an example of the execution of the algorithm. Suppose the latent Mixture has the following parameters:

$\pi_1 = [8,7,6,5,1,3,4,2],\ w_1 = 0.3,$
$\pi_2 = [1,2,3,4,6,7,8,5],\ w_2 = 0.1,$
$\pi_3 = [1,7,5,3,6,8,2,4],\ w_3 = 0.1,$
$\pi_4 = [2,7,8,4,5,1,3,6],\ w_4 = 0.2,$
$\pi_5 = [2,7,8,6,5,1,3,4],\ w_5 = 0.25,$
$\pi_6 = [2,7,8,6,5,1,4,3],\ w_6 = 0.05$

We start from the restriction of the mixture on items $\{1,2\}$. Collection C of marginalised centers contains both $[1,2]$ (stemming from $\pi 1, \pi_2, \pi_3$) and $[2,1]$ (stemming from π_4, π_5, π_6). Signature set $\{1,2\}$ is incorporated in the first query, along with the third item. The full query set is $J = \{1,2,3\}$. The marginalised model on these items is the following: $\{([1,2,3],0.1),\ ([1,3,2],0.4),\ ([2,1,3],0.5)\}$. In this case the query set was big enough to derive the marginalised mixture on items $1,2,...n$ explicitly.

The algorithm goes on to learn the marginal on items $1,2,3,4$. Collection C of marginalised centers on this items is equal to $\{[1,2,3],[1,3,2],[2,1,3]\}$. Signature $I = \{1,2\}$ isolates the third element of C: $\pi_{s^*}|_{[3]} = [2,1,3]$.
Item 4 along with signature I constitute the first query set $J = \{1,2,4\}$. This query aims to detect centers deriving from $\pi_{s^*}|_{[3]}$, where item 4 lies in the first place, right before 2, or between 2 and 1. The marginal mixture on J is the following:
$\{([1,2,4],0.2),\ ([1,4,2],0.3),\ ([2,1,4],0.3),\ ([2,4,1],0.2)\}$.
Component $([2,4,1],0.2)$, satisfies the search condition, because item 4 lies between 2 and 1. So we conclude that $([2,4,1,3],0.2)$ is a component of the mixture on items $1,2,3,4$.

The next query is somewhat trivial. Possible insertion positions are considered for item 4, either between 2 and 3 or after 3. This requires the addition of item 3 to J, so J trivially covers the full set of items $1,2,3,4$. From this query components $([2,1,4,3],\ 0.05)$ and $([2,1,3,4]\ 0.25)$ are learned. At this step all components derived from $[2,1,3]$ are identified. We delete this element from C and continue to the next candidate.

After the deletion C is equal to $\{[1,2,3],[1,3,2]\}$. Signature $I = \{2,3\}$ isolates the first element of C: $\pi_{s^*}|_{[3]} = [1,2,3]$.
In order to detect candidate centers $[4,1,2,3,]$ or $[1,4,2,3,]$ both items 1 and 4 are added to the signature so J covers the full range of items and the query is trivial again. None of these candidate centers is detected so we continue to candidates where 4 is either placed between 2 and 3 or after 3. In this case $J = \{2,3,4\}$. The marginal mixture on J is the

following: {([2, 3, 4], 0.35), ([2, 4, 3], 0.25), ([3, 2, 4], 0.1), ([3, 4, 2], 0.3)}. From these only the first two satisfy the restriction of the signature (2<3). The first one gives a new component of the mixture on four items, that is ([1, 2, 3, 4] 0.1). The weight of this component is 0.1 = 0.35 – 0.25, not 0.35 as it appears on the marginal mixture on J. The reason for this is that ([2, 3, 4], 0.35) includes the already learned component ([2, 1, 3, 4] 0.25). We need to exclude the weight of previously learned components in order to learn new ones. The signature $I = \{2, 3\}$ only works on the diminished set C. On the full set it also corresponds to [2, 1, 3], apart from [1, 2, 3].
Similarly, the second component of the marginal ([2, 4, 3], 0.25) appears to give a new component on four items, with weight 0.25. However this weight corresponds to previously learned components ([2, 4, 1, 3], 0.2) and ([2, 1, 4, 3], 0.05). Thus no new component is derived. At this point we have learned all the components derived from [1, 2, 3]. We delete it from C.

Now C contains a single element, that is [1, 3, 2]. No signature is needed. Queries only contain new item 4 and its candidate neighbours (1,4,3) or (3,4,2). Query $J = \{1, 4, 3\}$ gives no new components. Query $J = \{3, 4, 2\}$ detects the centers [1, 3, 2, 4] and [1, 3, 4, 2] with weights 0.1 and 0.3 respectively.

At this point the algorithm has successfully learned the marginal mixture on the first four items. In the next iterations the mixture is learned on items 1,2,..i, until the 8-th iteration, when the full mixture is learned.

To demonstrate a more interesting case of signature calculation we jump to the start of 8-th iteration. The collection C of centers on items 1,2,.. 7 is the following:
$C =$ {[7, 6, 5, 1, 3, 4, 2]
[1, 2, 3, 4, 6, 7, 5]
[1, 7, 5, 3, 6, 2, 4]
[2, 7, 4, 5, 1, 3, 6]
[2, 7, 6, 5, 1, 3, 4]
[2, 7, 6, 5, 1, 4, 3]}

Firstly, we use pairwise comparison (1,2). This way C is split into two halves. The first half contains permutations that place item 1 before item 2. In the second half $2 < 1$. We keep the first half:
$C_1 =$ {[7, 6, 5, 1, 3, 4, 2]
[1, 2, 3, 4, 6, 7, 5]
[1, 7, 5, 3, 6, 2, 4]}
Then we use pairwise comparison (1,5). This comparison splits C_1 into two parts. $C_1^- =$ {[7, 6, 5, 1, 3, 4, 2]}, where 5 is placed before 1 and $C_1^+ = \{[1, 2, 3, 4, 6, 7, 5], [1, 7, 5, 3, 6, 2, 4]\}$, where 5 is placed after 1. $C_2 = C_1^-$ is a unit set, so the procedure terminates. The isolated center $\pi_{s^*}|_{[7]}$ is [7, 6, 5, 1, 3, 4, 2] and the signature set I deriving from comparisons (1,2) and (1,5) is $\{1, 2, 5\}$.

5.3 Learning Mixtures of Two Mallows Models Using Pairwise Comparisons

Awasthi, Blum et al. in [23] give an algorithm for learning Mixtures of two Mallows Models using complete samples. Mao et al. in [6] propose an algorithm for learning mixtures of k Mallows Models that uses groups of pairwise comparisons. In the case of two centers each sample should contain at least two pairwise comparisons sampled simultaneously from the same center, or a 4-wise comparison. However, if as we saw earlier in the identifiability section pairwise comparisons should suffice. We will extend the ideas of the identifiability section to the setting of noisy oracles. We will try to simulate the strong noiseless oracle using samples of the noisy oracle.

We have seen that the cases where $\phi = 1$ or $w_1 = w_2 = 0.5$ are degenerate. We consider the level of degeneracy of the mixture.

Definition 5.3.1. *We say that the mixture $w_1 \cdot M(\phi, \pi_1) + w_2 \cdot M(\phi, \pi_2)$ is $a-$non degenerate iff $|w_i - 0.5| > a$, $w_i > a$, for i=1,2 and $\phi < 1 - a$, where a is some positive constant less than 1.*

We suppose that the mixture is $a-$non degenerate. Also, wlog we suppose that $w_1 < w_2$. Then we have $a < w_1 < 0.5 - a < 0.5 + a < w_2 < 1 - a$.
The probability mass function of the mixture takes values in the set $\{p_1, p_2, p_3, p_4\}$, where
$p_1 = \frac{\phi}{\phi+1}$,
$p_2 = w_1 \cdot \frac{1}{\phi+1} + w_2 \cdot \frac{\phi}{\phi+1}$,
$p_3 = w_1 \cdot \frac{\phi}{\phi+1} + w_2 \cdot \frac{1}{\phi+1}$,
$p_4 = \frac{1}{\phi+1}$
These four numbers are distinct and their order is $p_1 < p_2 < p_3 < p_4$ Depending on how items e_i, e_j compare with each other in each of the two central rankings, the query on the pairwise comparison (e_i, e_j) follows one of the four possible Bernoulli distributions $Be(p_l)$, l in $\{1, 2, 3, 4\}$.

For each pairwise comparison we will try to detect which Bernoulli it follows by estimating the Bernoulli parameter p_l empirically from the samples. We want the estimation to be close to the correct value of the parameter so that the detection (classification) is correct. The more samples we use for the estimation, the closer it gets to the correct value, with high probability. Another important factor is how close the latent parameters $\{p_1, p_2, p_3, p_4\}$ are to each other. The closer they are, the more difficult the detection.

The non degeneracy condition allows us to bound the difference between the latent parameters. In particular we have:
$p_2 - p_1 = w_1 \cdot \frac{1}{\phi+1} + w_2 \cdot \frac{\phi}{\phi+1} - \frac{\phi}{\phi+1} = \frac{w_1(1-\phi)}{\phi+1} > \frac{a^2}{2}$
$p_4 - p_3 = \frac{1}{\phi+1} - w_1 \cdot \frac{\phi}{\phi+1} - w_2 \cdot \frac{1}{\phi+1} = \frac{w_1(1-\phi)}{\phi+1} > \frac{a^2}{2}$
$p_3 - p_2 = \frac{2(1-\phi)(0.5-w_1)}{\phi+1} > a^2$

Learning Algorithm For Mixtures of Two Mallows Models using pairwise comparisons

Given our sample set Π consisting of pairwise comparisons $\{c_1, c_2, \ldots c_N\}$, where each c_l is of the form $(e_i < e_j)$, we compute the quantities $q(i<j)$ for all $i,j \in [n] \times [n]$: $q(i<j) = \sum_{c \in \Pi} \mathbb{1}\{c = (e_i < e_j)\}$

for (i,j) in $[n] \times [n]$ and $i \neq j$:
$\hat{p}_{i,j} := \frac{q(i<j)}{q(i<j)+q(i>j)}$

We define two parallel lists of clusters, one for the empirical frequencies and one for the corresponding pairwise comparisons.
frequency_clusters=$[[\hat{p}_{1,2}]]$
comparison_clusters=$[[(e_1 < e_2)]]$
threshold=$a^2/4$
k=1
for (i,j) in $[n] \times [n]$ and $i \neq j$:
  for l in $[1, k]$:
    choose a random element p'_l in frequency_cluster l
    $d_l := |\hat{p}_{i,j} - p'_l|$
  if $min\{d\}$ >threshold:
    k+=1
    frequency_clusters.append($[\hat{p}_{i,j}]$)
    comparison_clusters.append($[(e_i < e_j)]$)
  else:
    frequency_clusters[argmin$\{d\}$].append($\hat{p}_{i,j}$)
    comparison_clusters[argmin$\{d\}$].append($(e_i < e_j)$)

If the algorithm has executed correctly, that is all empirical frequencies are close to their theoretical values, then we expect either k=2 (if central permutations are reversals) or k=4 (if central permutations are not reversals)

WLOG we assume that frequency_clusters list is sorted in increasing order. If not, we sort it and ensure that comparison_clusters stays parallel to it.
if k=2:
  first cluster has comparisons $(e_i < e_j)$ st. $\pi_1(e_i) < \pi_1(e_j)$ and $\pi_2(e_i) > \pi_2(e_j)$
  second cluster has comparisons $(e_i < e_j)$ st. $\pi_1(e_i) > \pi_1(e_j)$ and $\pi_2(e_i) < \pi_2(e_j)$
if k=4:
  first cluster has comparisons $(e_i < e_j)$ st. $\pi_1(e_i) > \pi_1(e_j)$ and $\pi_2(e_i) > \pi_2(e_j)$
  second cluster has comparisons $(e_i < e_j)$ st. $\pi_1(e_i) < \pi_1(e_j)$ and $\pi_2(e_i) > \pi_2(e_j)$
  third cluster has comparisons $(e_i < e_j)$ st. $\pi_1(e_i) > \pi_1(e_j)$ and $\pi_2(e_i) < \pi_2(e_j)$
  fourth cluster has comparisons $(e_i < e_j)$ st. $\pi_1(e_i) < \pi_1(e_j)$ and $\pi_2(e_i) < \pi_2(e_j)$

The proposed algorithm succeeds to reconstruct the latent central permutations as long as each used pairwise comparison has an empirical frequency $\hat{p}_{i,j}$ close to the theoretical expected frequency $p_{i,j}$. It suffices to hold that for all pairwise comparisons the sample frequency has a difference of at most $a^2/8$ from the corresponding theoretical value. As we will see below this can be achieved using enough samples from each pairwise comparison. Let $N_{i,j}$ be the number of samples containing items i,j and N be the minimum of these numbers. That is, all pairwise comparisons are represented in at least N samples. Then, using the union bound and Hoeffding's inequality we obtain:

$$\begin{aligned}
\mathbb{P}\{\text{incorrect centers reconstruction}\} &\leq \mathbb{P}\{\bigcup_{(i,j)\in[n]^2} \{|\hat{p}_{i,j} - p_{i,j}| > a^2/8\} \\
&\leq \sum_{(i,j)\in[n]^2} \mathbb{P}\{|\hat{p}_{i,j} - p_{i,j}| > a^2/8\} \\
&\leq \sum_{(i,j)\in[n]^2} 2exp\left(-2N_{i,j}\frac{a^4}{64}\right) \\
&\leq 2n^2 exp\left(-2N\frac{a^4}{64}\right)
\end{aligned}$$

We demand $2n^2 exp\left(-2N\frac{a^4}{64}\right) \leq \epsilon \Leftrightarrow N \geq \frac{32log\left(\frac{2n^2}{\epsilon}\right)}{a^4}$.
If we suppose that queries are adaptive, then we can use an optimal number of comparisons (for example via Mergesort), so the total sample complexity will be $O(nlog(n) \cdot N)$. There are also more sophisticated methods to perform the estimation using noisy comparisons, e.g. Feige et al. in [49] and Davidson et al. in [50].

The weights and the spread parameter can be computed by solving the system
$p_1 = \frac{\varphi}{\varphi+1}, \quad p_2 = w_1 \cdot \frac{1}{\varphi+1} + w_2 \cdot \frac{\varphi}{\varphi+1}$. The system has a unique solution as long as p_1 is known, that is central rankings agree on some comparisons, which means that they are not reversals. p_1, p_2 are computed in the learning algorithm with an error tolerance of a^2. If greater precision is sought, then more samples should be used. Note that p_is could be calculated by aggregating the frequencies of different comparisons as long as they follow the same Bernoulli (this is detected by the clustering threshold with high probability). Then, tail bounds get tighter as more samples are used.

5.4 Learning Selective Mallows Mixtures-The General Case

5.4.1 The Effect of Selectivity On The Sample Complexity

In this section we try to perform parameter estimation of the Mallows Mixture, so identifiablity conditions must be satisfied. First we focus on the work of Mao et al. so we assume common spread parameters. To ensure identifiability we suppose all subsets of items with length $l = 2 \cdot \lfloor log_2(k) \rfloor + 3$, are p-frequent. This means that for each such set the

probability that either the set or a superset of it is selected is greater than p. We use the framework proposed by [6], where the "subroutine" aggregates samples in an empirical distribution and searches over a cover of the space of marginalised mixtures finds the correct marginalised mixture with high probability. This way it effectively simulates the noiseless oracle used by a meta-algorithm to inductively reconstruct the central permutation and approximate the corresponding weights with small error.

If the subroutine is proved to simulate the strong oracle, where both the marginalised centers and the corresponding weights are estimated, then the meta-algorithm performs optimal size queries, that fit the identifiablity conditions ($l = 2 \cdot \lfloor log_2(k) \rfloor + 3$). If the subroutine is only assumed to retrieve the marginalised centers with high probability and not the weights, then the query length grows and in [6] a meta-algorithm is proposed that uses queries of length l=2k-2 in the worst case, which is not tight with respect to the identiability conditions. However, Corollary 3 in [5] provides a stopping criterion, that if it were used by the subroutine it would guarantee that both the marginalised centers and the corresponding weights are ∂-close estimated.

The subroutine responds to the queries set by the meta-algorithm. Each query involves only a small subset of the items (0(log(k) or O(k) depending on the algorithm) so the samples are marginalised into this set of items. Even if samples are originally complete they are truncated in each call of the subroutine to match the queried subset. So the framework proposed in [6] is selection friendly. In particular, it is compatible with the version of the selective model that applies the selection mechanism after the (complete) Mallows sampling, that is selection that preserves the positions in the complete sample.

In the work of [6] and [5] sample complexity N is calculated on complete samples that are used in each call of the subroutine. In [6] N is found to be $poly_k(\frac{1}{1-\phi}, \frac{1}{\gamma}) \cdot log(n)$. In our setting the sample complexity is modified so that in each call of the subroutine enough samples of the corresponding subset are present. We assume that the required subsets are p-frequent so in each call of the subroutine we have to find an appropriate sample complexity N' so that N samples contain the queried subset. This is a case of binomial distribution, where p is the probability of success, N' is the total number of trials, N is the number of successful outcomes and the required probability of the event is set to $1 - \frac{\delta}{k \cdot n^2}$, because all calls of the subroutine must be successful and there are at most $k \cdot n^2$ such calls.

Let E_i be the event that less than N samples contain the subset J_i queried in i-th call of the subroutine and $p_i = \sum_{s \text{ contains } J_i} f(s)$. $\mathbb{P}[E_i] = \sum_{j=0}^{N-1} \binom{N'}{j} \cdot {p_i}^j \cdot (1-p_i)^{N'-j}$. Applying Hoeffding's inequality we yield $\mathbb{P}[E_i] \leq exp[-2N'(p_i - \frac{N}{N'})^2]$. But for all i $p_i \geq p$. Then $\mathbb{P}[E_i] \leq exp[-2N(p - \frac{N}{N'})^2]$, $\forall i$. $\mathbb{P}\{\text{each } J_i \text{ is contained in at least N samples }\} = 1 - \mathbb{P}\{\bigcup_i E_i\} \geq 1 - \sum_{i=1}^{k \cdot n^2} \mathbb{P}[E_i] \geq 1 - k \cdot n^2 \cdot exp[-2N'(p - \frac{N}{N'})^2]$. Setting $\delta = k \cdot n^2 \cdot exp[-2N'(p - \frac{N}{N'})^2] \Leftrightarrow N' = \frac{\sqrt{L^2 + 8 \cdot p \cdot N \cdot L} + L + 4 \cdot p \cdot N}{4p^2}, L = ln(\frac{k \cdot n^2}{\delta})$, we achieve our goal. $N' < \frac{\sqrt{8 \cdot p \cdot N \cdot L} + 2L + 4 \cdot p \cdot N}{4p^2} < \frac{2(2L + 4 \cdot p \cdot N)}{4p^2} =$

$O(\frac{N}{p} + \frac{L}{p^2})$, where $N = poly_k(\frac{1}{1-\phi}, \frac{1}{\gamma}) \cdot log(n)$, using the results of [6].

The above results simulate the noiseless oracle with the Subroutine function of [6]. The problem is that this function requires marginalised samples to preserve the information of the position of each item in the complete ranking. We are more interested in the selective mechanism that is bijective. In this case, we would use the complete learning algorithm (of Mao et al. or Moitra et al.) to simulate the noiseless oracle. If we assume common spread parameters we use the algorithm of Mao et al. (see theorem 4.3.5) and invoke it for a total of $k \cdot n^2$ selection sets with length logarithmic on k. The error probability tolerance in theorem 4.3.5 is set to $2n^{-10}$. We will demand it to be $k \cdot n^2$ times smaller, so that from union bound we can ensure that all oracle calls will be successful. Even with the new error probability tolerance, the sample complexity for each call remains $poly_k(\frac{1}{1-\phi}, \frac{1}{\gamma}) \cdot log(n)$. We assume that selection sets of length $2log(k) + 3$ are p-frequent. To ensure that for each queried subset the required number of samples will be available, it suffices to have a total sample complexity which is $O\left(\frac{poly_k(\frac{1}{1-\phi}, \frac{1}{\gamma}) \cdot log(n)}{p} + \frac{L}{p^2}\right)$, where $L = ln(\frac{k \cdot n^2}{\delta})$ as shown earlier. This way, with probability at least 1-δ, the learning algorithm succeeds.

If we assume that spread parameters are not known and not equal to each other, then we will use the algorithm of Moitra et al. to simulate the noiseless oracle (we use the algorithm given in 4.2.1, because the main algorithm has an unnecessary demand that $n > 10k^2$). The algorithm in 4.2.1 demands that the number of items is at least equal to $10k$. Thus, selection sets will have to be at least that long. Having selection sets this long, we can use the algorithm in 4.3.2, which has the advantage to not depend on mixing weights estimations. The algorithm has to be slightly modified because of the non common spread parameters. It will use the pairwise comparisons signature to distinguish distinct centers but it will also have to use spread parameters as signatures for concentric components. Concentric components might arise that can not be merged as in the case of equal spread parameters. To distinguish these components we have to use the spread parameter estimations as a signature. Thus, the spread parameter estimations have to be accurate to avoid confusion between different concentric components. In particular, the additive error for each spread parameter should be at most $\Delta\phi/4$, where $\Delta\phi = min_{i \neq j}\{|\phi_i - \phi_j|\}$. We can build the centers inductively in the logic of algorithm 4.3.2 with the requirement that in each query the returned marginalised centers are correct and the additive error for each spread parameter is at most $\Delta\phi/4$. Weights should be correct (up to a small additive ∂) only in one query, as the reconstruction algorithm does not depend on them. We decrease the error probability of the algorithm of Moitra et al. from δ to $\frac{\delta}{kn^2}$, because it will be invoked at most kn^2 times and this way a union bound ensures that, with probability at least $1 - \delta$, all queries will be successful. We also set the additive error equal to $min\{\Delta\phi/4, \partial\}$ to ensure that the user-defined additive error tolerance ∂ will be achieved and the implementation defined precision $\Delta\phi/4$ for the spread parameters will be achieved as well. Let $\mu = min\{min_{i \neq j}\{|\phi_i - \phi_j|\}, min_i\{|\phi_i - 1|\}\}$. Then, the sample complexity for each queried selection set is $poly_k(n, \frac{1}{\mu}, \frac{1}{\gamma}, \frac{1}{min\{\Delta\phi/4, \partial\}}, log(\frac{kn^2}{\delta})) = poly_k(n, \frac{1}{\mu}, \frac{1}{\gamma}, \frac{1}{\partial}, log(\frac{1}{\delta}))$,since $\Delta\phi < \mu$,

where μ is the non degeneracy condition. We assume that selection sets of length $10k+3$ are p-frequent.To ensure that for each queried subset the required of samples will be available, it suffices to have a total sample complexity which is $O\left(\frac{poly_k(n,\frac{1}{\mu},\frac{1}{\gamma},\frac{1}{\partial},log(\frac{1}{\delta}))}{p}+\frac{L}{p^2}\right)$, where $L=ln(\frac{k\cdot n^2}{\delta})$ as shown earlier. With this sample complexity we can learn the centers exactly and the weights and spread parameters up to an additive error ∂ with probability at least $1-\delta$, assuming that selection sets of length $10k+3$ are p-frequent.

5.4.2 Learning the Selective Mallows Mixture Model in TV Distance

In this chapter we define an empirical distribution based on selective samples drawn from the latent Mallows Mixture. Then we find the sample complexity needed to bound the TV distance between the empirical and the latent model.

We begin with the definition of the selective empirical model. One option is to keep the definition of [6], which is $M_N(\pi)=\frac{1}{N}\sum_{i=1}^{N}\mathbb{1}\{\pi=\sigma_i\}$. The argument π as well as the samples σ_i can be incomplete. Let S be the support set of selection mechanism f(s). For each s in S we define N_s as the number of samples that are permutations of the elements of s. Then M_N can be analysed into selections sets as follows: $M_N(\pi)=\sum_{s\in S}\frac{N_s}{N}\cdot\frac{1}{N_s}\sum_{i:set(\sigma_i)=s}\mathbb{1}\{\pi=\sigma_i\}$. The factor $\frac{N_s}{N}$ accounts for the term f(s) in the density of the Selective Mallows Mixture Model and quantity $M_{N_s}|_s=\frac{1}{N_s}\sum_{i:set(\sigma_i)=s}\mathbb{1}\{\pi=\sigma_i\}$ is equal to the empirical distribution of the complete (non selective) marginal of the latent mixture on set s. $M_{N_s}|_s$ is compatible with the definition of the empirical mixture model given in [6], so the results of Proposition 3.3 and Theorem 3.4 in [6] can be applied for $M_{N_s}|_s$, for all s in S.

Now we will find an analytical expression for the TV distance between the empirical model consisting of samples and the latent model that generated these samples.
$2TV(M,M_N|_J)=\sum_{\sigma\in S_{n,J}}\mid M(\sigma)-M_N|_J(\sigma)\mid=\sum_{s\in S}\sum_{\sigma\in S_{n,J\cap s}}\mid M(\sigma)-M_N|_{J\cap s}(\sigma)\mid=$
$\sum_{s\in S}\sum_{\sigma\in S_{n,J\cap s}}\mid f(J\cap s)\cdot M|_{J\cap s}(\sigma)-\frac{N_{J\cap s}}{N}\cdot\frac{1}{N_{J\cap s}}\sum_{i:set(\sigma_i)=J\cap s}\mathbb{1}\{\sigma=\sigma_i\}\mid$
By increasing the sample complexity N the TV distance between the latent model and its empirical decreases. In particular quantity $\frac{N_{J\cap s}}{N}$ approximates $f(J\cap s)$ and $M|_{J\cap s}(\sigma)$ approximates $\frac{1}{N_{J\cap s}}\sum_{i:set(\sigma_i)=J\cap s}\mathbb{1}\{\sigma=\sigma_i\}$.

If we suppose that the selection mechanism f is known, then the empirical distribution can be defined as a function of f as follows : $M_N(\pi)=\sum_{s\in S}f(s)\cdot\frac{1}{N_s}\sum_{i:set(\sigma_i)=s}\mathbb{1}\{\pi=\sigma_i\}$.
Then $2TV(M,M_N|_J)=\sum_{\sigma\in S_{n,J}}\mid M(\sigma)-M_N|_J(\sigma)\mid=$
$\sum_{s\in S}f(J\cap s)\cdot\sum_{\sigma\in S_{n,J\cap s}}\mid M|_{J\cap s}(\sigma)-\frac{1}{N_{J\cap s}}\sum_{i:set(\sigma_i)=J\cap s}\mathbb{1}\{\sigma=\sigma_i\}\mid$
We will use Proposition 3.3 of [6] to bound the TV distance on each subset $J\cap s$. According to this proposition $\mathbb{P}\{TV(M|_J,M_N|_J)>d\}\leq exp\left(-N\frac{3d}{10}\right)+2(2kq)^l\cdot exp\left(-N\frac{d^2}{(2kq)^{2l}}\right)$, where $l=|J|$ and $q=1+\frac{1}{1-\phi}log\left(\frac{8l}{d(1-\phi)}\right)$. By setting $N=N_0(d,\epsilon)=max\left\{\frac{10log(\frac{2}{\epsilon})}{3d},\frac{4^l(kq)^{2l}log(2^{l+2}(kq)^l/\epsilon)}{d^2}\right\}$ we achieve $\mathbb{P}\{TV(M|_J,M_N|_J)>d\}\leq\epsilon$.

Let N_D be the number of selection sets s such that $J\cap s\neq\emptyset$. For each such set we demand that $\mathbb{P}\{TV(M|_{J\cap s},M_{N_{J\cap s}}|_{J\cap s})>\frac{d}{N_D\cdot f(J\cap s)}\}\leq\epsilon/N_D$. Then by union bound we achieve

to bound the TV distance between the selective mixture model and its empirical with high probability: $\mathbb{P}\{TV(M, M_N|_J) \leq d\} \geq 1-\epsilon$. To meet the demand for each selection set we need $N_{J\cap s} \geq N_0\left(\frac{d}{N_D \cdot f(J\cap s)}, \frac{\epsilon}{N_D}\right)$. The number of samples drawn from each selection set follows binomial distribution $B(N, f(J \cap s))$, where N is the total sample complexity and $f(J \cap s)$ is the probability of selection of the set. We want that with high probability $(1 - \delta)$ all subsets are adequately represented in the samples. It suffices to demand for each subset that the probability it is underrepresented is less that δ/N_D.

Let E_i be the event that less than $N_1(s_i) = N_0\left(\frac{d}{N_D \cdot f(J\cap s_i)}, \frac{\epsilon}{N_D}\right)$ samples contain the subset s_i, $i \in [N_D]$ and $p_i = f(J \cap s_i)$.
$\mathbb{P}[E_i] = \sum_{j=0}^{N_1(s_i)-1} \binom{N}{j} \cdot {p_i}^j \cdot (1 - p_i)^{N-j}$. $\mathbb{P}\{$each s_i is contained in at least $N_1(s_i)$ samples $\}$ = $1 - \mathbb{P}\{\bigcup_i E_i\} \geq 1 - \sum_{i=1}^{N_D} \mathbb{P}[E_i]$.
In order to bound this probability from below by $1 - \delta$ we demand that $\mathbb{P}[E_i] \leq \frac{\delta}{N_D}$ for all $i \in [N_D]$. Hoeffding bounds for binomial variables E_i yield $\mathbb{P}[E_i] \leq exp\left[-2N\left(p_i - \frac{N_1(s_i)}{N}\right)^2\right]$. Thus, it suffices to demand $exp\left[-2N\left(p_i - \frac{N_1(s_i)}{N}\right)^2\right] \leq \frac{\delta}{N_D}$ for all $i \in [N_D]$. This is equivalent to $N \geq \frac{\sqrt{L^2+8\cdot p_i\cdot N_1(s_i)\cdot L}+L+4\cdot p_i\cdot N_1(s_i)}{4{p_i}^2}, L = ln\left(\frac{N_D}{\delta}\right)$, for all $i \in [N_D]$. **(1)**.
The RHS of inequality **(1)** is $O\left(\frac{N_1(s_i)}{p_i} + \frac{L}{{p_i}^2}\right)$. Thus, N is $O\left(max_{i\in[N_D]}\left\{\frac{N_1(s_i)}{p_i} + \frac{L}{{p_i}^2}\right\}\right)$.
We substitute $N_1(s_i)$ with its formula and turn the *max* operator in the formula of $N_0(d, \epsilon)$ into a summation because adding two quantities is asymptotically the same as taking the maximum of them.
Then, sample complexity N is $O\left(max_{i\in[N_D]}\left\{\frac{log\left(\frac{2N_D}{\epsilon}\right)\cdot N_D}{d} + \frac{4^l(kq)^{2l}log\left(2^l(kq)^l N_D/\epsilon\right)\cdot N_D^2\cdot p_i}{d^2} + \frac{log\left(\frac{N_D}{\epsilon}\right)}{{p_i}^2}\right\}\right)$.

In this work, as well as in the literature, the length l of selection sets and the number k of distinct centers are supposed to be small. As a result, N_D, the number of possible selection sets is $poly(n)$ and polynomial quantities raised to the power of l remain polynomial. Thus, sample complexity N is $poly(n, \frac{1}{\epsilon}, \frac{1}{d}, \frac{1}{p})$, where n is the number of items, d is the TV-distance error margin, ϵ is the error probability margin and selection mechanism $f(s)$ is assumed to be p-frequent.

5.4.3 Sample Grouping vs Parameter Cover

All known methods for learning the Mallows Mixture in the general case involve some kind of exhaustive search over candidate models. For each candidate a criterion is applied that compares the candidate model with the latent mixture model in terms of TV distance. Because the parameters of the hidden model are unknown, the TV distance is calculated between empirical models, constructed from samples of the latent model and synthetic samples generated from the candidate model. There are theoretical guarantees that if the candidate and the latent model are close in terms of the empirical TV distance, then they are also close in terms of their parameters. One way to generate a set of candidate models that includes some model that is appropriately close to the latent model is to perform a cover over the space of k-mixtures, considering all possible combinations of central per-

mutations and gridsearching over the mixing weights using a step that is determined by the sensitivity we aim at. Here we propose an alternative method of exhaustive search that combines samples rather than candidate parameters.

This method has some advantages compared to the method of cover. It directs the search only to candidate centers than are implied from the samples. For example, if no samples contain the comparison $e_i < e_j$, then no candidate central permutation containing this comparison will be considered. Moreover it effectively parameterizes the size of the search space in terms of the spread parameter ϕ and the minimal weight γ. If ϕ is small or γ is big, then the search space is decreased, taking these parameters into consideration. On the other hand the cover fails to adapt to these parameters and searches over the same space irrespective of them. The proposed method also features a better dependency on the number of items n that is $O\left(log(n)^{k\frac{log(k\cdot n/\epsilon)}{(1-\phi)^2}}\right)$ compared to that of the cover method that is $n!^k$.

We will now describe the proposed method of generating candidate models.Using the results of Caragiannis et al (2013) we have that given $O\left(\frac{log(k\cdot n/\epsilon_3)}{(1-\phi)^2}\right)$ samples from a mallows model on $\mathbb{S}_n$ we can retrieve its central ranking with probability at least $1-\epsilon$ using a positional estimator. The time complexity of the pos. est. is $O(r\cdot n^2)$ where r is the number of samples drawn from the mallows model.Suppose we draw N (complete) samples from a mallows mixture model and the number r_i of samples drawn from the i-th cluster, $i \in [k]$, is at least equal to some value r which is $O\left(\frac{log(k\cdot n/\epsilon_3)}{(1-\phi)^2}\right)$. Then we could perform an exhaustive search over all possible k-tuples of disjoint subsets of the samples of length r. For each such k-tuple we will assume that its i-th element is a set that contains samples drawn the same cluster and we will try to retrieve each cluster via a positional estimator. We will assume that this holds $\forall i \in [k]$ and different sets contain samples from different clusters. For some k-tuple this assumption will be true and then with probability $1-\mathbb{P}\{\bigcup_{i=1}^{k}\pi_i \text{ is wrongly estimated}\} \geq 1-k\cdot\epsilon/k = 1-\epsilon$ all central permutations of the mixture will be correctly estimated.

It remains to find a value for N, such that with probability at least $1-\delta$, the number r_i of samples drawn from the i-th cluster, is at least equal to r for all clusters and r is $O\left(\frac{log(k\cdot n/\epsilon_3)}{(1-\phi)^2}\right)$. Let E_i be the event that less than r samples are drawn from cluster i. $\mathbb{P}[E_i] = \sum_{j=0}^{r-1}\binom{N}{j}\cdot {w_i}^j\cdot(1-w_i)^{N-j}$. Applying Hoeffding's inequality we yield $\mathbb{P}[E_i] \leq exp\left[-2N(w_i-\frac{r}{N})^2\right]$. Let γ be the weight w_i that maximizes the quantity $exp\left[-2N(w_i-\frac{r}{N})^2\right]$. Then $\mathbb{P}[E_i] \leq exp\left[-2N(\gamma-\frac{r}{N})^2\right]$, $\forall i \in [k]$. $\mathbb{P}\{\text{at least r samples are drawn from each cluster}\} = 1-\mathbb{P}\{\bigcup_{i=1}^{k}E_i\} \geq 1-\sum_{i=1}^{k}\mathbb{P}[E_i] \geq 1-k\cdot exp\left[-2N(\gamma-\frac{r}{N})^2\right]$. Setting $\delta = k\cdot exp\left[-2N(\gamma-\frac{r}{N})^2\right] \Leftrightarrow N = \frac{\sqrt{L^2+8\cdot\gamma\cdot r\cdot L}+L+4\cdot\gamma\cdot r}{4\gamma^2}, L = ln\left(\frac{k}{\delta}\right)$, we achieve our goal. $N < \frac{\sqrt{8\cdot\gamma\cdot r\cdot L}+2L+4\cdot\gamma\cdot r}{4\gamma^2} < \frac{2(2L+4\cdot\gamma\cdot r)}{4\gamma^2} = O\left(\frac{r}{\gamma}+\frac{L}{\gamma^2}\right)$.

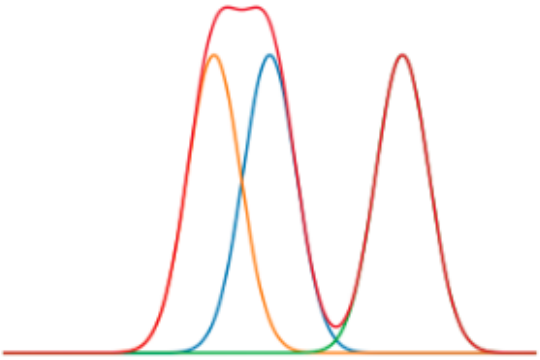
We want the proposed algorithm to retrieve the set of central permutations of the mixture with probability at least $1-\epsilon_0 \Leftrightarrow \mathbb{P}\{\text{algorithm fails}\} \leq \epsilon_0$. The algorithm could fail

either because there are not enough samples in some cluster (less than r) or because some estimator fails despite at least r samples are drawn. Summing over those two cases we have: $\mathbb{P}\{\text{algorithm fails}\} \leq \delta + (1-\delta)\cdot\epsilon$. We could set $\delta = \epsilon = \epsilon_0/2$ and achieve our goal.

Now we will analyse the time complexity of this algorithm. The number of all possible k-tuples of disjoint subsets of length r is equal to $\frac{N!}{(N-r)!r!} \cdot \frac{(N-r)!}{(N-2r)!r!} \cdot \ldots \frac{(N-(k-1)r)!}{(N-kr)!r!} = \frac{N!}{(N-kr)!\cdot(r!)^k} = \frac{(N-kr+1)\cdot(N-kr+2)\cdot\ldots N}{(r!)^k} < \frac{N^{kr}}{(r!)^k}$. For every such k-tuple we apply k positional estimators. The pos. est. takes $O(r\cdot n^2)$ time, so the total time complexity is $O\left(k\cdot r\cdot n^2\cdot\frac{N^{kr}}{(r!)^k}\right)$.

5.5 Learning Separable Mallows Mixture Models

5.5.1 Learning Clusters Based On Empirical Modes



We assume that a separation condition of the form $(d_{KT}(\pi_i,\pi_j) > a, \forall i \neq j)$ is satisfied for the the central permutations of the mixture. We will try to detect the central permutations by looking at the modes of the empirical distribution (the local maxima of the empirical pdf). We know that the pdf of a mallows model is maximised at its center. We want to find a value for a such that $\forall j \in [k]\ \ w_j\cdot\phi^{d_{KT}(\pi_j,\pi)} - \sum_{i\neq j} w_i\cdot\phi^{d_{KT}(\pi_i,\pi)} > p(d)$, for π in the neighbourhood of π_j, that is π st $d_{KT}(\pi_j,\pi) \leq d$. This condition implies that the density of a component of the mixture in an area close to the center of the component (called the neighbourhood of that center) is significantly higher that the total density of all the other components summed in this particular area. Their difference is a function of d, the distance from the center of the component. As d increases the rest of the components may dominate over the single component. We will now find a lower bound $p(d)$ for the density domination of a single component over the rest of the components in the neighbourhood of width d of the single component.

From the triangle inequality we have $d_{KT}(\pi_i,\pi) \geq d_{KT}(\pi_i,\pi_j) - d_{KT}(\pi_j,\pi) > a - d_{KT}(\pi_j,\pi)$. So we obtain the following lower bound for the density domination of j-th component: $w_j\cdot\phi^{d_{KT}(\pi_j,\pi)} - \sum_{i\neq j} w_i\cdot\phi^{d_{KT}(\pi_i,\pi)} > w_j\cdot\phi^{d_{KT}(\pi_j,\pi)} - \sum_{i\neq j} w_i\cdot\phi^a\cdot\phi^{-d_{KT}(\pi_j,\pi)}$. **(1)**

But the RHS of **(1)** is equal to $w_j\cdot\phi^{d_{KT}(\pi_j,\pi)} - \phi^a\cdot\phi^{-d_{KT}(\pi_j,\pi)}\cdot\sum_{i\neq j} w_i =$
$w_j\cdot\phi^{d_{KT}(\pi_j,\pi)} - \phi^a\cdot\phi^{-d_{KT}(\pi_j,\pi)}\cdot(1-w_j) > w_j\cdot\phi^d + \phi^a\cdot\phi^{-d}\cdot(w_j-1) > \gamma\cdot\phi^d + \phi^a\cdot\phi^{-d}\cdot(\gamma-1)$, for π in the neighbourhood of π_j (we name γ the minimal weight of the mixture).

Returning to **(1)** we obtain $w_j \cdot \phi^{d_{KT}(\pi_j,\pi)} - \sum_{i \neq j} w_i \cdot \phi^{d_{KT}(\pi_i,\pi)} > \gamma \cdot \phi^d + \phi^a \cdot \phi^{-d} \cdot (\gamma - 1)$, for π in the neighbourhood of π_j. We set $p(d) = \gamma \cdot \phi^d + \phi^a \cdot \phi^{-d} \cdot (\gamma - 1)$.

We want local maxima to only exist at the neighbourhoods of the centers. We express this as follows: for all π st $\frac{a}{2} > d_{KT}(\pi_{j^*}, \pi) > d$, where j^* is the central permutation closet to π, it holds that $\sum_{i=1}^{k} w_i \cdot \phi^{d_{KT}(\pi_i,\pi)} < min_{\{\pi: d_{KT}(\pi_{j^*},\pi) \leq d\}}(\sum_{i=1}^{k} w_i \cdot \phi^{d_{KT}(\pi_i,\pi)})$.

For π st $\frac{a}{2} > d_{KT}(\pi_{j^*}, \pi) > d$, where j^* is the central permutation closet to π, we have that $\sum_{i=1}^{k} w_i \cdot \phi^{d_{KT}(\pi_i,\pi)} < w_{j^*} \cdot \phi^{d_{KT}(\pi_{j^*},\pi)} + \sum_{i \neq j^*} w_i \cdot \phi^a \cdot \phi^{-d_{KT}(\pi_{j^*},\pi)} \leq w_{j^*} \cdot \phi^{d+1} + \sum_{i \neq j^*} w_i \cdot \phi^a \cdot \phi^{-a/2} = w_{j^*} \cdot \phi^{d+1} + (1 - w_{j^*}) \cdot \phi^{a/2}$.

We also have that $min_{\{\pi': d_{KT}(\pi_{j^*},\pi') \leq d\}}(\sum_{i=1}^{k} w_i \cdot \phi^{d_{KT}(\pi_i,\pi')}) > w_{j^*} \cdot \phi^d + (1 - w_{j^*}) \cdot \phi^{a+d}$.

For values of a such that $w_{j^*} \cdot \phi^{d+1} + (1 - w_{j^*}) \cdot \phi^{a/2} \leq w_{j^*} \cdot \phi^d + (1 - w_{j^*}) \cdot \phi^{a+d} \Leftrightarrow \phi^d((\phi^a + \phi - 1)w_{j^*} - \phi^a) \leq \phi^{a/2}(w_{j^*} - 1)$ **(2)** the local maxima requirement is satisfied.
(2) can only be true if $(\phi^a + \phi - 1)w_{j^*} - \phi^a < 0$, because the RHS of the inequality is negative and ϕ^d is positive. But this constraint is always satisfied because $(\phi^a + \phi - 1)w_{j^*} - \phi^a = (w_{j^*} - 1)\phi^a + (\phi - 1)w_{j^*} < 0$, as $0 < w_{j^*} < 1$, $0 < \phi < 1$.
Thus, **(2)** is equivalent to $\phi^d \geq \frac{\phi^{a/2}(w_{j^*}-1)}{(\phi^a+\phi-1)w_{j^*}-\phi^a}$. **(2′)**
The greatest possible value of ϕ^d is 1, so a must be such that $1 \geq \frac{\phi^{a/2}(w_{j^*}-1)}{(\phi^a+\phi-1)w_{j^*}-\phi^a} \Leftrightarrow (\phi^a + \phi - 1)w_{j^*} - \phi^a \leq \phi^{a/2}(w_{j^*} - 1) \Leftrightarrow$
$0 \leq (1 - w_{j^*})\phi^a + (w_{j^*} - 1)\phi^{a/2} + (1 - \phi)w_{j^*}$. **(3)**
This is a quadratic expression of $\phi^{a/2}$. If the determinant is negative, then the expression is always positive and **(3)** is satisfied.
If the determinant is non negative, then two solutions exist for the corresponding quadratic equation. $\phi^{a/2}$ should either be above the greater of the two solutions or below the smaller one. In the second case a restriction of the form $a > a_{min}$ arises. If we tighten the restriction in **(2′)** by considering values of d greater than zero then the restriction for a becomes more strict (a_{min} increases).

At this point we are going to introduce the observations about the modes into the framework proposed in [6]. The learning algorithm that uses noiseless queries to the "weak oracle" can be used to reconstruct the central permutations as long as there is some way to simulate the noiseless "weak oracle" using noisy samples. In the general case the weak oracle is simulated using an exhaustive algorithm that checks all possible candidate models and selects the one that is closest in TV distance to the available samples. This computationally expensive procedure could be bypassed if a simple criterion could be applied to detect target subpermutations.
In particular we would like the probability mass of the mixture to be higher at samples that are fully concordant to one of the central rankings than at samples that do not fully agree

with neither of the central rankings. The queries used by the weak oracle contain a "signature" set of items, such that all centers (marginalised on items $[1, 2..., n]$ have different images on this set. As a result, it is guaranteed that at most one of the marginalised centers could agree with some sample on the queried set. The query set also includes three more items that are used to infer the position of an item in a latent marginalised center isolated by the "signature". Two of these items are consecutive in the target (marginalised) center and the third is a new item (item id equals $n + 1$) that is checked as to whether it can be placed in between the two consecutives in some latent center extending the marginalised one. If the guess of the position of the new item is correct, then a sample is produced that is totally concordant to a latent center on items $[1, 2, .., n+1]$. The sample is a permutation of the items in set $J = \{\textit{items in signature set}\} \cup \{\textit{the two consecutive items}\} \cup \{\textit{item } n+1\}$. Without a condition about the spread parameters we need to check all samples supported on set J in order to effectively simulate the weak oracle (we have to learn the marginal mixture on set J). However, if the spread parameters are small enough, then we can only check the probability mass on the sample that consists of the signature ranking, the two consecutive items in the correct relevant position to the signature and the new item $(n+1)$ placed between the consecutive items. If the probability mass on this sample is above some threshold, then the guess is correct and the position of item $(n+1)$ in a latent center has been learned. Now we will formulate the condition for the spread parameters and the threshold.

Suppose that we know that there is some constant a such that $\phi < a$ and $w_i > a$ for all $i \in [k]$. Also suppose that this constant is known to us.

• For the samples π^* that contain a correct guess for the position of item $n + 1$ the (theoretical) probability mass is $M(\pi^*) = w_{i^*} + \sum_{i \neq i^*} w_i \cdot \phi^{d_i}$, where i^* is the index of the center for which we made the correct guess. π^* is in total agreement with this center, so their KT-distance is zero and the corresponding term of the mixture is equal to w_{i^*}, which is its maximal value. The other terms are equal to $w_i \cdot \phi^{d_i}$, where d_i is at least one, because of the disagreements on the signature ranking. Ignoring these terms we have $M(\pi^*) > w_{i^*} > a$.
• For the samples π that do not contain a correct guess for the position of item $n + 1$ the (theoretical) probability mass is $M(\pi) = \sum_i w_i \cdot \phi^{d_i}$, where each d_i is at least one, either due to disagreement on the signature ranking or because of incorrect guess of the position of item $n + 1$. Thus, we have $M(\pi) \leq \sum_i w_i \cdot \phi < \sum_i w_i \cdot a = a$

Using the above observation we could set the threshold equal to a and decide that a guess is correct iff the frequency of the corresponding sample is at least equal to a. The absolute difference between the sample frequency and the theoretical probability mass decreases exponentially to the number of samples due to Hoeffding's inequality, so using enough samples and assuming that there is a non zero gap g between $min\{w_i\}$ and ϕ the above greedy rule is correct with high probability. In particular, if we have at

least $log(2/\epsilon)/(2g^2)$ samples from some selection set that contains the signature, the two consecutive items and item $n+1$ then with probability at least $1-\epsilon$ the greedy rule *frequency* $>?$ $a/Z(|J|)$ for finding the correct position of item $n+1$ in some latent center in $[n+1]$ works correctly. We assume that queries are adaptive. The greedy rule has to be applied $O(n^2 \cdot k)$ times so for an overall error tolerance ϵ the sample complexity should be $log(2n^2 \cdot k/\epsilon)/(2g^2)$ for the subset used in each query, so $n^2 \cdot k \cdot log(2n^2 \cdot k/\epsilon)/(2g^2)$ in total.

Note that this analysis is relevant to the selective Mallows Mixture setting, because the required samples have length $J = O(k)$ and because longer samples can also be used, as long as there are enough samples from the corresponding selection set. The low spread parameter condition helps us avoid costly histogram approximation methods and bridge the gap between the single and the mixture Mallows learning. The majority rule on pairwise comparisons is replaced by a "dominance" rule on samples that are subpermutations of some latent central ranking.

5.5.2 Clustering Algorithm for Learning Separable Mallows Mixtures and Conditions for the Success of the Algorithm

We assume that a separation condition of the form $(d_{KT}(\pi_i, \pi_j) > a, \forall i \neq j)$ is satisfied for the the central permutations of the mixture. We have a set Π of N complete samples drawn from a mixture of k Mallows models. We aim to divide them into k groups (clusters), such that permutations that belong to the same group come from the same component of the mixture. Firstly, we propose a simple clustering algorithm and analyse the probability of success of the algorithm. A sufficient separation condition is provided that guarantees the success of the clustering algorithm with high probability. Then we use the clustered samples to estimate the central rankings of the mixture. The required sample complexity for this estimation is calculated.

Clustering Algorithm For Separable Mallows Mixtures

```
clusters=[[Π[0]]]
threshold=a/2
k=1
for n in [1, N − 1] :
  π = Π[n]
  for i in [0, k − 1] :
    choose a random element σ_i in cluster i
    d_i := D_KT(σ_i, π)
  if min{d} >threshold:
    k+=1
    clusters.append([π])
  else:
    clusters[argmin{d}].append(π)
```

We suppose that the above algorithm has performed a correct clustering of the samples 1,2 .. n-1. We will study the probability of error on sample $\pi = \Pi[n]$.

If π comes from a cluster j already seen ($j \leq k$) we have:
$D_{KT}(\pi, \sigma_j) \leq D_{KT}(\pi_j, \sigma_j) + D_{KT}(\pi_j, \pi)$, where π_j is the latent central permutation of cluster j.
$\mathbb{P}[D_{KT}(\pi_j, \sigma_j) \geq a/2] = \sum_{d=a/2}^{\frac{n(n-1)}{2}} \frac{A(n,d)\cdot\phi^d}{Z(\phi)} \leq \sum_{d=a/2}^{\frac{n(n-1)}{2}} \frac{A(n,d)\cdot\phi^{a/2}}{Z(\phi)} = \frac{\phi^{a/2}}{Z(\phi)} \cdot \sum_{d=a/2}^{\frac{n(n-1)}{2}} A(n,d) \leq \frac{\phi^{a/2}}{Z(\phi)} \cdot n!$
Similarly we have $\mathbb{P}[D_{KT}(\pi_j, \pi) > a/2] < \frac{\phi^{a/2}}{Z(\phi)} \cdot n!$,
so $\mathbb{P}[D_{KT}(\pi_j, \sigma_j) + D_{KT}(\pi_j, \pi) > a/2] < 2\frac{\phi^{a/2}}{Z(\phi)} \cdot n!$. We want $D_{KT}(\pi, \sigma_j) \leq a/2$ with probability greater than $1 - \delta/k$, so we demand $\delta/k > 2\frac{\phi^{a/2}}{Z(\phi)} \cdot n! \Rightarrow a > a_{min} = 2 \cdot \frac{\log\left(\frac{\delta \cdot Z(\phi)}{2k \cdot n!}\right)}{\log(\phi)}$.
We also have that $D_{KT}(\pi, \sigma_{j'}) \geq D_{KT}(\pi_{j'}, \pi_j) - D_{KT}(\pi_{j'}, \sigma_{j'}) - D_{KT}(\pi_j, \pi) \geq a - D_{KT}(\pi_{j'}, \sigma_{j'}) - D_{KT}(\pi_j, \pi)$ and with probability at least $1 - \delta/k$ it holds that $D_{KT}(\pi, \sigma_{j'}) > a/2$=threshold, where j' are the cluster ids of the other clusters seen so far ($j' \neq j$).
The algorithm could perform a misclassification of π either by creating a new cluster containing π or by assigning it to a false existing cluster (there are at most k-1 such clusters). Taking a union bound over all the error events we have that with probability at least $1 - \delta$ the algorithm assigns π to its correct cluster.

If π comes from a cluster j that has not been seen by the algorithm so far ($j = k$) we have:
$D_{KT}(\pi, \sigma_{j'}) \geq D_{KT}(\pi_{j'}, \pi_k) - D_{KT}(\pi_{j'}, \sigma_{j'}) - D_{KT}(\pi_k, \pi) \geq a - D_{KT}(\pi_{j'}, \sigma_{j'}) - D_{KT}(\pi_k, \pi)$ and with probability at least $1 - \delta/k$ it holds that $D_{KT}(\pi, \sigma_{j'}) > a/2$=threshold, $j' < k$. Again with union bound over all j' we get that with probability at least $1 - \delta$ the algorithm correctly creates a new cluster containing π.

$P_n = \mathbb{P}$\{no errors in the first n iterations\} =
$\mathbb{P}$\{no errors in the first n iterations\}$\cdot\mathbb{P}$\{ no error at iteration n | no errors in the first n-1 iterations\} $\geq$
$(1 - \delta)P_{n-1}$
$P_0 = 1$
$P_n \geq (1 - \delta)^n$
$\mathbb{P}$\{clustering algorithm succeeds \} $= P_{N-1} \geq (1 - \delta)^{N-1}$
We want $\mathbb{P}$\{clustering algorithm succeeds \} $\geq 1 - \epsilon_2$ so we set $1 - \epsilon_2 = (1 - \delta)^{N-1} \Leftrightarrow \delta = 1 - (1 - \epsilon_2)^{\frac{1}{N-1}}$.
$a_{min} = 2 \cdot \frac{\log\left(\frac{\delta \cdot Z(\phi)}{2k \cdot n!}\right)}{\log(\phi)} = 2 \cdot \frac{\log\left(1-(1-\epsilon_2)^{\frac{1}{N-1}}\right)+\log(Z(\phi))-\log(2k \cdot n!)}{\log(\phi)}$

Once we finish the clustering of the samples we apply the positional estimator on each cluster to estimate the central permutations of the mixture.

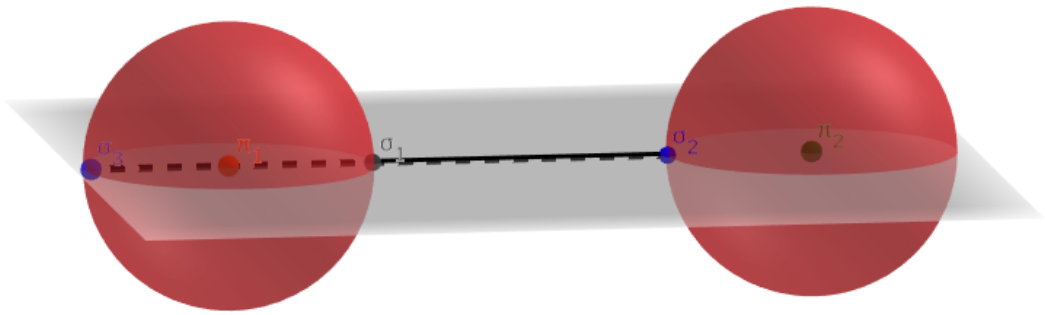
• By setting the value for N equal to $\frac{\sqrt{L^2+8\cdot\gamma\cdot r\cdot L}+L+4\cdot\gamma\cdot r}{4\gamma^2} = O\left(\frac{r}{\gamma} + \frac{L}{\gamma^2}\right), L = ln\left(\frac{k}{\epsilon_1}\right), \gamma = min_i\{w_i\}$ we ensure that with probability at least $1 - \epsilon_1$ the number r_i of samples drawn from the i-th cluster, is at least equal to r for all clusters (it has been proved previously in this

work).

• If r is at least equal to some value which is $O\left(\frac{log(k \cdot n/\epsilon_3)}{(1-\phi)^2}\right)$, and we have at least r correctly labeled samples from each cluster, then with probability at least $1 - \frac{\epsilon_3}{k}$ i-th center is correctly estimated and thus, from union bound, with probability at least $1 - \epsilon_3$ all central permutations of the mixture are correctly estimated.

• If $a \geq 2 \cdot \frac{log\left(1-(1-\epsilon_2)^{\frac{1}{N-1}}\right)+log(Z(\phi))-log(2k \cdot n!)}{log(\phi)}$, then with probability at least $1-\epsilon_2$ the clustering algorithm successfully partitions the set of N samples into their correct clusters.

The above analysis focuses on the probability of success in each step of the algorithm. This analysis is too detailed for the simple algorithm we proposed. A more elegant analysis will be made in the next paragraph . However the above analysis would make more sense in some other more complicated version of the algorithm. For example, an improvement of the algorithm would be to construct an estimator of the latent center of each cluster using the samples that have been assigned to the cluster so far. Then, new samples would be compared to the estimations of the centers rather than random samples from the cluster. Supposing that no (or few) misclassifications have been made at the first t steps of the algorithm, the expected distance between each estimator and its corresponding center falls as t increases, so the probability of misclassification of the remaining samples decreases. The above analysis could take this decrease into consideration and be useful in such a scenario. However for the simple algorithm we proposed the following simpler analysis is more suitable.



An alternative way to guarantee the success of the algorithm at each step is to demand that with high probability all samples lie within a radius equal to a/4 around their corresponding central permutation. This is a sufficient condition for the success of the clustering algorithm because the distance between points of the same cluster is bounded above by $2d < a/2$ and the distance between points that belong to different clusters is bounded below by $a - 2d > a/2$. Let π be a sample generated by center π_i. We want that with probability at least $1 - \frac{\epsilon_2}{N}$ distance $D_{KT}(\pi, \pi_i)$ is not greater than $\frac{a}{4}$. We know that $\mathbb{P}[D_{KT}(\pi, \pi_i) \geq a/4] \leq \frac{\phi^{a/4}}{Z(\phi)} \cdot n!$ so we set $\frac{\epsilon_2}{N} = \frac{\phi^{a_{min}/4}}{Z(\phi)} \cdot n! \Rightarrow log\left(\frac{\epsilon_2 \cdot Z(\phi)}{N \cdot n!}\right) = log(\phi) \cdot a_{min}/4 \Rightarrow$
$a_{min} = 4 \cdot log\left(\frac{\epsilon_2 \cdot Z(\phi)}{N \cdot n!}\right) / log(\phi) = 4 \cdot log\left(\frac{N \cdot n!}{\epsilon_2 \cdot Z(\phi)}\right) / log(\frac{1}{\phi}) =$
$O([log(N) + nlog(n) - log(\epsilon_2) - (n-1)log(\phi + 1)]/log(\frac{1}{\phi}))$

From union bound over N samples we yield:

$\mathbb{P}$[learning algorithm fails] $\leq \mathbb{P}$[not enough samples are drawn from some cluster]+$\mathbb{P}$[clustering algorithm fails]+$\mathbb{P}$[central permutation estimator fails in some cluster] $\leq \epsilon_1 + \epsilon_2 + \epsilon_3$. Setting $\epsilon_1 = \epsilon_2 = \epsilon_3 = \epsilon/3$ we bound the error probability of the learning algorithm by ϵ.

5.5.3 Robustness of Learning Separable Mallows Mixture Models Under Selection Noise

In this chapter we extend the work of the previous chapter to the case where samples are incomplete. We suppose that a selection mechanism $p(m)$ drops m items with probability $p(m)$. The selection affects the distance between dissimilar rankings, because discordant pairs may be discarded and the resulting permutations may be closer to each other than the initial complete ones. The effect of selectivity should be taken into account into the separation condition.

Let J be the selection set and m be the number of missing items. $m = n - |J|$.

A lower bound for the distance after selection would be the following:
$D_{KT}(\pi|J, \sigma|J) \geq D_{KT}(\pi, \sigma) - n - (n-1) - \ldots - (n-m+1) = D_{KT}(\pi, \sigma) - (2n - m + 1)m/2$.
The first item that gets dropped by the selection mechanism gives at most n discordant pairs between π and σ. The second gives at most $n-1$ new pairs, and in general the m-th gives $n - (m-)$ new pairs. The m-th item is in n discordant pairs at most, but at least $m-1$ are common with the previous $m-1$ items. Thus, the effect of selectivity on the KT distance is a decrease less or equal to $\frac{(2n-m+1)m}{2}$.

An upper bound for the distance after selection would be:
$D_{KT}(\pi|J, \sigma|J) \leq D_{KT}(\pi, \sigma)$.
This would be the case if all discarded items participated only in concordant comparisons, so the total number of discordant pairs would remain unchanged.

Let d be the maximal radius of the clusters. For all pairs (i,j) of clusters with latent centers (π_i, π_j) we require that with high probability $D_{KT}(\sigma_{i,1}|J, \sigma_{i,2}|J) < D_{KT}(\sigma_{i,1}|J, \sigma_j|J)$, where $\sigma_{i,1}, \sigma_{i,2}$ are random samples drawn from cluster i and σ_j is a random sample drawn from cluster j.
Using the lower and upper bounds discussed above we have:
$D_{KT}(\sigma_{i,1}|J, \sigma_{i,2}|J) \leq D_{KT}(\sigma_{i,1}, \sigma_{i,2}) \leq 2d$
$D_{KT}(\sigma_{i,1}|J, \sigma_j|J) \geq D_{KT}(\sigma_{i,1}, \sigma_j) - n - (n-1)\ldots - (n-m+1) \geq a - 2d - (2n - m + 1)m/2$.

To guarantee that $D_{KT}(\sigma_{i,1}|J, \sigma_{i,2}|J) < D_{KT}(\sigma_{i,1}|J, \sigma_j|J)$ with high probability, with demand that with high probability it holds that $2d < a - 2d - (2n - m + 1)m/2$
$\Leftrightarrow (2n - m + 1)m/2 < a - 4d \Leftrightarrow a > (2n - m + 1)m/2 + 4d$ **(1)**.

We know that $\mathbb{P}[D_{KT}(\pi, \pi_i) \geq d] = \sum_{l=d}^{\frac{n(n-1)}{2}} \frac{A(n,l)\cdot\varphi^l}{Z(\varphi)} \leq \frac{\varphi^d}{Z(\varphi)} \cdot \sum_{l=d}^{\frac{n(n-1)}{2}} A(n,l) \leq \frac{\varphi^d}{Z(\varphi)} \cdot n!$

We want that with high probability (at least $1-\epsilon_1$) all samples lie within a radius d_{max} of the center that generated them. If the probability that one samples violates the condition is less than $\frac{\epsilon_1}{N}$, then from union bound over all N samples the total condition is satisfied with probability at least $1-\epsilon_1$. We set
$\frac{\epsilon_1}{N} = \frac{\phi^{d_{max}}}{Z(\phi)} \cdot n! = \frac{\phi^{d_{max}}}{Z(\phi)} \cdot n! \Rightarrow$
$log\left(\frac{\epsilon_1 \cdot Z(\phi)}{N \cdot n!}\right) = log(\phi) \cdot d_{max} \Rightarrow$
$d_{max} = log\left(\frac{\epsilon_1 \cdot Z(\phi)}{N \cdot n!}\right) / log(\phi) = log\left(\frac{N \cdot n!}{\epsilon_1 \cdot Z(\phi)}\right) / log\left(\frac{1}{\phi}\right) =$
$O\Big([log(N) + nlog(n) - log(\epsilon_1) - (n-1)log(\phi+1)] \, / log(\frac{1}{\phi})\Big)$

p(m) depends on the selection mechanism. For example if each element is dropped independently from the others with probability p_d, then $p(m) = \binom{n}{m} \cdot {p_d}^m \cdot (1-p_d)^{n-m}$.

In order to guarantee that **(1)** is satisfied for all samples with high probability (at least $1-\epsilon_2$) we demand that $a > (2n - m_{cr} + 1)m_{cr}/2 + 4d_{max}$, where m_{cr} is a critical selection length, such that $\mathbb{P}[m \leq m_{cr}] > \epsilon_2/N \Leftrightarrow \sum_{m=0}^{m_{cr}} p(m) > \epsilon_2/N$. This way, from union bound we yield that with probability at least $1-\epsilon_2$ it holds for all samples σ_i that $a > (2n - m_i + 1)m_i/2 + 4d_{max}$, where m_i is the number of missing items from sample σ_i. We we set $\epsilon_1 = \epsilon_2 = \epsilon/2$. Then the clustering algorithm succeeds with probability greater than $1-(\epsilon_1+\epsilon_2) = 1-\epsilon$.

5.5.4 Concentration of Mass of the Mallows Distribution Inside the Sphere of Radius d

In this chapter we study the way in which the mass of a Mallows model is distributed at different distances around the central ranking. This analysis is connected with the analysis of separable mixtures in this work, because the latter demands that each component is restricted inside a sphere with a small radius compared to the distance between different centers of the mixture. Thus, we need to know how large the radius of a Mallows hyper-sphere should be in order to enclose a high proportion of the total mass of the model.
The probability that a sample drawn from a Mallows model $\mathcal{M}(\pi_0, \phi)$ lies within a hypersphere of radius a is equal to $\sum_{d=0}^{a} \frac{A(n,d)}{Z(\phi,n)}\phi^d$, where n is the number of items in π_0 and $A(n,d)$ is the sequence of Mahonian numbers, introduced in subsection 2.2.1 . We observe that this probability only depends on n, ϕ and a. If we see it as a function of the radius x, then $p(x) = \sum_{d=0}^{x} \frac{A(n,d)}{Z(\phi,n)}\phi^d$ and $p(x)$ is parameterized by n and ϕ. We would like to know how these parameters affect the form of $p(x)$. The only obstacle to this is the absence of a closed form formula for the Mahonian numbers $A(n,d)$. Thus, we will need to find closed form bounds for $A(n,d)$. In this chapter we provide such bounds and we also study the figure of $p(x)$ experimentally to gain intuition.

Theoretical Bounds

Lemma 5.5.1. *Mahonian numbers A(n,k) are symmetric on k:* $A(n,k) = A\left(n, \binom{n}{2} - k\right)$.

Proof. Base case: $A(2,0) = A(2,1) = 1$

Induction hypothesis:
$A(n-1,k) = A\left(n-1, \binom{n-1}{2} - k\right) \quad \forall k \in \left\{0, 1, \ldots \binom{n-1}{2}\right\}, n > 2$

Induction step:
For all k $\in \left\{0, 1, \ldots \binom{n}{2}\right\}$ we have: $A\left(n, \binom{n}{2} - k\right) = \sum_{j=0}^{n-1} A\left(n-1, \binom{n}{2} - k - j\right) =$
$\sum_{j=0}^{n-1} A\left(n-1, \binom{n-1}{2} - \binom{n}{2} + k + j\right) = \sum_{j=0}^{n-1} A(n-1, 1-n+k+j) =$
$\sum_{j'=1-n}^{0} A(n-1, k+j') = \sum_{i=0}^{n-1} A(n-1, k-i) = A(n,k).$

Note that for $k < 0$ or $k > \binom{n-1}{2}$ $A(n-1,k) = A\left(n-1, \binom{n-1}{2} - k\right) = 0$ from the recursive formula of the Mahonian numbers, so we only make use of the induction hypothesis for non trivial values of k $\left(0 \le k \le \binom{n-1}{2}\right)$. □

Lemma 5.5.2. *Mahonian numbers A(n,k) are increasing on k for $k \le \left\lceil \binom{n}{2}/2 \right\rceil$.*

Proof. Base case: $A(3,0) = 1 < A(3,1) = 2$

Induction hypothesis:
$A(n-1,k) < A(n-1,k+1), \quad \forall k \in \left\{0, 1, \ldots \left\lceil \binom{n-1}{2}/2 \right\rceil - 1\right\}, n > 3$

Induction step:
For all k $\in \left\{0, 1, \ldots \left\lceil \binom{n}{2}/2 \right\rceil - 1\right\}$ we have: $A(n,k+1) = \sum_{j=0}^{n-1} A(n-1, k+1-j) > \sum_{j=0}^{n-1} A(n-1, k-j) = A(n,k)$, because $A(n-1,k+1-j) > A(n-1,k-j)$ for $j \le k$ from the induction hypothesis and $A(n-1,k+1-j) \ge 0 = A(n-1,k-j)$ for $j > k$. □

Combining the two above lemmas we can see that the Mahonian numbers are decreasing for $k > \left\lfloor \binom{n}{2}/2 \right\rfloor$ and are maximised at $k = \left\lceil \binom{n}{2}/2 \right\rceil$. Moreover, a random permutation (uniform distribution on $\mathbb{S}_n$) has expected distance from another fixed permutation equal to $\left\lceil \binom{n}{2}/2 \right\rceil$.

Lemma 5.5.3. *Mahonian numbers A(n,k) are greater than $\binom{n}{k}$ for $k > 2$ and $n > 2$. For all n and k $A(n,k) \ge \binom{n}{k} - 1$.*

From the above lemma we have for the probability mass inside the Mallows sphere of radius d=n: $\mathbb{P}[d \le n] = \sum_{d=0}^{n} \frac{A(n,d)}{Z(\phi,n)} \phi^d \ge \frac{1}{Z(\phi,n)} \left(\sum_{d=0}^{n} \binom{n}{d} \phi^d - 1 - \phi - \phi^2\right) =$
$\frac{1}{Z(\phi,n)} \left((\phi+1)^n - 1 - \phi - \phi^2\right)$

Lemma 5.5.4. *Mahonian numbers A(n,d) are greater than $n \cdot (n-1) \ldots \cdot (n - (\frac{d}{n} - 1))$ for $d \le \frac{n(n-1)}{4}$*

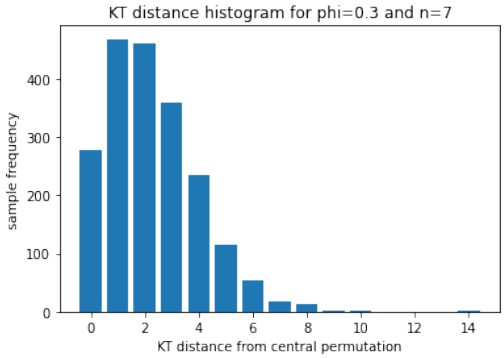
Proof. $A(n,d) > n \cdot (n-1) \ldots \cdot (n-x)$, where x is some recursion depth such that recursion tree is complete (each node has only non zero children). Restrictions for x:
$0 \le d - x \cdot n + (x+1)x/2 < (n-x)(n-x-1)/2$

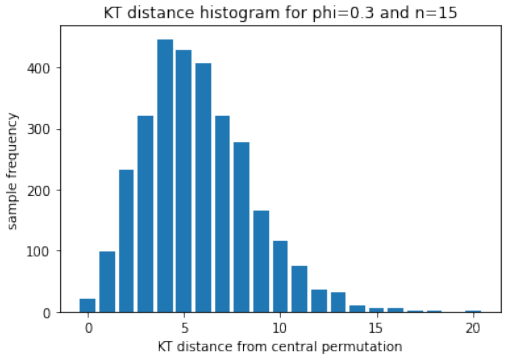
$0 < n - x$
$d < (n-x)(n-x-1)/2$
$x_{max} = n - 1 - \sqrt{(n-\frac{1}{2})^2 - 2d}, d \leq \frac{n(n-1)}{4}$
We set $a = n - \frac{1}{2}$, $b = 2d$.
$x_{max} + \frac{1}{2} = a - \sqrt{a^2 - b} = \frac{b}{a + \sqrt{a^2-b}} > \frac{b}{2a}$
So we yield $x_{max} > \frac{d}{n-\frac{1}{2}} - \frac{1}{2}$.
Plugging this into the inequality $A(n,d) > n \cdot (n-1) \ldots \cdot (n - x_{max})$ completes the proof. □

From the above lemma we have for the probability mass inside the Mallows sphere of radius d=x: $\mathbb{P}[d \leq x] = \sum_{d=0}^{x} \frac{A(n,d)}{Z(\phi,n)} \phi^d \geq \frac{1}{Z(\phi,n)} \left(\sum_{d=0}^{x} n \cdot (n-1) \ldots \cdot \left(n - (\frac{d}{n} - 1) \right) \phi^d \right)$.

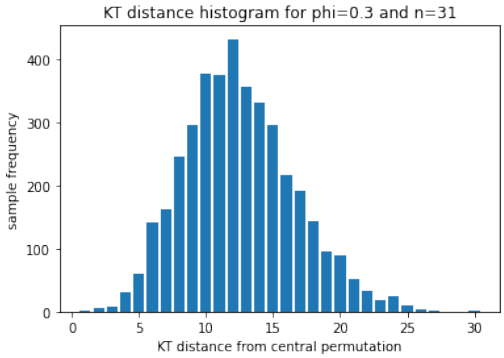
Experimental Results

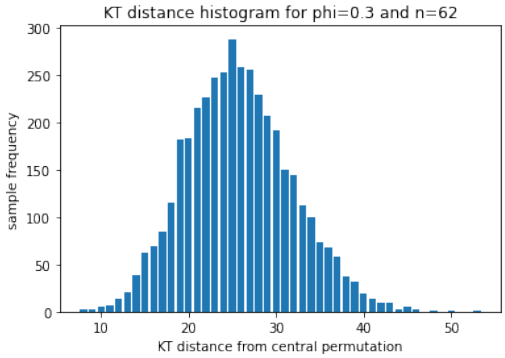
We plot the area of concentration of the Mallows model as a function of n and ϕ. When we say "area of concentration" we mean a distance interval $[d_{min}, d_{max}]$, such that with high probability a random sample drawn from the Mallows model has distance d, with $d_{min} \leq d \leq d_{max}$ from the central permutation. Both d_{min} and d_{max} seem to scale almost linearly on n. The slope of these two linear functions depends on the spread parameter ϕ. The higher the spread parameter, the greater the values of d_{min} and d_{max}, leading to a more diffuse area of concentration.



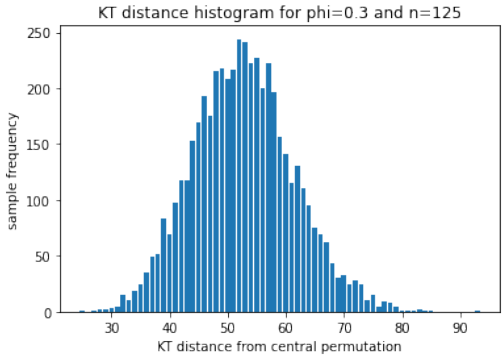


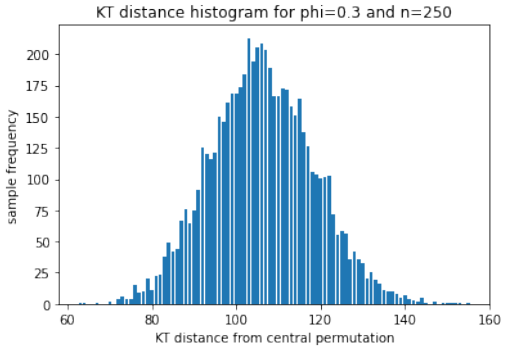
(a) *n=7* | *n=15*



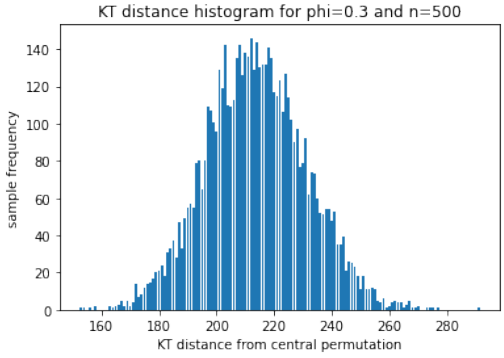


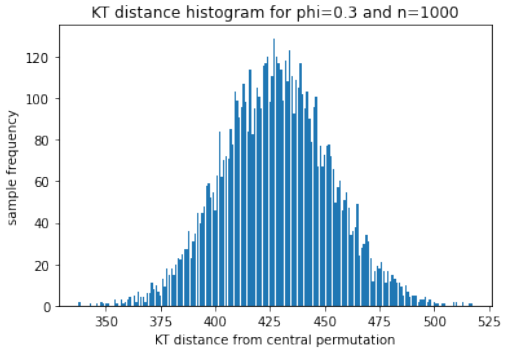
(b) *n=31* | *n=62*





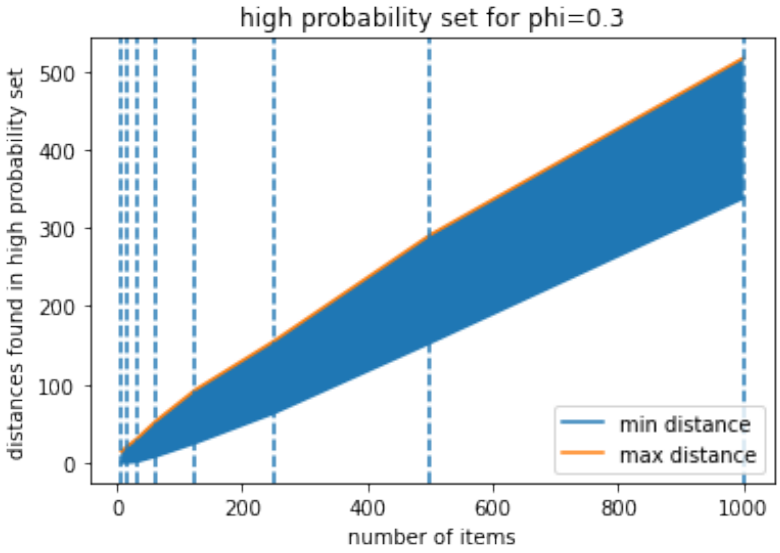
(c) *n=125* | *n=250*



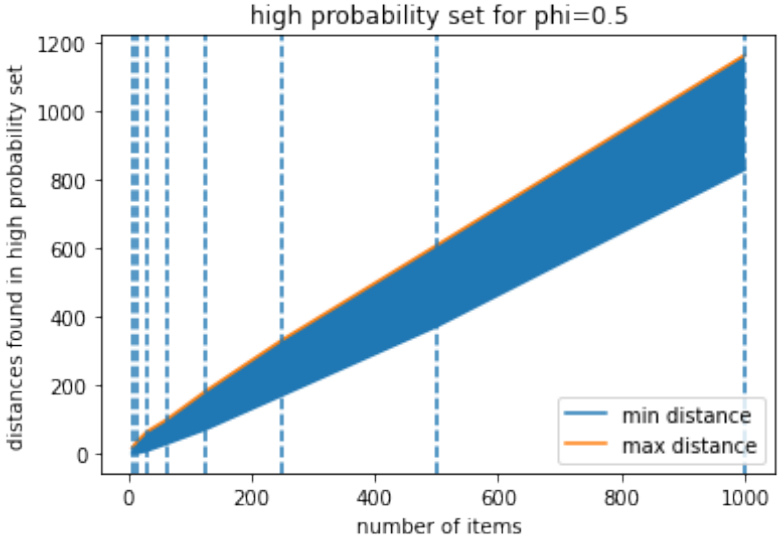


(d) *n=500* | *n=1000*

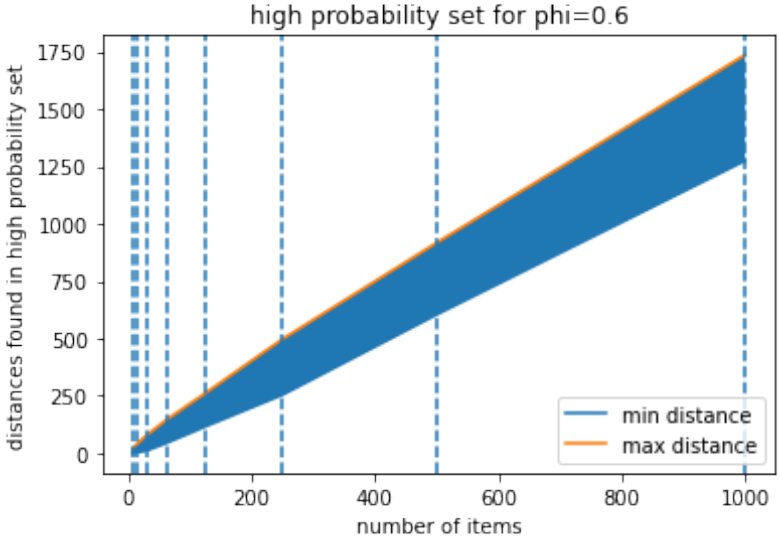
Figure 5.1. *Area of concentration of the Mallows mass for $\varphi = 0.3$ and different values of n*



(a) $\phi = 0.3$



(b) $\phi = 0.5$



(c) $\phi = 0.6$

Figure 5.2. *range of concentration as a function of n for different values of φ*

Chapter 6

Conclusion-Future Work

In this thesis we presented the most important theoretical results in the field of Mallows Mixture Learning. Building on previous work, we prove a tight condition on the minimal sample length that preserves identifiability. We propose an algorithm that learns mixtures of two Mallows models using exclusively pairwise comparisons, which is the most extreme case of selectivity. We show how the existing algorithms that use complete samples can be used as a subroutine for a learning algorithm that uses incomplete samples. Then we focus on separable mixtures, where we can detect central permutations by looking at the pmf modes or most strongly when components are so far from each other, that samples from one component are most likely closer to each other than to samples from another component. In the general case, the sample and time complexity are proved to be polynomial on all parameters except the number k of components and in the case of separable mixtures it is polynomial to k as well.

One important limitation of our results is that in the non separable case we can not use samples from a specific selection set, unless many other samples from this selection set are present, so that an empirical histogram on the selection set is formed. One potential solution would be to replace the histogram criterion (such as 4.3.1) with a bayesian likelihood criterion. In particular, given a collection of incomplete samples we could choose the candidate model that maximises the likelihood of the sample collection. This could definitely work as a heuristic and it could work theoretically if we find an upper bound for the likelihood of the samples under a candidate model with different parameters than the original ones and a lower bound for the likelihood under the correct candidate model that is greater than the aforementioned upper bound.

Another direction would be finding a tight condition on the minimal sample length that preserves identifiability in mixtures with different spread parameters. Using the results of [5] we show that it suffices to have samples from all subsets of length $10k+3$. There is gap between this bound and the bound $2log(k)+3$ that holds in the case of equal spread parameters. The problem arises from the fact that in the case of equal spread parameters the complete mixture of k distinct components is identifibale for all numbers n of items, due to the non zero determinant of Zagier. However, in the case of non equal spread parameters the existing literature proves identifiability assuming that $n \geq 10k$. It would

be interesting to reduce this bound to $O(log(k))$.

Bibliography

[1] Wassily Hoeffding. *Probability Inequalities for Sums of Bounded Random Variables*. *Journal of the American Statistical Association*, 58(301):13–30, 1963.

[2] Ioannis Caragiannis, Ariel D. Procaccia and Nisarg Shah. *When Do Noisy Votes Reveal the Truth?* *ACM Trans. Econ. Comput.*, 4(3), 2016.

[3] Róbert Busa-Fekete, Dimitris Fotakis, Balázs Szörényi and Manolis Zampetakis. *Optimal Learning of Mallows Block Model*, 2019.

[4] Don Zagier. *Realizability of a model in infinite statistics*. *Communications in Mathematical Physics*, 147(1):199–210, 1992.

[5] Allen Liu and Ankur Moitra. *Efficiently Learning Mixtures of Mallows Models*. *2018 IEEE 59th Annual Symposium on Foundations of Computer Science (FOCS)*, 2018.

[6] Cheng Mao and Yihong Wu. *Learning mixtures of permutations: Groups of pairwise comparisons and combinatorial method of moments*, 2020.

[7] Michael A. Fligner. *Probability models and statistical analyses for ranking data*. Springer-Verl., 1993.

[8] John I. Marden. *Analyzing and modeling rank data*. Chapman amp; Hall, 1995.

[9] Lirong Xia. *Learning and decision-making from Rank Data*. Morgan amp; Claypool Publishers, 2019.

[10] C. L. MALLOWS. *NON-NULL RANKING MODELS. I*. *Biometrika*, 44(1-2):114–130, 1957.

[11] Louis Leon Thurstone. *A Law of Comparative Judgement*. *Psychological Review*, 34:278–286, 1927.

[12] *Discussion on Professor Ross's Paper*. *Journal of the Royal Statistical Society: Series B (Methodological)*, 12(1):41–59, 1950.

[13] Ralph Allan Bradley and Milton E. Terry. *Rank Analysis of Incomplete Block Designs: I. The Method of Paired Comparisons*. *Biometrika*, 39(3/4):324–345, 1952.

[14] R. L. Plackett. *The Analysis of Permutations*. *Applied Statistics*, 24(2):193, 1975.

[15] R.duncan Luce. *The choice axiom after twenty years*. *Journal of Mathematical Psychology*, 15(3):215–233, 1977.

[16] Mark Braverman and Elchanan Mossel. *Sorting from Noisy Information*, 2009.

[17] Wenpin Tang. *Mallows ranking models: maximum likelihood estimate and regeneration*. *ICML*, 2019.

[18] F. Chierichetti, Anirban Dasgupta, R. Kumar and S. Lattanzi. *On reconstructing a hidden permutation*. *Leibniz International Proceedings in Informatics, LIPIcs*, 28:604–617, 2014.

[19] Dimitris Fotakis, Alkis Kalavasis and Konstantinos Stavropoulos. *Aggregating Incomplete and Noisy Rankings*, 2020.

[20] Stavropoulos Konstantinos. *Μάθηση διατάξεων από δείγματα με ελλιπή πληροφορία*. Διπλωματική εργασία, NTUA, 2020.

[21] Bruce Hajek, Sewoong Oh and Jiaming Xu. *Minimax-optimal Inference from Partial Rankings*. *Advances in Neural Information Processing SystemsZ*. Ghahramani, M. Welling, C. Cortes, N. Lawrence and K. Q. Weinberger, editors, volume 27. Curran Associates, Inc., 2014.

[22] Nihar B. Shah, Sivaraman Balakrishnan, Joseph Bradley, Abhay Parekh, Kannan Ramchandran and Martin J. Wainwright. *Estimation from Pairwise Comparisons: Sharp Minimax Bounds with Topology Dependence*, 2015.

[23] Pranjal Awasthi, Avrim Blum, Or Sheffet and Aravindan Vijayaraghavan. *Learning Mixtures of Ranking Models*. *Advances in Neural Information Processing Systems*, volume 27, 2014.

[24] Flavio Chierichetti, Anirban Dasgupta, Ravi Kumar and Silvio Lattanzi. *On Learning Mixture Models for Permutations*. *Proceedings of the 2015 Conference on Innovations in Theoretical Computer Science*, ITCS '15, page 85–92, New York, NY, USA, 2015. Association for Computing Machinery.

[25] Thomas Brendan Murphy and Donal Martin. *Mixtures of distance-based models for ranking data*. *Computational Statistics Data Analysis*, 41(3):645–655, 2003. Recent Developments in Mixture Model.

[26] Tyler Lu and Craig Boutilier. *Learning Mallows Models with Pairwise Preferences*. *Proceedings of the 28th International Conference on International Conference on Machine Learning*, ICML'11, page 145–152, Madison, WI, USA, 2011. Omnipress.

[27] Radford M. Neal and Geoffrey E. Hinton. *A View of the Em Algorithm that Justifies Incremental, Sparse, and other Variants*, pages 355–368. Springer Netherlands, Dordrecht, 1998.

[28] Julia Stoyanovich, Lovro Ilijasic and Haoyue Ping. *Workload-Driven Learning of Mallows Mixtures with Pairwise Preference Data*. *Proceedings of the 19th International Workshop on Web and Databases*, WebDB '16, New York, NY, USA, 2016. Association for Computing Machinery.

[29] Brendan J. Frey and Delbert Dueck. *Clustering by Passing Messages Between Data Points*. *Science*, 315(5814):972–976, 2007.

[30] Cynthia Dwork, Ravi Kumar, Moni Naor and D. Sivakumar. *Rank Aggregation Methods for the Web*. *Proceedings of the 10th International Conference on World Wide Web*, WWW '01, page 613–622, New York, NY, USA, 2001. Association for Computing Machinery.

[31] Marina Meilă and Harr Chen. *Dirichlet Process Mixtures of Generalized Mallows Models*. *Proceedings of the Twenty-Sixth Conference on Uncertainty in Artificial Intelligence*, UAI'10, page 358–367, Arlington, Virginia, USA, 2010. AUAI Press.

[32] Flavio Chierichetti, Anirban Dasgupta, Shahrzad Haddadan, Ravi Kumar and Silvio Lattanzi. *Mallows Models for Top-k Lists*. *Advances in Neural Information Processing Systems*S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi and R. Garnett, editors, volume 31. Curran Associates, Inc., 2018.

[33] Collas Fabien and Irurozki Ekhine. *Concentric mixtures of Mallows models for top-k rankings: sampling and identifiability*, 2020.

[34] Zhibing Zhao, Peter Piech and Lirong Xia. *Learning Mixtures of Plackett-Luce Models*, 2020.

[35] Xiaomin Zhang, Xucheng Zhang, Po Ling Loh and Yingyu Liang. *On the identifiability of mixtures of ranking models*, 2022.

[36] Anindya De, Ryan O'Donnell and Rocco Servedio. *Learning sparse mixtures of rankings from noisy information*, 2018.

[37] Henry Teicher. *Identifiability of Finite Mixtures*. *The Annals of Mathematical Statistics*, 34(4):1265 – 1269, 1963.

[38] Henry Teicher. *Identifiability of Mixtures*. *The Annals of Mathematical Statistics*, 32(1):244 – 248, 1961.

[39] Ankur Moitra and Gregory Valiant. *Settling the Polynomial Learnability of Mixtures of Gaussians*, 2010.

[40] Rares Darius Buhai and David Steurer. *Beyond Parallel Pancakes: Quasi-Polynomial Time Guarantees for Non-Spherical Gaussian Mixtures*, 2021.

[41] Persi Diaconis and R. L. Graham. *Spearmans Footrule as a Measure of Disarray*. *Journal of the Royal Statistical Society: Series B (Methodological)*, 39(2):262–268, 1977.

[42] L. G. Valiant. *A Theory of the Learnable*, 1984.

[43] A. B. Tsybakov. *Introduction to nonparametric estimation*, page 132. Springer, 2010.

[44] Jean Bretagnolle and Catherine Huber. *Estimation des densités : risque minimax*. *Séminaire de probabilités de Strasbourg*, 12:342–363, 1978.

[45] Robert M. Fano. *Transmission of information: A statistical theory of communications*. M.I.T. Press, 1968.

[46] Herman Chernoff. *A Measure of Asymptotic Efficiency for Tests of a Hypothesis Based on the sum of Observations*. *The Annals of Mathematical Statistics*, 23(4):493 – 507, 1952.

[47] Mark Braverman and Elchanan Mossel. *Sorting from Noisy Information*. *CoRR*, abs/0910.1191, 2009.

[48] Aditya Bhaskara, Moses Charikar and Aravindan Vijayaraghavan. *Uniqueness of Tensor Decompositions with Applications to Polynomial Identifiability*, 2013.

[49] Uriel Feige, Prabhakar Raghavan and Eli Upfal. *Computing with Noisy Information*. *SIAM J. Comput.*, 23:1001–1018, 1994.

[50] Susan Davidson, Sanjeev Khanna, Tova Milo and Sudeepa Roy. *Top-k and Clustering with Noisy Comparisons*. *ACM Trans. Database Syst.*, 39(4), 2015.