University of Helsinki
Faculty of Science
Master's Programme in Mathematics and Statistics



Master's thesis

Robust Clustering of Multinomial Count Data via Dirichlet-Multinomial Mixture Models

Tianyu Liu

Supervisor: Leena Kalliovirta May 2026

Title: Robust Clustering of Multinomial Count Data via Dirichlet–Multinomial Mixture Models
Author: Tianyu Liu
Month and year: May 2026
Page count: 51 pp.

Abstract:

Multivariate count data are widely encountered in applications such as text analysis, biological studies, and image processing. The multinomial model is commonly used for such data but fails to account for heterogeneity and overdispersion. The Dirichlet–Multinomial distribution addresses this limitation by introducing random variation in the underlying probabilities, allowing for more flexible modeling of count data. In many applications, observations arise from multiple latent subpopulations, motivating the use of Dirichlet–Multinomial mixture models for clustering. Parameter estimation is typically performed using the Expectation–Maximization algorithm. However, standard mixture models are sensitive to contaminated observations, which can distort parameter estimates and reduce clustering accuracy. This thesis considers a robust extension of the Dirichlet–Multinomial mixture model by incorporating a noise component to capture atypical observations. A simulation study is conducted under varying data structures and contamination levels to evaluate model performance. The results show that while the standard model performs well in clean settings, its performance deteriorates under contamination. In contrast, the robust model maintains greater stability and achieves improved clustering accuracy by effectively isolating outliers.

Keywords: Multinomial distribution, mixture models, robust clustering, count data, EM algorithm

Contents

1 Intoduction **1**

2 Models for categorical data **3**
2.1 Multinomial distribution 3
2.2 Dirichlet-Multinomial distribution 6
2.2.1 Dirichlet distribution 7
2.2.2 Derivation of Dirichlet distribution 7
2.2.3 Dirichlet-Multinomial distribution 9
2.2.4 Binomial-Beta distribution with an example on capture-recapture data 11
2.3 Limitations of Dirichlet-Multinomial distribution 11
2.4 Generalized Dirichlet-Multinomial distribution 12
2.5 Dirichlet-Multinomial mixture model 15
2.5.1 Latent Heterogeneity and Mixture Modeling 15
2.5.2 Interpretation of Model Parameters 16
2.5.3 Likelihood representation 17

3 Estimation **18**
3.1 Maximum likelihood estimation 18
3.1.1 Latent Variable Representation 19
3.1.2 Motivation for the EM Algorithm 20
3.2 The Expectation-Maximization (EM) Algorithm 20
3.2.1 Complete data formulation 20
3.2.2 Expectation step 21
3.2.3 Maximization step 21
3.2.4 Convergence Properties 22
3.3 Robust Extension: Noise Component 22
3.3.1 Mixture Model with a Noise Component 23
3.3.2 Parameter Estimation under the Robust Model 23
3.3.3 Interpretation of the Noise Component 24
3.4 Evaluation Metrics 24
3.4.1 Clustering accuracy 25
3.4.2 Clustering robustness 26

3.4.3 Computational time 26

4 Simulation Studies **27**
4.1 Data generation 27
4.2 Contamination mechanism 28
4.3 Experiments 29
4.4 Simulation results 29
4.4.1 Single dominant index 29
4.4.2 Double dominant indices 33
4.4.3 Orthogonal structure 37

5 Conclusion **43**

Bibliography **45**

A Proofs for section 2.2.2 **48**

B Pseudo-code of simulation **50**

Chapter 1

Intoduction

Multivariate count data arise in a wide range of scientific applications, including text mining, biological studies, ecology and image analysis. In many of these context, observations are naturally represented as vectors of counts over a finite set of categories. For example, documents can be represented by word frequency vectors, and images could be histograms of visual features. A common characteristic of such data is that the counts are constrained to sum to a fixed total number, making probabilistic modeling inherently compositional.

A fundamental model for this type of data is the multinomial distribution, which describes the outcome of repeated independent samples over a fixed set of categories. While the multinomial model is mathematically convenient and widely used, it relies on a strong assumption that all observations share a common underlying probability vector. In practice, this assumption is often violated. Real-world data frequently exhibit heterogeneity across observations, which implies that different samples may have their own underlying probabilistic distribution. The variability will also induce the problem that the observed variance is often larger than the expected value. The phenomenon is referred to as overdispersion. To address this limitation, compound distributions are often used. In this thesis, we will focus on Dirichlet-Multinomial distribution, where we treat the probability vector as a random variable drawn from a Dirichlet distribution. This framework provides a natural way to account for overdispersion and has been successfully applied in various domains. However, despite its flexibility, the Dirichlet-Multinomial model still assumes that all observations arise from a single homogeneous population, which restricts the observed data to have similar behaviors. In many applications, observed data often contain subpopulations. For instance, documents may belong to different topics, which may exhibit different word usage patterns. In such settings, the observed variability is driven not only by random fluctuations but also by the presence of multiple latent structures. This motivates the use of finite mixture models,

where the population is modeled as a combination of several subpopulations, each characterized by its own parameter distribution.

The Dirichlet-Multinomial mixture model provides a natural framework for clustering multivariate count data. By combining multiple Dirichlet-Multinomial components, the model could simultaneously capture within-cluster variability and between-cluster heterogeneity. Parameter estimation in this framework is typically performed by numerical iterative methods since there are hardly closed-form solution in compound models. In this thesis, we adopt Expectation-Maximization algorithm, where the model iteratively infers latent cluster assignments and updates model parameters. Despite its flexibility, the standard mixture modeling approach is sensitive to contaminated observations or outliers. In practical datasets, it is common to encounter observations that do not conform to any of the underlying cluster structures. These atypical observations may arise due to measurement errors, rare events, or fundamentally different generating mechanisms. Since the likelihood-based estimation procedure attempts to fit all observations, even a small proportion of outliers can significantly distort the parameter estimates and degrade latent cluster assignment accuracy. To address this issue, a robust extension of mixture models will be proposed in Chapter 3. We introduce an additional noise component into the mixture model, which is designed to capture or absorb observations that are not well explained by the primary clusters (Coretto & Hennig, 2016; Fraley & Raftery, 2002), thereby reducing their influence on parameter estimation and providing a more meaningful cluster interpretation given the presence of contamination.

We develop a simulation framework, in which both the structural complexity of the data and the level of contamination can be systematically controlled. We will consider three increasingly harder scenarios to categorize well-separated latent clusters and more complex structures. Within each scenario, we introduce a stepped increase in contamination levels followed by three main evaluation metrics. In Chapter 2, we will introduce the establishment process the Dirichlet-Multinomial mixture model. In Chapter 3, the estimation methods and robust extension will be elaborated, followed by simulation design and result in Chapter 4 and conclusions in Chapter 5.

Chapter 2

Models for categorical data

This chapter introduces the probabilistic models for categorical and count data that form the foundation of the subsequent analysis. We begin with the multinomial distribution, which serves as the basic model for multivariate count observations under the assumption of a common probability vector. We then extend this framework by introducing the Dirichlet distribution and the Dirichlet–Multinomial model, which allow for heterogeneity and overdispersion across observations. The limitations of these models are discussed, motivating the consideration of more flexible extensions such as the generalized Dirichlet–Multinomial distribution. Finally, we present the Dirichlet–Multinomial mixture model, which provides a natural framework for modeling latent heterogeneity and forms the basis for the clustering methodology developed in later chapters.

2.1 Multinomial distribution

Suppose we have n independent trials and each trial results in one of K exclusive categories. The probability that ith-outcome happens is p_j for $j = 1, \ldots, K$, and p_j is called success probability with $\sum_{j=1}^{K} p_j = 1$.

Let X_j denote the number of occurrence of outcome j for $j = 1, \ldots, K$. Each of the count takes place in $\{0, 1, 2, \ldots, n\}$ and $\sum_{j=1}^{K} X_j = n$. Then we get a cell count vector $\boldsymbol{X}$ and its corresponding success probability vector $\boldsymbol{p}$

$$\boldsymbol{X} = (X_1, \ldots, X_K), \quad \boldsymbol{p} = (p_1, \ldots, p_K).$$

Conditioning $\boldsymbol{X}$ on $\boldsymbol{p}$ follows the Multinomial distribution

$$\boldsymbol{X} \mid \boldsymbol{p} \sim Multinomial(n, \boldsymbol{p}).$$

with probability mass function

$$f(\boldsymbol{X} \mid \boldsymbol{p}) = P(X_1 = x_1, \ldots, X_K = x_K \mid p_1, \ldots, p_K)$$

Table 2.1: ABO group frequencies in blood donors in Finland based on the study of Eriksson et al. (1962).

Blood type	O	A	B	AB
Probabilities	0.332	0.432	0.155	0.081

$$
\begin{aligned}
&= \binom{n}{x_1} p_1^{x_1} \binom{n-x_1}{x_2} p_2^{x_2} \dots \binom{n-\sum_{j=1}^{K-1} x_j}{x_K} p_K^{x_K} \\
&= \frac{n!}{x_1! x_2! \dots x_K!} p_1^{x_1} p_2^{x_2} \dots p_K^{x_K}. \qquad (2.1)
\end{aligned}
$$

The expectation and variance for an element X_j are given by

$$E(X_j) = np_j,\ Var(X_j) = np_j(1-p_j),\ j = 1, \dots, K. \qquad (2.2)$$

After having a brief view of the distribution, we now use a simple example to illustrate where this distribution can be applied. In research on blood groups in Finland (Eriksson et al., 1962), blood types of 6474 blood donors are collected and the result is shown in Table 1. The probability estimates obtained using the study indicate that in a random sample of 10 donors in Finland, the probability of an event of 6 with blood type O, 2 with type A, 1 with type B, and 1 with type AB, is

$$
\begin{aligned}
P(X_1 = 6, X_2 = 2, X_3 = 1, X_4 = 1) &= \frac{10!}{6!2!1!1!} 0.332^6 0.432^2 0.155^1 0.081^1 \\
&= 0.0079.
\end{aligned}
$$

In practice, we observe not only a single multinomial count vector but a collection of independent multinomial samples. Suppose we have N independent samples drawn from a single, homogeneous population. We assume that the probability of observing category j is constant across all samples and is denoted by p_j. Thus, the entire dataset is characterized by a single parameter vector $\boldsymbol{p} = (p_1, \dots, p_K)$, where $\sum_{j=1}^{K} p_j = 1$. For each sample i, the probability mass function is given by

$$f(\boldsymbol{X}_i \mid \boldsymbol{p}) = \frac{n_i!}{x_{i1}! x_{i2}! \cdots x_{iK}!} \prod_{j=1}^{K} p_j^{xij}. \qquad (2.3)$$

Provided the $N \in \mathbb{N}$ independent samples, we define the full sample matrix $\mathcal{X}$ as the collection of all observed independent count vectors as

$$\mathcal{X} = (\boldsymbol{X}_1, \dots, \boldsymbol{X}_N).$$

Additionally, with the corresponding success probability vector, we can derive the log-likelihood function for success probabilities given sample matrix

and sample size

$$l(\boldsymbol{p}; \mathcal{X}, n_1, \ldots, n_N) = log(\prod_{i=1}^{N} f(\boldsymbol{X}_i \mid \boldsymbol{p})) \\ = \sum_{i=1}^{N} log(n_i!) - \sum_{i=1}^{N}\sum_{j=1}^{K} log(x_{ij}!) + \sum_{i=1}^{N}\sum_{j=1}^{K} x_{ij} log(p_j). \tag{2.4}$$

Note that n_i and x_{ij} are independent of p_j for $i = 1, \ldots, N$ and $j = 1, \ldots, k$. Therefore, we can simplify the log-likelihood function to

$$l(\boldsymbol{p}; \mathcal{X}, n_1 \ldots, n_N) = \sum_{i=1}^{N}\sum_{j=1}^{K} x_{ij} log(p_j)$$

based on the proportional relationship. Then the maximum likelihood estimators for p_j from sample i is given by

$$\hat{p}_j^{(i)} = \frac{x_{ij}}{n_i} \quad \text{for i} = 1, 2, \ldots, N,$$

where x_{ij} denotes the number of ocurrences in category j for sample i and n_i denotes the size of i-th sample but the estimates will vary in practice.

Looking back to the blood type example, researchers records birth county of each donor, which is shown in Table 2.2. Then we derive a simple contingency table. Using the MLE estimatior above, we can find that the probabilities of a donor with blood type O and born in Vaasa, Kuopio and Uusimaa are

$$\hat{p}_{31} = \frac{x_{31}}{n_3} = \frac{199}{485} = 0.410,$$
$$\hat{p}_{41} = \frac{x_{41}}{n_4} = \frac{328}{732} = 0.448,$$
$$\hat{p}_{61} = \frac{x_{61}}{n_6} = \frac{693}{2072} = 0.334.$$

However, we notice that the proportions of type O blood vary significantly between counties or group of counties. The simple model fails to capture this heterogeneity. Therefore, the simple multinomial model is insufficient.To resolve this, we must relax the assumption of a fixed $\boldsymbol{p}$, which will be discussed in section 2.2.

Another limitation is that under this model, the expectation and covariance structure is fully determined by the probabilities with

$$E(\boldsymbol{X}_i) = n_i \boldsymbol{p}, \text{and } Cov(X_{ik}, X_{il}) = -n_i p_{ik} p_{il} \quad \text{for } k \neq l,$$

where $k, l \in \{1, \ldots, K\}$. The multinomial distribution introduces a rigid covariance structure, which implies that the cross-category are negatively

Table 2.2: ABO group frequencies in blood donors born in different counties in Finland Eriksson et al. (1962).

County		Number investigated				
		O	A	B	AB	n
Lappi	L	22	26	8	6	62
Oulu	O	100	99	31	16	246
Vaasa	V	199	183	65	38	485
Kuopio	K	196	328	128	80	732
Turku and Pori	T	144	205	91	29	469
Uusimaa	U	693	922	312	145	2072
Häme	H	222	279	106	53	660
Mikkeli	M	114	161	54	39	368
Viipuri	Vi	309	418	154	88	969
Kymi	Ky	152	176	54	29	411
Total		2151	2797	1003	523	6474
		$\hat{p}_O$	$\hat{p}_A$	$\hat{p}_B$	$\hat{p}_{AB}$	
Estimates		0.332	0.432	0.155	0.082	

related due to the constraint $\sum_{l=1}^{K} x_{il} = n_i$. Moreover, in many real applications, the observed count variability exceeds the multinomial variance, which is commonly known as overdispersion (Budig et al., 2026). This typically arises from heterogeneity across sampling units, where underlying probability vector varies from one observation to another. For example, in the context of text modeling (Elkan, 2006), multinomial distribution works poorly to address the problem of word burstiness (Church & Gale, 1995; Katz, 1996), which describes the phenomenon that in a certain document, the occurrence of a certain word are more likely to appear in the same document than other words in the same topic. These limitations motivate compound multinomial models such as Dirichlet multinomial distribution, which explicitly accounts for random variation in $\boldsymbol{p}_i$ and provides increased flexibility for modeling correlated and overdispersed data.

2.2 Dirichlet-Multinomial distribution

To address the limitation of the simple Multinomial model and accommodate the observed heterogeneity and overdispersion, we must relax the assumption that the probability vector is constant across all samples. Instead of a fixed parameter, we treat the probability vector associated with the i-th sample, denoted as $\boldsymbol{p}_i$, as a realization of a random variable. This approach assumes that each sample has its own unique underlying probability distribution, drawn from a random distribution. A common choice is Dirichlet distribution since it is the conjugate prior to the multinomial distribution. This algebraic

property allows us to integrate out $\boldsymbol{p}_i$ analytically, resulting in a closed-form probability mass function for the marginal distribution.

2.2.1 Dirichlet distribution

The Dirichlet distribution, denoted as $Dir(\alpha)$, is defined over the standard K-simplex, Δ_K (Frigyik et al., 2010). This simplex represents the space of all possible probability distributions over K outcomes. A random vector $\boldsymbol{p} = (p_1, p_2, \ldots, p_K)$ is said to follow a Dirichlet distribution if its elements are non-negative and sum to unity

$$\Delta_K = \{\boldsymbol{p} \in \mathbb{R}^K : p_j \geq 0 \text{ for all } j \text{ and} \sum_{j=1}^{K} p_j = 1\}. \tag{2.5}$$

The distribution is parametrized by a vector of positive concentration parameter $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$, where $\alpha_j \geq 0$ for $j = 1, \ldots, K$. The probability mass function is given by

$$f(\boldsymbol{p}; \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{p})} \prod_{j=1}^{K} p_j^{\alpha_j - 1}$$

where

$$B(\boldsymbol{\alpha}) = \frac{\prod_{j=1}^{K} \Gamma(\alpha_j)}{\Gamma(\alpha_0)}, \ \alpha_0 = \sum_{j=1}^{K} \alpha_j.$$

2.2.2 Derivation of Dirichlet distribution

Consider $K + 1$ independent random variables, $Z_1 \ldots Z_{K+1}$, where each Z_l is distributed according to a Gamma distribution with a common rate parameter β and its shape parameter α_l

$$Z_l \sim Gamma(\alpha_l, \beta). \tag{2.6}$$

The joint probability density function of $\boldsymbol{Z} = (Z_1, \ldots, Z_{K+1})$ is the product of the individual Gamma probability density functions

$$f(\boldsymbol{Z}; \boldsymbol{\alpha}, \beta) = \prod_{l=1}^{K+1} \left[\frac{\beta^{\alpha_l}}{\Gamma(\alpha_l)} z_l^{\alpha_l - 1} e^{-\beta Z_l} \right]. \tag{2.7}$$

We define two new variables

$$p_j = \frac{Z_j}{\sum_{l=1}^{K+1} Z_l}, \ j = 1, \ldots, K \quad \text{and} \quad Y = \sum_{l=1}^{K+1} Z_l$$

The set of variables $\boldsymbol{p} = (p_1, \ldots, p_K)$ along with Y forms a one-to-one transformation from the original $\boldsymbol{Z}$ variables. It is a key result that $\boldsymbol{p}$ and Y

are independent random variables. The distribution of the sum Y is also a Gamma distribution

$$Y \sim \text{Gamma}\left(\sum_{l=1}^{K+1} \alpha_l, \beta\right) = \text{Gamma}(\alpha_0, \beta), \tag{2.8}$$

where $\alpha_0 = \sum_{l=1}^{K+1} \alpha_l$. The independence of $\boldsymbol{p}$ and Y gives their joint distribution equal to the product of two marginal distribution $f(\boldsymbol{p}, Y) = f(\boldsymbol{p}) \cdot f(Y)$. The marginal distribution of $\boldsymbol{p}$, which is the Dirichlet distribution, $f(\boldsymbol{p})$, is found by taking the ratio of the joint density of $\boldsymbol{Z}$ to the marginal density of Y.

Theorem 2.2.1 (Rudin, 1976, Theorem 10.9)**.** *Suppose g is a continuously differentiable 1-1 $\mathscr{L}$-1 (Lebesgue integrable) mapping of an open set $E \subset R^k$ into R^k such that the determinant of Jacobian matrix $J \neq 0$ for all $\boldsymbol{x} \in E$. If f is a continuous function on R^k whose support is compact and lies in $g(E)$, then*

$$\int_{R^k} f(\boldsymbol{y}) d\boldsymbol{y} = \int_{R^k} f(g(\boldsymbol{x}))|J(\boldsymbol{x})| d\boldsymbol{x}$$

With this theorem, we can do this change of variables operation to derive the joint probability function $f(\boldsymbol{p}, Y)$ with the mapping relationship

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_{K+1} \end{bmatrix} \xrightarrow{g} \begin{bmatrix} \frac{Z_1}{\sum_{l=1}^{K+1} Z_l} \\ \frac{Z_2}{\sum_{l=1}^{K+1} Z_l} \\ \vdots \\ Z_{K+1} \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ Y \end{bmatrix}.$$

The Jacobian matrix that we need to do the transformation is calculated by

$$J = \begin{vmatrix} \frac{\partial Z_1}{\partial p_1} & \cdots & \frac{\partial Z_1}{\partial p_K} & \frac{\partial Z_1}{\partial Y} \\ \vdots & \ddots & \vdots & \vdots \\ \frac{\partial Z_K}{\partial p_1} & \cdots & \frac{\partial Z_K}{\partial p_K} & \frac{\partial Z_K}{\partial Y} \\ \frac{\partial Z_{K+1}}{\partial p_1} & \cdots & \frac{\partial Z_{K+1}}{\partial p_K} & \frac{\partial Z_{K+1}}{\partial Y} \end{vmatrix} = \begin{vmatrix} Y & \cdots & 0 & p_1 \\ \vdots & \ddots & \vdots & \vdots \\ 0 & \cdots & Y & p_K \\ -Y & \cdots & -Y & 1 - \sum_{j=1}^{K} p_j \end{vmatrix} = Y^K,$$

where $\frac{\partial Z_l}{\partial p_l} = Y$, $\frac{\partial Z_l}{\partial p_j} = 0$, for $l, j \in \{1, 2, \ldots, K\}$ and $l \neq j$, $\frac{\partial Z_l}{\partial Y} = p_l$. Here we do not uncover the derivation and it will be shown in appendices. With all the information that we need, we can have the marginal probability distribution function of $\boldsymbol{p}$

$$f(\boldsymbol{p}) = \frac{\Gamma(\alpha_0)}{\prod_{l=1}^{K} \Gamma(\alpha_l)} \prod_{l=1}^{K} p_l^{\alpha_l - 1} = \frac{1}{B(\boldsymbol{\alpha})} \prod_{l=1}^{K} p_l^{\alpha_l - 1}, \; B(\boldsymbol{\alpha}) = \frac{\prod_{l=1}^{K} \Gamma(\alpha_l)}{\Gamma(\alpha_0)}$$

where $\alpha_0 = \sum_{l=1}^{K} \alpha_l$ and the term $B(\boldsymbol{\alpha})$ is the multivariate Beta function, or the normalizing constant, which ensures the total probability over the simplex integrates to unity. Critically, the common rate parameter β in equation

(2.6) and (2.7) from cancels out in the normalization process, which is why the distribution is independent of the choice of β. This function provides the continuous density over the standard simplex Δ_K for the probability vector $\boldsymbol{p}$.

The Dirichlet distribution provides a flexible probabilistic model for random composition $\boldsymbol{p} = (p_1, \ldots, p_K)$ on the simplex. The parameters give an intuitive interpretation in terms of mean and concentration. If our probability function follows a Dirichlet distribution, then the expectation and variance of each component are

$$E(p_j) = \frac{\alpha_j}{\alpha_0},\ Var(p_j) = \frac{\alpha_j(\alpha_0 - \alpha_j)}{\alpha_0^2(\alpha_0 + 1)}.$$

The Dirichlet distribution is centered around the normalized parameter vector. Moreover, the total concentration α_0 determines the amount of the variability of $\boldsymbol{p}$. Larger values of α_0 yield smaller variances, which means that probability vectors are tightly concentrated near the mean. The interpretation is useful in count data modeling, as it provides a way to represent heterogeneity in the underlying category probabilities across observational units.

2.2.3 Dirichlet-Multinomial distribution

After having a brief view of Dirichlet distribution, now we introduce Dirichlet-Multinomial distribution.

Definition 2.2.2. If the probability vector $\boldsymbol{p}$ is random and follows a Dirichlet distribution with parameters $\boldsymbol{\alpha} = (\alpha_1, \ldots, \alpha_K)$

$$\boldsymbol{p} \mid \boldsymbol{\alpha} \sim Dir(\boldsymbol{\alpha}),$$

then the marginal distribution of $\boldsymbol{X}_i = (X_{i1}, \ldots, X_{iK})$ conditioned on $\boldsymbol{\alpha}$ is the Multinomial-Dirichlet distribution:

$$\boldsymbol{X}_i \mid \boldsymbol{\alpha} \sim \text{Multinomial-Dirichlet}(n_i, \boldsymbol{\alpha}),$$

where $n_i = \sum_{j=1}^{K} X_{ij}$, which is the i-th sample size for $i = 1, \ldots, N$.

The Dirichlet has good property of being conjugate to multinomial distribution, which makes the marginalization tractable. Thus, we have a closed form of probability mass function

$$\begin{aligned} f(\boldsymbol{X}_i \mid \boldsymbol{\alpha}) &= \int_{\Delta_K} f(\boldsymbol{X}_i \mid \boldsymbol{p}_i) \cdot f(\boldsymbol{p}_i \mid \boldsymbol{\alpha})\, d\boldsymbol{p}_i \\ &= \int_{\Delta_K} \left[\frac{n_i!}{\prod_{j=1}^{K} X_{ij}!} \prod_{j=1}^{K} p_{ij}^{X_{ij}} \right] \left[\frac{1}{B(\boldsymbol{\alpha})} \prod_{j=1}^{K} p_{ij}^{\alpha_j - 1} \right] d\boldsymbol{p}_i \end{aligned}$$

$$
\begin{aligned}
&= \left[\frac{n_i!}{\prod_{j=1}^{K} X_{ij}!} \cdot \frac{1}{B(\boldsymbol{\alpha})}\right] \int_{\Delta_K} \prod_{j=1}^{K} p_{ij}^{X_{ij}+\alpha_j-1}\, d\boldsymbol{p}_i \\
&= \frac{n_i!}{\prod_{j=1}^{K} X_{ij}!} \cdot \frac{B(\boldsymbol{\alpha}+\boldsymbol{X}_i)}{B(\boldsymbol{\alpha})} \\
&= \frac{n_i!}{x_{i1}!x_{i2}!\cdots x_{iK}!} \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_0+n_i)} \prod_{j=1}^{K} \frac{\Gamma(X_{ij}+\alpha_j)}{\Gamma(\alpha_j)} \\
&= \frac{\Gamma(\alpha_0)\Gamma(n_i+1)}{\Gamma(\alpha_0+n_i)} \prod_{j=1}^{K} \frac{\Gamma(X_{ij}+\alpha_j)}{\Gamma(\alpha_j)\Gamma(X_{ij}+1)}, \text{with } \alpha_0 = \sum_{j=1}^{K} \alpha_j. \quad (2.9)
\end{aligned}
$$

This compound structure provides an explanation for overdispersion in multinomial count data. Compared with multinomial model, the Dirichlet-Multinomial distribution allows the outcome or category probabilities to vary across observational units, resulting in increased variance. By law of total variance,

$$
\begin{aligned}
Var(X_{ij}) &= E(Var(X_{ij} \mid p_j)) + Var(E(X_{ij}) \mid p_j) \\
&= n_i \frac{\alpha_j}{\alpha_0}(1-\frac{\alpha_j}{\alpha_0})\frac{\alpha_0+n_i}{\alpha_0+1} \\
&= Var_{Mul}(X_{ij})\frac{\alpha_0+n_i}{\alpha_0+1},
\end{aligned} \tag{2.10}
$$

where $Var_{Mul}(X_{ij})$ is the same as in formula (2.2). For a single sample, the variance expression derived above reveals the key advantage of Dirichlet-Multinomial distribution over multinomial distribution. The term

$$
C = \frac{n_i+\alpha_0}{1+\alpha_0}
$$

acts as a variance inflation factor. Since $n_i > 1$ and $\alpha_0 > 0$, this factor is strictly greater than 1. This implies that the Dirichlet-Multinomial distribution inherently accommodates extra-binomial variation.

Benefited from its conjugacy, we can also derive the posterior distribution of $\boldsymbol{p}_i$ by

$$
\begin{aligned}
f(\boldsymbol{p}_i \mid \boldsymbol{X}_i, \boldsymbol{\alpha}) &= \frac{f(\boldsymbol{p}_i, \boldsymbol{X}_i \mid \boldsymbol{\alpha})}{f(\boldsymbol{X}_i \mid \boldsymbol{\alpha})} \\
&= \frac{f(\boldsymbol{X}_i \mid \boldsymbol{p}_i, \boldsymbol{\alpha}) f(\boldsymbol{p}_i \mid \boldsymbol{\alpha})}{f(\boldsymbol{X}_i \mid \boldsymbol{\alpha})} \\
&= \frac{\Gamma(\alpha_0+n_i)}{\prod_{j=1}^{K}\Gamma(X_{ij}+\alpha_j)} \prod_{j=1}^{K} p_{ij}^{X_{ij}+\alpha_j-1} \sim Dir(X_{ij}+\alpha_j),
\end{aligned} \tag{2.11}
$$

where $n_i = \sum_{j=1}^{K} X_{ij}$ and $\alpha_0 = \sum_{l=1}^{K} \alpha_l$. It also follows a Dirichlet distribution, but with the parameters containing our data.

2.2.4 Binomial-Beta distribution with an example on capture-recapture data

To illustrate the practical necessity of the Dirichlet-Multinomial framework, we consider the analysis of capture-recapture data, specifically the Snowshoe Hare dataset examined by Dorazio & Royle (2003). This example demonstrates how the Dirichlet prior effectively models individual heterogeneity, a common source of overdispersion in biological datasets.

The objective of this study was to estimate the population size based on repeated sampling. For a standard Multinomial distribution model to be valid, one must assume that every individual I has the exact same probability of capture p across all trapping occasions. Under this assumption, the count of captures for any individual should follow a strict Binomial distribution, which is a Multinomial distribution with $K = 2$. However, empirical data frequently violates this assumption. In ecological terms, the population often consists of distinct behavioral subgroups. In this study, two subgroups are defined as trap-happy individuals, which describes animals with a high innate probability of capture, i.e. $p_I \gg \bar{p}$, where $\bar{p}$ stands for the average capture rate, and trap-shy individuals, which describes animals that avoid traps, i.e. $p_I \ll \bar{p}$.

The article demonstrated fitting a standard model with a fixed p leads to significant bias. Because the simple model underestimates the variance since it assumes Var $\propto p(1-p)$, it fails to account for the "heavy tails" of the distribution—the existence of individuals who are caught far more often or far less often than a random process would predict. In the context of the thesis, this is a classic case of overdispersion where the observed variance exceeds the theoretical variance of the simple model.

To resolve this, the authors employ a Multinomial distribution based mixture model equivalent to the Dirichlet-Multinomial, specifically, the Beta-Binomial for $K = 2$. Instead of a fixed parameter, the capture probability for the I-th individual is treated as a random variable drawn from a Beta distribution:

$$p_I \sim \text{Beta}(a, b). \tag{2.12}$$

By estimating the hyperparameters a and b, or equivalently, the precision $\alpha_0 = a + b$, the model explicitly accounts for the variation in catchability among individuals.

2.3 Limitations of Dirichlet-Multinomial distribution

Despite its flexibility in modeling overdispersion, the Dirichlet-Multinomial distribtuion retains several structural limitations. Firstly, the induced covariance structure remains fully determined by the mean parameters and

the total concentration parameter α_0. To calculate the covariance, we need to first derive the cross moments of Dirichlet distribution

$$E(\prod_{j=1}^{K} p_j^{\beta_j}) = \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K}\Gamma(\alpha_j)} \frac{\prod_{j=1}^{K}\Gamma(\beta_j+\alpha_j)}{\Gamma(\sum_{j=1}^{K}(\beta_j+\alpha_j))}. \tag{2.13}$$

The detailed derivation is shown in Appendix A. After this, we can have the covariance of Dirichlet distribution by taking $\beta_k = \beta_l = 1$, $\beta_j = 0$ for elsewhere,

$$Cov(p_k, p_l) = -\frac{\alpha_k \alpha_l}{\alpha_0^2(1+\alpha_0)},\ \alpha_0 = \sum_{j=1}^{K}\alpha_j, \tag{2.14}$$

where $k \neq l$. This negative covariance structure makes the Dirichlet-Multinomial model weak for positively correlated categories. For example, Blei & Lafferty (2005) illustrated this inability in the context of topic modeling. In analyzing a corpus of articles from the journal Science, they identified latent topics such as "Genetics" and "Evolutionary Biology". Intuitively, these topics are positively correlated. An article discussing genetics is more likely to discuss evolution than an article about fossils or geology. However, under a Dirichlet prior, the components of the probability vector are competing for a fixed probability mass. It enforces the logic that if the proportion of "Genetics" words increases, the proportion of "Evolution" words must typically decrease to satisfy the constraint. Lochner (1975) also points out another limitation of Dirichlet distribution that the same variables with same expectations must have same variances. This restricts heterogeneity across outcomes or gcategories even if they have equal mean frequencies. While negative correlation is a natural property of compositional data, the Dirichlet-Multinomial distribution still has a rigid structure allowing only negative correlation. It implies that the correlation between any two categories depends entirely on their means and the global precision α_0.

In such cases, a single Dirichlet–Multinomial distribution is insufficient to capture the heterogeneity present in the data. A natural direction is to consider more flexible extensions that relax the structural constraints of the model. One possibility is to generalize the Dirichlet distribution itself, while another is to introduce additional structure through mixture modeling.

2.4 Generalized Dirichlet-Multinomial distribution

To address the problems existing in multinomial and Dirichlet-Multinomial distribution, Connor & Mosimann (1969) introduced a generalization of Dirichlet distribution by introducing the concept of neutrality.

Definition 2.4.1. Given a random vector of proportions $(P_1, P_2, \ldots, P_K)$, the proportion P_1 is said to be neutral if P_1 is independent of the vector $(\frac{P_2}{1-P_1}, \frac{P_3}{1-P_1}, \ldots, \frac{P_K}{1-P_1})$

In this context, the success probabilities follow a generalized Dirichlet distribution with parameters $\boldsymbol{\alpha}_g = (\alpha_1, \beta_1, \ldots, \alpha_K, \beta_K)$ and has probability mass function

$$f(\boldsymbol{p}_i) = \prod_{j=1}^{K} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} p_{ij}^{\alpha_j - 1} \left(1 - \sum_{l=1}^{j} p_l\right)^{\gamma_j} \tag{2.15}$$

for $\sum_{j=1}^{K} p_{ij} < 1$ and $0 < p_{ij} < 1$ for $j = 1, 2, \ldots K$, where $\alpha_j > 0, \beta_j > 0, \gamma_j = \beta_j - \alpha_{j+1} - \beta_{j+1}$ for $j = 1, \ldots, K-1$ and $\gamma_K = \beta_K - 1$. This will be reduced to Dirichlet distribution when $\beta_j = \alpha_j + \beta_{j+1}$. Lochner (1975) notes that in generalized Dirichlet distribution, the variances do not need to stay the same for variables with equal means, which gives flexibility of establishing the statistical model. The flexible covariance matrix is given based on the general moment function (Connor & Mosimann, 1969; Wong, 1998)

$$E(\prod_{j=1}^{K} p_{ij}^{r_j}) = \prod_{j=1}^{K} \frac{\Gamma(\alpha_j + \beta_j)\Gamma(\alpha_j + r_j)\Gamma(\beta_j + \delta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)\Gamma(\alpha_j + \beta_j + r_j + \delta_j)},$$

where $\delta_j = r_{j+1} + r_{j+2} + \ldots + r_K$ for $j = 1, 2, \ldots, K-1$. The covariance is then derived

$$Cov(p_k, p_l) = E(p_l) \left(\frac{\alpha_k}{\alpha_k + \beta_k + 1} \prod_{j=1}^{k-1} \frac{\beta_j + 1}{\alpha_j + \beta_j + 1} - E(p_k) \right),$$

where $k \neq l$ and $k, l \in \{1, \ldots, K\}$. This structure strengthens ability to model both positive or negative correlations and different variances even when means are equal. Another good property is that the generalized Dirichlet distribution is also conjugate to multinomial distribution (Bouguila, 2008). The joint probability mass function of $\boldsymbol{X}_i, \boldsymbol{p}_i$ is calculated by

$$\begin{aligned} p(\boldsymbol{X}_i, \boldsymbol{p}_i \mid \boldsymbol{\alpha}_g) &= p(\boldsymbol{X}_i \mid \boldsymbol{p}_i)\, p(\boldsymbol{p}_i \mid \boldsymbol{\alpha}_g) \\ &= \frac{\Gamma\left(\sum_{j=1}^{K} X_{ij} + 1\right)}{\prod_{j=1}^{K} \Gamma(X_{ij} + 1)} \prod_{j=1}^{K} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)}\, p_{ij}^{\alpha_j^* - 1} \left(1 - \sum_{l=1}^{j} p_{il}\right)^{\gamma_j^*} \end{aligned} \tag{2.16}$$

where $\alpha_j^* = \alpha_j + X_{ij}$ and $\beta_j^* = \beta_j + X_{i,j+1} + \ldots + X_{i,K}$ for $j = 1, \ldots, K$, $\gamma_j^* = \beta_j - \alpha_{j+1}^* - \beta_{j+1}^*$ for $j = 1, \ldots, K-1$ and $\gamma_K^* = \beta_K^* - 1$. Based on the joint probability mass function, the marginal probability mass function of $\boldsymbol{X}_i$ conditioned on generalized parameter space $\boldsymbol{\alpha}_g$ is obtained by integrating

over $\boldsymbol{p}_i$

$$
\begin{aligned}
p(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_g) &= \int_{\boldsymbol{p_i}} p(\boldsymbol{X}_i, \boldsymbol{p}_i \mid \boldsymbol{\alpha}_g) d\boldsymbol{p}_i \\
&= \frac{\Gamma((\sum_{j=1}^{K} X_{ij}) + 1)}{\prod_{j=1}^{K} \Gamma(X_{ij} + 1)} \prod_{j=1}^{K} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \\
&\int_{\boldsymbol{p}_i} p_{ij}^{\alpha_j^* - 1} \left(1 - \sum_{l=1}^{j} p_{il}\right)^{\gamma_j^*} d\boldsymbol{p}_i \\
&= \frac{\Gamma((\sum_{j=1}^{K} X_{ij}) + 1)}{\prod_{j=1}^{K} \Gamma(X_{ij} + 1)} \prod_{j=1}^{K} \frac{\Gamma(\alpha_j + \beta_j)}{\Gamma(\alpha_j)\Gamma(\beta_j)} \prod_{j=1}^{K} \frac{\Gamma(\alpha_j^*)\Gamma(\beta_j^*)}{\Gamma(\alpha_j^* + \beta_j^*)}.
\end{aligned}
\tag{2.17}
$$

Then getting the posterior of generalized Dirichlet distribution follows similar procedure as Dirichlet distribution

$$
\begin{aligned}
f(\boldsymbol{p}_i \mid \boldsymbol{X}_i, \boldsymbol{\alpha}_g) &= \frac{f(\boldsymbol{p}_i, \boldsymbol{X}_i \mid \boldsymbol{\alpha}_g)}{f(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_g)} \\
&= \prod_{j=1}^{K} \frac{\Gamma(\alpha_j^* + \beta_j^*)}{\Gamma(\alpha_j^*)\Gamma(\beta_j^*)} p_{ij}^{\alpha_j^* - 1} \left(1 - \sum_{l=1}^{j} p_{il}\right)^{\gamma_j^*}
\end{aligned}
\tag{2.18}
$$

which follows a generalized Dirichlet distribution with parameters $\boldsymbol{\alpha^*} = (\alpha_1^*, \beta_1^*, \ldots, \alpha_K^*, \beta_K^*)$.

While the generalized Dirichlet-Multinomial distribution offers superior modeling flexibility, it introduces significant complexity in estimation. The likelihood function in equation (2.17) does not yield a closed-form solution for the maximum likelihood estimates of $\boldsymbol{\alpha}_g$. The number of parameters increases doubly since the rate parameter β_j in $\boldsymbol{\alpha}_g$ could not be canceled out as in Dirichlet-Multinomial distribution. This leads to a much more complex likelihood function and substantially higher computational burden during estimation.

In preliminary experiments conducted for the next chapter, Chapter 4, the generalized model was implemented and evaluated through simulation studies. However, the results indicated that, despite its increased flexibility in modeling covariance structures, the improvement in model performance was limited in the considered settings. In contrast, the computational cost and estimation instability increased noticeably, especially when the sample size was moderate and the number of categories was large. Furthermore, due to the trade-off between model complexity and practical performance, we will not further discuss the simulation details on the generalized Dirichlet-Multinomial distribution based model. This generalized model remains an important theoretical extension and serves as a useful reference for understanding the limitations of simpler models. Instead, we focus on the mixture model of Dirichlet-Multinomial distribution in Chapter 4.

2.5 Dirichlet-Multinomial mixture model

As we have mentioned in section 2.2, the Dirichlet-Multinomial (DM) distribution serves as a flexible extension of the Multinomial model capable of accommodating overdispersion. While the model captures heterogeneity in category or outcome probabilities through random variation in the latent probability vector, it still assumes that all observations arise from a single homogeneous population. In many real-world applications, however, such an assumption is restrictive. In text analysis, for example, documents may belong to different latent topics (Blei et al., 2003). In ecological studies, individuals may form behavioral subgroups (Farine & Whitehead, 2015). In image processing, histograms may correspond to different object classes (Sivic & Zisserman, 2003). In these scenarios, the observed variability is not merely due to random fluctuation around a common mean but is instead driven by the presence of multiple latent subpopulations. To model this structural heterogeneity, we adopt a finite mixture modeling framework.

2.5.1 Latent Heterogeneity and Mixture Modeling

Let

$$\boldsymbol{X}_i = \{X_{i1}, X_{i2}, \ldots, X_{iK}\}, \quad i = 1, 2, \ldots, N$$

denote independent multivariate count vectors, where D represents the number of categories. Each observation satisfies the constraint

$$\sum_{j=1}^{K} X_{ij} = n_i,$$

where n_i is the total count associated with observation i. Examples of such vectors include word frequency counts in documents (Blei et al., 2003), categorical event counts in biological studies (Royle & Dorazio, 2008), or binned feature histograms in image processing (Sivic & Zisserman, 2003).

To capture latent population heterogeneity, we assume that each observation belongs to one of D unobserved clusters. Let the latent variable Z_i denote the cluster assignment of observation i, where

$$Z_i \in \{1, 2, \ldots, D\}.$$

The generative process of the model can be described in two stages. First, a cluster index is sampled from a categorical distribution with probabilities $(\pi_1, \pi_2, \ldots, \pi_D)$, where

$$P(Z_i = d) = \pi_d, \quad d = 1, 2, \ldots, D,$$

and the mixture proportions satisfy

$$\pi_d > 0, \sum_{d=1}^{D} \pi_d = 1.$$

Conditional on the cluster membership $Z_i = d$, the observation $\boldsymbol{X}_i$ is assumed to follow Dirichlet-Multinomial distribution with parameter vector $\boldsymbol{\alpha}_d = (\alpha_{d1}, \ldots, \alpha_{dK})$. Thus,

$$\boldsymbol{X}_i \mid Z_i = d \sim \text{Dirichlet-Multinomial}(n_i, \boldsymbol{\alpha}_d).$$

Under this formulation, each cluster is characterized by its own concentration parameter vector, which determines both the expected category proportions and the dispersion structure within that cluster.

By marginalizing over the latent variable Z_i, the probability mass function of an observation can be expressed as a mixture of Dirichlet-Multinomial components

$$f(\boldsymbol{X}_i \mid \Theta) = \sum_{d=1}^{D} \pi_d f_{DM}(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_d), \tag{2.19}$$

where $f_{DM}(\cdot)$ denotes the Dirichlet-Multinomial probability mass function defined in equation (2.9), and

$$\Theta = \{\pi_1, \ldots, \pi_D, \boldsymbol{\alpha}_1, \ldots, \boldsymbol{\alpha}_D\}. \tag{2.20}$$

represents the complete set of unknown model parameters of observation i. This formulation allows the model to represent complex data distributions by combining several simpler component distributions.

2.5.2 Interpretation of Model Parameters

The parameters of the mixture model have clear statistical interpretations. The mixing weights π_d describe the relative prevalence of D latent clusters within the population. A larger value of π_d indicates that observations belonging to cluster d more frequently. The cluster-specific parameter vector $\boldsymbol{\alpha}_d$ determines the characteristic composition of categories or outcomes within cluster d. The expected category proportion within cluster d is given by

$$E(p_{ij}^{(d)}) = \frac{\alpha_{dj}}{\alpha_{d0}}, \ \alpha_{d0} = \sum_{j=1}^{D} \alpha_{dj},$$

where K is the number of category of outcomes. This is similar to single Dirichlet-Multinomial model from equation (2.13). This expression shows that the normalized vector $\frac{\alpha_{dj}}{\alpha_{d0}}$ represents the mean category distribution for cluster d. In addition to determining the mean composition, the parameter α_{d0} also controls the level of dispersion within the cluster. Larger values of α_{d0} correspond to lower variability, meaning that observations within the cluster tend to concentrate around the mean proportions. Conversely, smaller values of α_{d0} imply greater variability among observations belonging to the same cluster. Consequently, the mixture model captures two different

types of variation in the data. The parameters $\boldsymbol{\alpha}_d$ describe within-cluster variation, while the mixing weights π_d and the differences between the vectors $\boldsymbol{\alpha}_d$ describe variation between clusters.

2.5.3 Likelihood representation

Under the assumption that observations are independent and identically distributed, the log-likelihood function of the mixture model can be written as

$$\mathcal{L}(\Theta, \mathcal{X}) = \log \left(\prod_{i=1}^{N} P(\boldsymbol{X}_i \mid \Theta) \right) = \sum_{i=1}^{N} \log \left(\sum_{d=1}^{D} \pi_d P(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_d) \right). \quad (2.21)$$

This expression contains a summation inside the logarithm, which prevents the objective function from being decomposed into independent components. Direct maximization of equation (2.21) is analytically intractable. This difficulty is a common feature of mixture models and arises from the presence of latent cluster assignments. In practice, the model parameters are therefore estimated using iterative procedures that explicitly account for the unobserved cluster memberships. Among these methods, the Expectation-Maximization algorithm provides a principled and widely used framework for parameter estimation in latent variable models (Dempster et al., 1977). In the next chapter, we will introduce the estimation method based on Expectation-Maximization algorithm and propose an robust extension on this method to show how it can be applied to estimate the parameters of the Dirichlet-Multinomial mixture model.

Chapter 3

Estimation

Having established the Dirichlet–Multinomial mixture model in Chapter 2, the next step is to develop estimation procedures for the unknown model parameters. The presence of latent cluster memberships introduces additional complexity, as the likelihood function involves a mixture structure that cannot be directly optimized in closed form. This motivates the use of iterative estimation methods that explicitly account for the latent variable structure. In this chapter, we adopt the Expectation–Maximization (EM) algorithm, which provides a principled framework for maximum likelihood estimation in models with incomplete data.

3.1 Maximum likelihood estimation

With the parameter space in equation (2.20) and the log-likelihood structure in equation (2.21), the Maximum Likelihood Estimator is formally defined as

$$\hat{\Theta}_{ML} = \arg\max_{\Theta} \mathcal{L}(\Theta, \mathcal{X}). \tag{3.1}$$

Although the likelihood function of the Dirichlet-Multinomial mixture model can be written explicitly, obtaining the maximum likelihood estimator is far from straightforward. The optimization problem presents several challenges that make direct maximization impractical.

The fundamental difficulty arises from the fact that cluster memberships are not directly observed. Each observation could potentially belong to any of the D clusters, and the uncertainty in these assignments must be accounted for during estimation. Ignoring this latent structure would lead to biased parameter estimates and incorrect inference about the cluster structure.

Secondly, mixture models generally exhibit non-convex likelihood surfaces. The objective function may contain multiple local maxima corresponding to different cluster configurations. As a consequence, standard gradient-based optimization methods do not guarantee convergence to the

global optimum and may instead converge to suboptimal solutions depending on the initial parameter values. This issue has been documented on mixture models and latent variable estimation (Wu, 1983).

Another challenge is that the parametrization of the Dirichlet-Multinomial distribution introduces additional computational complexity. Even when the cluster memberships are known, the likelihood equations for the concentration parameters $\boldsymbol{\alpha}_d$ do not admit closed-form solutions. The derivatives of the log-likelihood involve special functions such as the digamma function, which complicates the analytical solutions. Consequently, numerical optimization techniques are typically required to estimate these parameters.

3.1.1 Latent Variable Representation

A key intuition for simplifying the estimation problem is to explicitly introduce the latent cluster membership variables discussed in Section 3.1.1. Instead of viewing each observation as arising from a mixture distribution, we associate each observation $\boldsymbol{X}_i$ with an unobserved cluster indicator $\boldsymbol{Z}_i$. This perspective simplifies the structure of the likelihood function as it replaces the mixture with a conditional model given the latent variables. Let

$$\boldsymbol{Z}_i = \{Z_{i1}, \ldots, Z_{iD}\}$$

denote a vector of binary indicators such that

$$Z_{id} = \begin{cases} 1, & \text{if observation } i \text{ originates from cluster } d \\ 0, & \text{otherwise,} \end{cases} \tag{3.2}$$

where $d = 1, \ldots, D$. Under this formulation, each observation is associated with exactly one cluster with the constraint

$$\sum_{d=1}^{D} Z_{id} = 1.$$

If the cluster assignment $\boldsymbol{Z}_i$ was observed, the likelihood function would take a much simpler form. The complete-data likelihood would be

$$L_c(\Theta) = \prod_{i=1}^{N} \prod_{d=1}^{D} \left[\pi_d \, f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_d)\right]^{Z_{id}}.$$

Taking logarithms yields the complete-data log-likelihood

$$\ell_c(\Theta) = \sum_{i=1}^{N} \sum_{d=1}^{D} Z_{id} \left(\log \pi_d + \log f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_d)\right). \tag{3.3}$$

Compared with the observed-data likelihood, this expression has a much simpler additive structure. Each observation contributes only to the component corresponding to its cluster assignment, which significantly simplifies the optimization problem. The cluster indicators Z_{id} are not observed in practice, which must be inferred from the datasets.

3.1.2 Motivation for the EM Algorithm

Finally, the fundamental difficulty arises from the fact that latent cluster assignments are not directly observed. Each observation could potentially belong to any of the D clusters, and the uncertainty in these assignments must be accounted for during estimation. Ignoring this latent structure would lead to biased parameter estimates and incorrect inference about the cluster structure. As already mentioned the Expectation–Maximization (EM) algorithm provides such a framework. Introduced by Dempster et al. (1977), the EM algorithm is specifically designed for maximum likelihood estimation in the presence of incomplete or missing data. In the context of mixture models, the missing data correspond precisely to the unobserved cluster memberships.

The EM algorithm proceeds by iteratively performing two steps. In the first step, the expected values of the latent variables are computed given the current parameter estimates. In the second step, the parameters are updated by maximizing the expected complete-data log-likelihood. Under mild regularity conditions, this procedure produces a sequence of parameter estimates that monotonically increases the observed-data likelihood (Wu, 1983). The next section describes the EM algorithm in detail and derives the specific update equations used for estimating the parameters of the Dirichlet–Multinomial mixture model.

3.2 The Expectation-Maximization (EM) Algorithm

In the context of the Dirichlet–Multinomial mixture model, the missing data correspond to the unknown cluster memberships of the observations. The EM algorithm alternates between estimating the expected cluster memberships given the current parameter values and updating the model parameters based on these estimated assignments. This iterative process continues until convergence of the likelihood function.

3.2.1 Complete data formulation

To apply the EM algorithm, we consider the complete-data formulation of the model by introducing latent indicator variables Z_{id} in equation (3.2) with the constraint $\sum_{d=1}^{D} Z_{id} = 1$. If the latent variables Z_{id} were observed, the likelihood function would take a simpler form. The complete-data likelihood is given by

$$L_c(\Theta) = \prod_{i=1}^{N} \prod_{d=1}^{D} \left[\pi_d \, f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_d)\right]^{Z_{id}} . \tag{3.4}$$

Taking logarithms yields the complete-data log-likelihood

$$\ell_c(\Theta) = \sum_{i=1}^{N} \sum_{d=1}^{D} Z_{id} \left(\log \pi_d + \log f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_d)\right) . \tag{3.5}$$

This separable structure makes each observation contribute only to the component corresponding to its cluster assignment. The EM algorithm exploits this structure by replacing the unobserved variables Z_{ik} with their conditional expectations.

3.2.2 Expectation step

In the expectation step, the algorithm computes the conditional expectation of the latent variable Z_{id} given observed data and the current parameter estimates of $\hat{\boldsymbol{\alpha}}_d$. These expectations correspond to the posterior probabilities that observation i belongs to cluster d.

Let $\hat{\Theta}^{(t)}$ denote the parameter estimate at iteration t. The posterior responsibility of cluster d for observation i is given by

$$\hat{\tau}_{id}^{(t)} = P(Z_{id} = 1 \mid \boldsymbol{X}_i, \hat{\Theta}^{(t)}) = \frac{\hat{\pi}_d^{(t)} f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \hat{\boldsymbol{\alpha}}_d^{(t)})}{\sum_{l=1}^{D} \hat{\pi}_l^{(t)} f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \hat{\boldsymbol{\alpha}}_l^{(t)})}. \tag{3.6}$$

The quantity $\hat{\tau}_{ik}$ is often referred to as the responsibility, as it measures the degree to which cluster k explains observation i. The EM algorithm performs soft clustering, allowing each observation to belong partially to multiple clusters. With the posterior probability in equation (2.11), the expected complete-data log-likelihood becomes

$$Q(\Theta \mid \hat{\Theta}^{(t)}) = \sum_{i=1}^{N} \sum_{d=1}^{D} \hat{\tau}_{id}^{(t)} \left(\log \hat{\pi}_d + \log f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \hat{\boldsymbol{\alpha}}_d^{(t)}) \right). \tag{3.7}$$

The function $Q(\Theta \mid \hat{\Theta}^{(t)})$ serves as the objective function for the next maximization step.

3.2.3 Maximization step

In the maximization step, the parameter estimates are updated by maximizing the expected complete-data log-likelihood obtained in the E-step. The mixing proportions can be updated in closed form. Let

$$N_d = \sum_{i=1}^{N} \hat{\tau}_{id}^{(t)}$$

denote the expected number of observations assigned to cluster d. The updated mixing weights are then given by

$$\hat{\pi}_d^{(t+1)} = \frac{N_d}{N}.$$

These updates ensure that the mixing proportions remain non-negative and sum to one.

The update of the concentration parameters $\hat{\boldsymbol{\alpha}}_d$ is more involved. Substituting the Dirichlet-Multinomial density into the objective function yiels

$$\sum_{i=1}^{N} \hat{\tau}_{id}^{(t)} \log f_{\mathrm{DM}}(\boldsymbol{X}_i \mid \hat{\boldsymbol{\alpha}}_d^{(t)}).$$

Maximizing this expression does not yield a closed-form solution. The derivatives involve digamma functions arising from the derivatives of the gamma function in the Dirichlet-Multinomial distribution. Thus, the parameters must be estimated using numerical optimization techniques.

In practice, we employ a constrained optimization method to ensure that the concentration parameters remain positive. In the implementation described in the next chapter, the optimization is performed using L-BFGS-B algorithm (Hubert & Arabie, 1985) , which is well suited for large-scaled optimization problems with simple bound constraints.

3.2.4 Convergence Properties

A key property of EM algorithm is that the observed data likelihood does not decrease at each iteration, i.e.

$$\ell(\hat{\Theta}^{(t+1)}) \geq \ell(\hat{\Theta}^{(t)}).$$

This monotonic improvement property guarantees stable progress of the algorithm toward a local maximum of the likelihood function (Wu, 1983). However, the algorithm may converge to different local maxima depending on the initialization of the parameters. For this reason, multiple random initializations are often used in practice to improve the chances of finding a better solution. Involving such a strategy would introduce an additional layer of computational complexity in the simulation study, and therefore it is not considered in this work due to constraint time. The iterative procedure continues until a convergence criterion is satisfied. Common stopping rules include monitoring the change in the log-likelihood or the change in parameter values between successive iterations.

3.3 Robust Extension: Noise Component

While mixture models provide a flexible framework for modeling heterogeneous populations, standard maximum likelihood estimation remains sensitive to contaminated observations. In clustering applications, even a small proportion of atypical data points can substantially influence parameter estimates and distort the resulting cluster structure. This issue arises because the likelihood function assigns substantial influence to observations that are

poorly explained by existing clusters. During EM procedure, such observations may pull cluster parameters toward regions of the parameter space that do not represent the underlying population structure.

The sensitivity to outliers is also mentioned on mixture models and robust statistics (McLachlan & Peel, 2000). In the context of count data clustering, extreme observations may occur for several reasons. For instance, some observations may correspond to rare categories, unusual combinations of counts or measurement errors.

To mitigate these effects, it is useful to introduce a mechanism that allows atypical observations to be accommodated without distorting the primary cluster structure. One widely adopted strategy is to augment the mixture model with an additional noise component that captures observations that do not fit well into any of the primary clusters.

3.3.1 Mixture Model with a Noise Component

The robust model extends the Dirichlet-Multinomial mixture by introducing an additional component designed to accommodate outliers. Let $d = 0$ denote this noise component. The model then becomes

$$P(\boldsymbol{X}_i \mid \Theta) = \pi_0 f(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_0) + \sum_{d=1}^{D} \pi_d f_{DM}(\boldsymbol{X}_i \mid \boldsymbol{\alpha}_d). \tag{3.8}$$

In this formulation, the parameters $\pi_0, \pi_1 \ldots, \pi_D$ represent the mixing proportions of the noise component and the K primary clusters. These proportions satisfy the constraint

$$\sum_{d=1}^{D} \pi_d = 1.$$

The parameter vector $\boldsymbol{\alpha}_0$ defines the distribution of the noise component. The noise parameter is typically fixed rather than estimated from the data. A common choice is to use a relatively diffuse parameter vector such as $\boldsymbol{\alpha}_0 = (1, 1, \ldots, 1)$, which corresponds to a broad distribution over the categorical simplex. This choice ensures that the noise component provides a low-informative baseline capable of accommodating observations that are not well explained by primary clusters.

The interpretation of this model is straightforward. Observations that resemble one of the cluster distributions are assigned high posterior probability for that cluster. Observations that are poorly explained by all clusters instead receive higher probability from the noise component.

3.3.2 Parameter Estimation under the Robust Model

In the extended model, parameter estimation can also be performed by EM algorithm with the modification that the mixture contains $D+1$ components

now with a specific noise component. In the expectation step, posterior probabilities are computed for all components, including the noise component. The responsibility estimate of latent cluster d for observation i is

$$\hat{\tau}_{id} = \frac{\hat{\pi}_d f_{DM}(\boldsymbol{X}_i \mid \hat{\boldsymbol{\alpha}}_d)}{\sum_{l=0}^{D} \hat{\pi}_l f_{DM}(\boldsymbol{X}_i \mid \hat{\boldsymbol{\alpha}}_l)}, \quad d = 0, 1, \ldots, D. \tag{3.9}$$

The quantity τ_{i0} represents the posterior probability that observation i belongs to the noise components. In the maximization step, the update of the mixing proportions follows the same structure as in the standard EM algorithm but extending the component index to 0

$$\hat{\pi}_d^{(t+1)} = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_{id}^{(t)}, \quad d = 0, 1, \ldots, D. \tag{3.10}$$

The cluster-specific parameter $\hat{\boldsymbol{\alpha}}_1, \ldots, \hat{\boldsymbol{\alpha}}_D$ are updated by maximizing the expected complete-data log-likelihood as shown in section 3.3. The parameter $\boldsymbol{\alpha}_0$ associated with the noise component is fixed and does not need optimization. This design ensures that the noise component works as a flexible distribution rather than competing directly with the cluster components.

3.3.3 Interpretation of the Noise Component

The inclusion of the noise component provides two important benefits. First, it reduces the influence of atypical observations on the cluster parameter estimates. When an observation is poorly explained by the cluster distributions, the model can assign it to the noise component instead of forcing it into one of the clusters. As a result, the parameter estimates of the cluster components remain more stable. Secondly, the model provides a natural estimate of the contamination level in the data. The estimate

$$\hat{\pi}_0 = \frac{1}{N} \sum_{i=1}^{N} \hat{\tau}_{i0}$$

is the expected proportion of observations that belong to the noise component. In practice, the estimation procedure may assign an excessive number of observations to the noise component if the noise component is too flexible. To avoid this situation, regularization or constraints on the mixing weight π_0 can be introduced during estimation. In this thesis, we manually control the maximum contaminated proportion to be 0.2 so that the noise component does not dominate the mixture model.

3.4 Evaluation Metrics

To assess the performance of the proposed clustering methods, we consider both statistical accuracy and computational efficiency. In simulations, the

true cluster memberships are known in advance. This allows direct comparison between the estimated clustering results and the underlying generated results. At the same time, the computational cost of the estimation procedure will also be taken into account, since iterative mixture-model algorithms may become expensive as the number of clusters or the contamination level increases. For these reasons, the simulation study in Chapter 4 evaluates the methods using three complementary aspects: clustering accuracy, robustness to contamination, and computational time.

3.4.1 Clustering accuracy

For clustering accuracy, we use the measure adjusted rand index (ARI). The ARI quantifies the true proportion between the real cluster labels and the estimated cluster assignments while correcting for chance agreement. Unlike raw classification accuracy, the ARI is invariant to permutation of cluster labels, which is essential in mixture modeling because component labels are not identifiable in absolute sense.

Definition 3.4.1. (Hubert & Arabie, 1985). Let U denote the true clustering space and V denote the estimated clustering space. Let r_{ij} denote the number of observations that are simultaneously assigned to cluster i in U and cluster j in V. The marginal sums are defined by

$$a_i = \sum_j r_{ij}, b_j = \sum_i r_{ij},$$

which represents the sized of the clusters in the true and estimated partitions. The Adjusted Rand Index is defined as

$$\text{ARI} = \frac{\sum_{i,j} \binom{r_{ij}}{2} - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}{\frac{1}{2}\left[\sum_i \binom{a_i}{2} + \sum_j \binom{b_j}{2}\right] - \frac{\sum_i \binom{a_i}{2} \sum_j \binom{b_j}{2}}{\binom{n}{2}}}.$$

This expression adjusts the classical Rand Index by subtracting the expected agreement under random partitioning. As a result, the ARI takes values in the interval $[-1, 1]$. A value close to 1 indicates strong agreement between the true and estimated clustering, whereas values near 0 indicate performance comparable to random assignment. The use of ARI is appropriate in simulation studies because the true group memberships are known and the index provides a standardized way to compare clustering results across different experimental settings. In this thesis, ARI is used as the main performance measure for both the standard EM algorithm and the robust extension with a noise component. For the robust model, some observations may be assigned to the additional noise component rather than to one of the primary clusters. In such cases, the ARI is computed using

the observations assigned to the substantive clusters only. This reflects the fact that the purpose of the noise component is not to define an additional cluster of interest, but to absorb atypical observations that would otherwise distort the clustering structure.

3.4.2 Clustering robustness

In addition to clustering accuracy, it is important to evaluate how the robust model behaves under contamination. The key quantity in this context is the estimated mixing proportion of the noise component, denoted by $\hat{\pi}_0$. Within the robust mixture model introduced in equation (3.8), π_0 represents the proportion of observations assigned to the noise component. Its estimate therefore provides an interpretable indicator of the amount of contamination detected by the model. If the data are approximately clean, $\hat{\pi}_0$ should remain small. As the contamination level increases, the estimated value of $\hat{\pi}_0$ is expected to increase correspondingly, although in practice regularization constraints may prevent this estimate from growing without bound.

This quantity is not used as a clustering accuracy measure in the strict sense, but it provides important diagnostic information about the robustness mechanism. In the simulation study, the behavior of $\hat{\pi}_0$ across different contamination levels is therefore examined alongside ARI. Together, these two quantities provide a more complete picture: ARI measures how well the cluster structure is recovered, while $\hat{\pi}_0$ describes how strongly the model relies on the noise component to handle atypical observations.

3.4.3 Computational time

Since both the standard and robust methods are estimated using iterative EM-type procedures, computational efficiency is also considered. The total runtime required to fit the model to a single dataset is recorded using system timing functions. In the simulation study, runtime is compared across methods and contamination settings in order to assess the practical computational cost of robustness. The recorded runtime corresponds to the full execution time of the algorithm for one dataset, including all EM iterations and all internal numerical optimization steps required for updating the Dirichlet–Multinomial parameters. It is therefore a measure of total computational effort rather than time per iteration.

Chapter 4

Simulation Studies

The probabilistic generation of the data can be viewed as a two-stage process. We adopt a hierarchical framework where the probability vector for each observation is treated as a latent variable. This generative process occurs in two stages. In Stage 1 called latent process, a specific probability vector $\boldsymbol{p}_i = (p_{i1}, \ldots, p_{iK})$ is drawn from a Dirichlet distribution $Dir(\boldsymbol{\alpha}_i)$ for each sample $i = 1, \ldots, N$. In Stage 2 called observation process, the observed count vector $\boldsymbol{X}_i$ is generated from conditional on the realization of $\boldsymbol{p}_i$. The resulting marginal distribution of the observed counts $\boldsymbol{X}_i$, unconditional on the latent $\boldsymbol{p}_i$, is obtained by integrating out the probability vector over the D-dimensional simplex Δ_K.

4.1 Data generation

The datasets are generated according to a Dirichlet-Multinomial mixture model with D latent clusters and K categories or outcomes. For each observation $i = 1, \ldots, N$, a latent cluster assignment $Z_i \in \{1, \ldots, D\}$ is first generated such that each cluster contains an equal number of observations. Conditional on $Z_i = d$, the observation $X_i = (X_{i1}, \ldots, X_{iK})$ follows a Dirichlet-Multinomial distribution with parameter $(N_i, \boldsymbol{\alpha}_i)$, where $\boldsymbol{\alpha}_i \in \mathbb{R}_+^K$ is the cluster-specific concentration parameter. To introduce heterogeneity in sample sizes, the total count n_i is randomly drawn for each observation from a discrete uniform distribution over a fixed range. This reflects practical settings in which observations may correspond to different total counts.

To introduce variability across observations, the total count n_i is randomly generated for each observation. Specifically, n_i is drawn independently from a discrete uniform distribution over a fixed range. This avoids overly restrictive settings in which all observations have identical total counts and better reflects practical data scenarios. For each configuration, datasets are generated independently and both the standard Dirichlet–Multinomial mixture model and its robust extension are fitted using the estimation pro-

cedures described in Chapter 3. Each method is initialized randomly and applied once per dataset, consistent with the computational constraints of the study.

The simulation is repeated over multiple independent runs for each parameter configuration. Clustering performance is evaluated using the adjusted Rand index, which compares the estimated clustering with the true cluster assignments. Robustness is assessed through the estimated noise proportion $\hat{\pi}_0$, and computational efficiency is measured by the runtime required for model estimation. The results are aggregated across runs to assess both the average performance and variability of the methods.

4.2 Contamination mechanism

In this thesis, we introduce a controlled proportion of contaminated observations to evaluate the robustness of the proposed methods. This allows assessment of model performance under deviations from the assumed Dirichlet-Multinomial mixture structure. Let $c \in [0, 1)$ denote the contamination rate. For each generated dataset, a proportion c of the observations is selected at random and replaced by outliers. These contaminated observations are generated independently from a distribution that does not correspond to any of the cluster-specific Dirichlet-Multinomial components used in the data-generating process.

In the implementation, for each contaminated observation i, the outliers are drawn from a Multinomial distribution with size $n_i^{(0)}$ and a probability vector $\boldsymbol{p}_i^{(0)}$, which is different from its ground true $\boldsymbol{p}_i$. The probability vector is drawn as the weighted average of two probability vectors with a magnet parameter, default to be 0.5 or from a symmetric Dirichlet

$$\boldsymbol{p}_i^{(0)} \sim \text{Dir}(\delta \mathbb{I}), \tag{4.1}$$

where $\delta > 0$ is a small constant and $\mathbb{I}$ denotes the K-dimensional all-one vector. In this thesis, we take $\delta = \frac{1}{K}$, which satisfies the condition of simplex in equation (2.5). Conditional on $\boldsymbol{p}_i^{(0)}$, the count vector is generated as

$$\boldsymbol{X}_i^{(0)} \sim \text{Multinomial}(n_i^{(0)}, \boldsymbol{p}_i^{(0)}).$$

The contaminated sample size is randomly drawn from $[60, 120)$ with probability 0.6, which aligns with the clean data or a predefined larger number with probability 0.4, i.e.

$$n_i^{(0)} = \begin{cases} \text{random draw from [60,120)}, & \text{with probability 0.6} \\ \text{pre-defined N}, & \text{with probability 0.4.} \end{cases} \tag{4.2}$$

To add randomness and evaluate its ability when dealing with extremely contaminated data, the pre-defined N is chosen to be much larger than

the normal simulation sample size, for example, $N = 800$ in our simulation. This construction produces approximately uniform count vectors across categories, with limited variability driven only by sampling noise. Consequently, these observations do not align with any of the cluster-specific Dirichlet-Multinomial components and cannot be well represented by the mixture model.

4.3 Experiments

The experiments are conducted across a range of values for the number of clusters D, with the sample size per cluster fixed to ensure a balanced design. For each configuration, independent data sets are generated and both the standard Dirichlet-Multinomial mixture model and its robust extension are fitted using the estimation procedures described in Chapter 3. The performance metrics are then computed for each run and aggregated across repetitions. To investigate the effect of contamination, the contamination rate c is varied over a predefined grid, including the uncontaminated case. In this thesis, the rates are chosen to be $\{0, 0.05, 0.1, 0.15, 0.2\}$. For each value of c, a new set of datasets is generated and simulated, allowing the degradation of clustering performance as contamination rate increases.

In addition to contamination, the structural complexity of the data-generating model is controlled through three distinct configurations, referred to Tier 1, Tier 2 and Tier 3 in this thesis. These tiers correspond to progressively more challenging clustering scenarios. In Tier 1, clusters are well separated and differ primarily through a single dominant category. In Tier 2, clusters exhibit multiple dominant categories, leading to partial overlap between cluster distributions. In Tier 3, cluster differences are defined in a high-dimensional latent space, resulting in more subtle and less interpretable distinctions. For each tier and each contamination level, the simulation is repeated over multiple independent runs. The adjusted rand index is used to evaluate clustering accuracy, while the estimated noise proportion $\hat{\pi}_0$ provides a measure of robustness for the proposed model. The runtime of the estimation procedure is recorded to assess computational efficiency. The pseudo-code of the simulation is given in Appendix B.

4.4 Simulation results

4.4.1 Single dominant index

The results for Tier 1 across $D = 3, \ldots, 14$ are presented in Figure 4.1. This setting corresponds to well-separated clusters characterized by a single dominant category. When there is no contamination, both the standard model based on EM algorithm and the robust model, which contains the

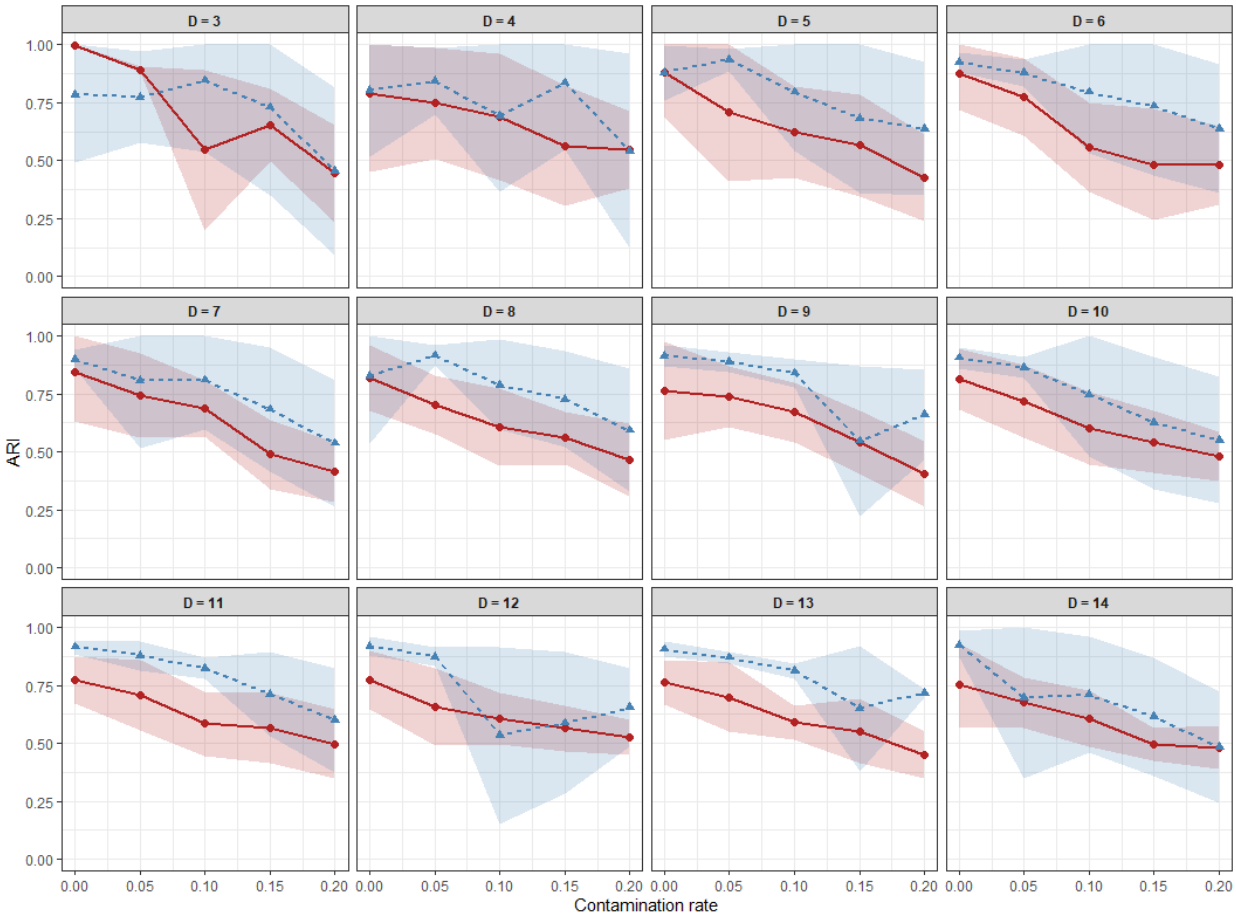


Figure 4.1: The lines are the average of ARI in standard model (red) and in the robust model (blue) in Tier 1. The shaded area describes the variation over the simulated ARI values in the standard model (red) and the robust model (blue). The number of latent clusters varies from $D = 3$ to $D = 14$.

extra category for capturing noise, achieve high latent clustering accuracy across all values of D. The ARI is close to one in most cases, indicating that the true cluster structure is effectively recovered. This confirms that under ideal conditions, the additional noise component does not negatively affect performance.

As contamination rate increases, a consistent pattern emerges across all D values. The latent cluster assignment accuracy of the standard model decreases steadily as contamination rate increases. The decline is sharper at higher contamination levels. This behavior reflects the sensitivity of the standard mixture model to outliers. In contrast, the robust model shows a slower decreasing trend. This demonstrates the effectiveness of the noise component in absorbing atypical observations and preserving the integrity of

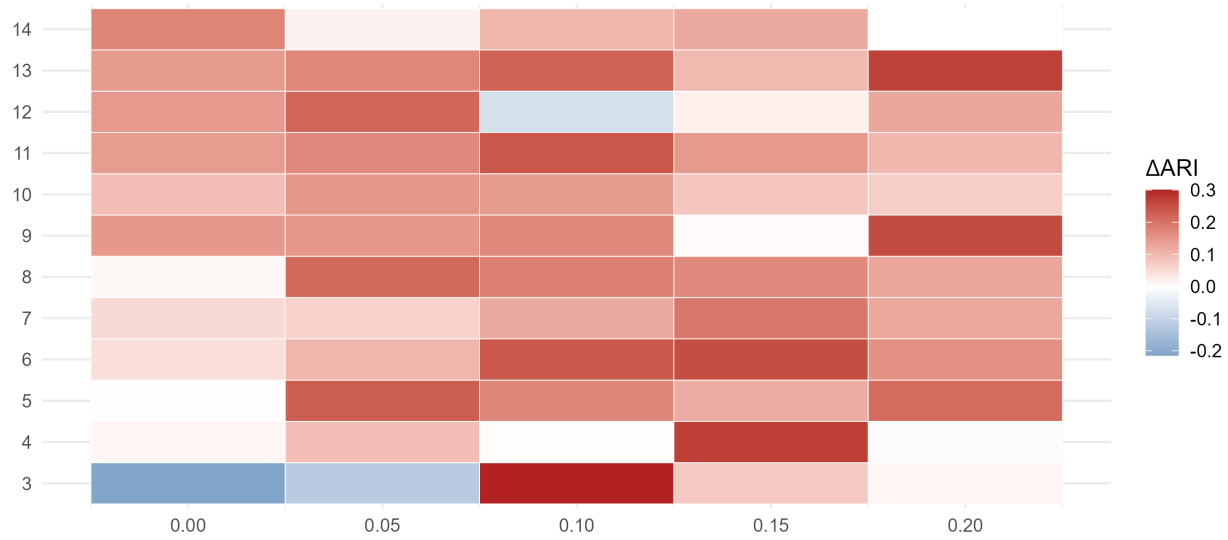


Figure 4.2: The difference between ARI of the robust model and the standard model (ΔARI) in Tier 1. Positive values (red) indicate that the robust model outperforms the standard model, while negative values (blue) indicate the opposite. The horizontal axis represents the contamination level, and the vertical axis represents the number of latent clusters D.

the cluster-specific parameters. The variability of the ARI in the simulated runs, represented by the shaded bands, is generally larger for the robust model. This effect is visible at moderate contamination level, where uncertainty in assigning borderline observations to either clusters or the noise component leads to increased variability across simulation runs. The increased variability observed for the robust model is consistent with the theoretical properties of finite mixture models and robust estimators. The inclusion of a noise component introduces additional flexibility and latent structure, which reduces bias under contamination but increases variability due to uncertainty in cluster assignments and the presence of multiple local optima in the likelihood function. In contrast, the standard model exhibits more stable but systematically biased performance under contamination.

The runtime results show a clear increase in computational cost for the robust model compared to the standard algorithm. The robust model consistently requires more computation due to the additional parameters and the more complex likelihood structure. For larger values of categories, The difference in runtime becomes substantial, indicating a trade-off between robustness and computational efficiency.

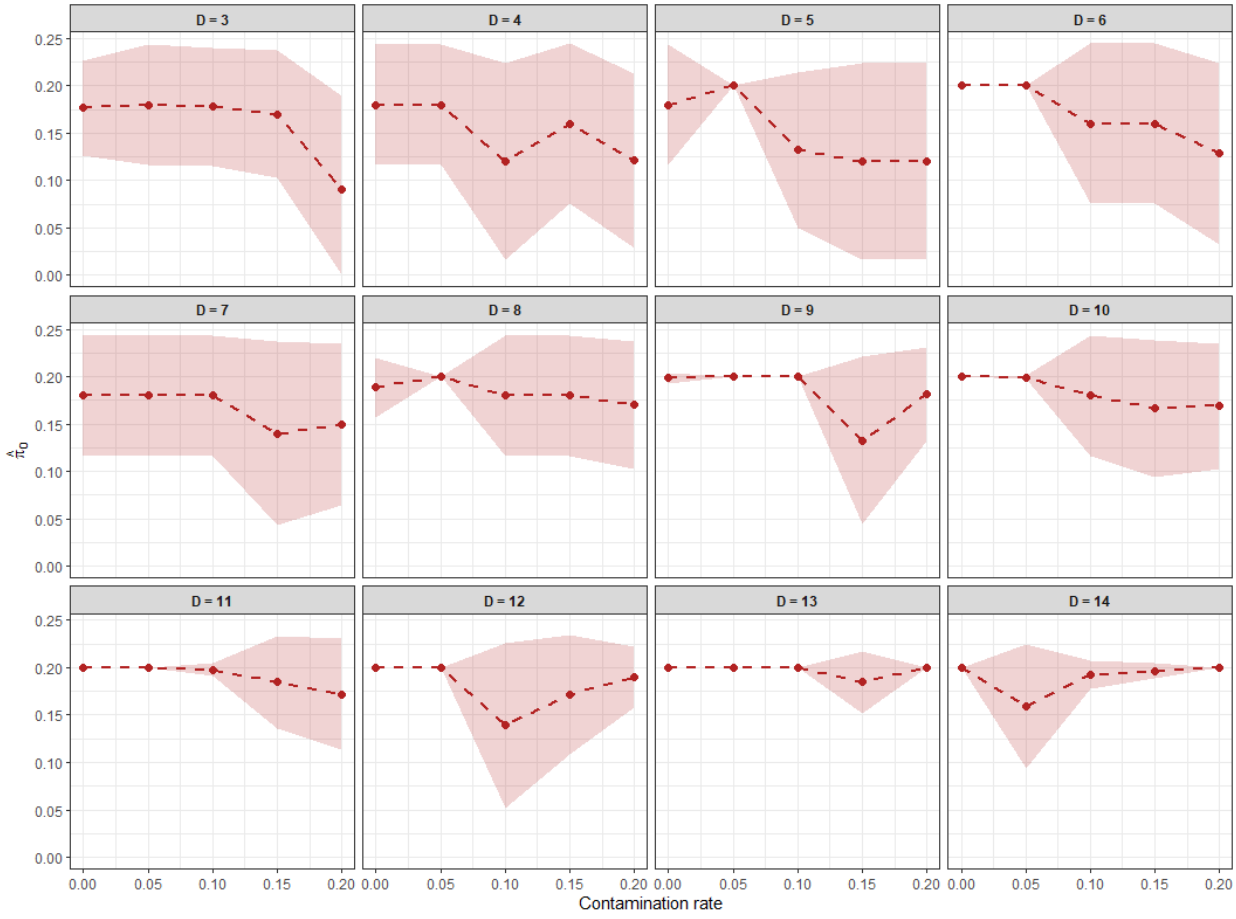


Figure 4.3: Estimated contaminated proportion $\hat{\pi}_0$ in Tier 1. The dashed red lines show the average estimated contamination proportion, while the shaded regions represent variation across simulated values.

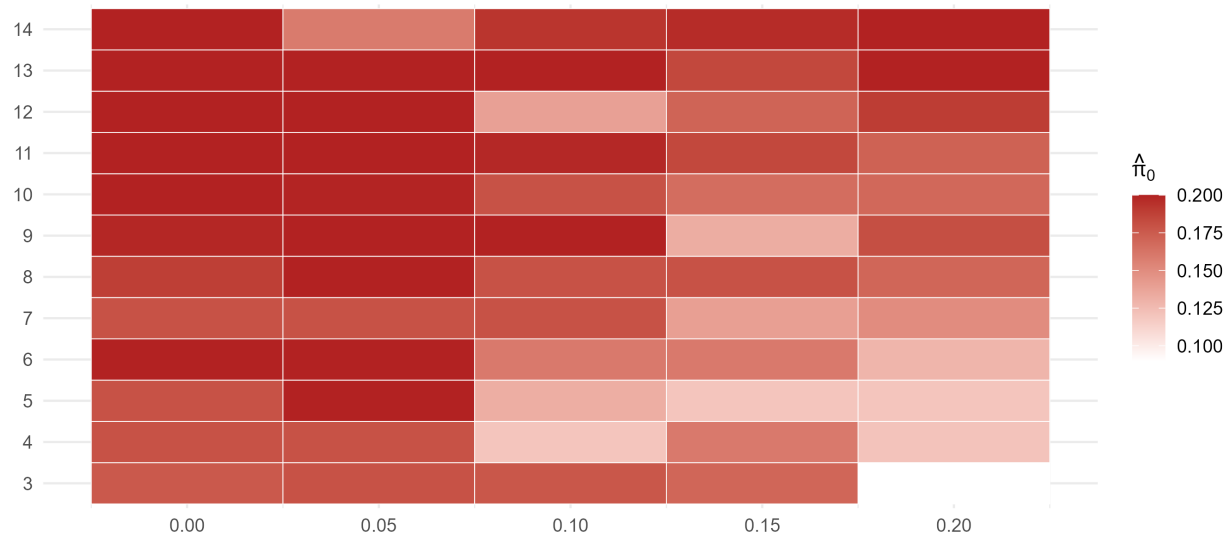


Figure 4.4: Estimated contaminated proportion $\hat{\pi}_0$ heatmap across all simulated runs in Tier 1. Darker red cells indicate larger estimated contaminated proportions. The horizontal axis represents the contamination level, and the vertical axis represents the number of latent clusters D.

4.4.2 Double dominant indices

The simulation Tier 2 represents an intermediate level of complexity, where clusters of latent variables are defined by block-wise structures rather than a single dominant category. In this setting, multiple categories contribute to each latent cluster, leading to partial overlap between latent distributions. Compared to the previous simulation, the separation between clusters is weaker, but the structure remain more interpretable.

The results in Figure 4.5 indicate that, in the absence of contamination, both standard and robust model achieve high category assignment accuracy across all values of D. However, the ARI values are consistently lower than those observed in Tier 1 simulation, implying that the increased difficulty due to overlapping structures. As the number of latent clusters increases, a gradual decline in category assignment accuracy is observed, especially for the standard model. From Figure 4.5 and 4.6, we can also notice that as the contamination rate increases, the performance of the two methods diverges more clearly. The standard model exhibits a steady decrease trend across all D values, which indicates its sensitivity to contaminated observations. In contrast, the robust model shows a more stable performance across all contamination levels. This advantage is more obvious when the contamination level exceeds 10%. To make this advantage more visible, Figure 4.6 provides the difference of adjusted rand index between the robust and standard

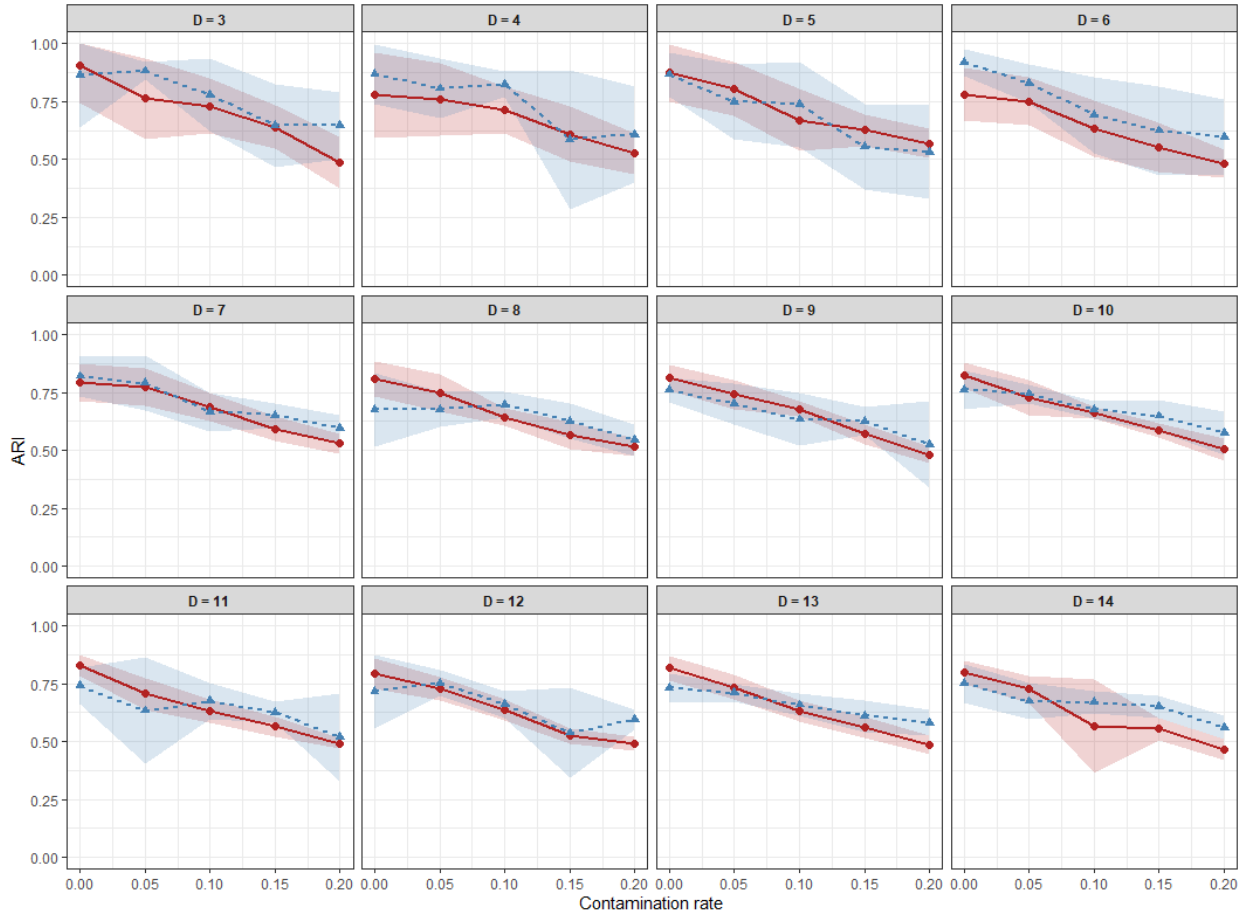


Figure 4.5: The lines are the average of ARI in standard model (red) and in the robust model (blue) in Tier 2. The shaded area describes the variation over the simulated ARI values in the standard model (red) and the robust model (blue). The number of latent clusters varies from $D = 3$ to $D = 14$.

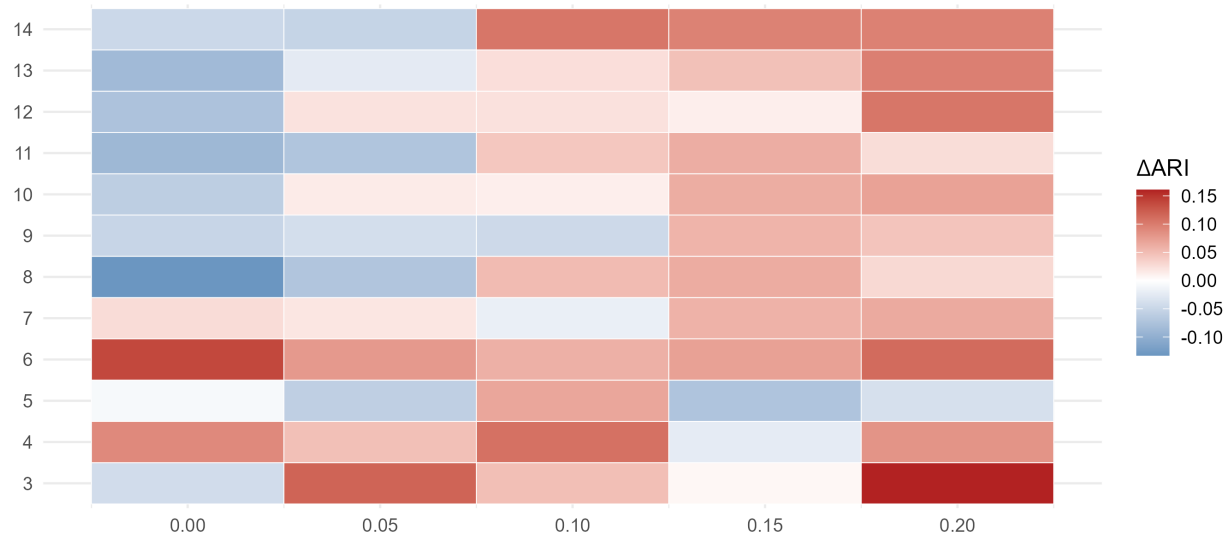


Figure 4.6: The difference between ARI of the robust model and the standard model (ΔARI) in Tier 2. Positive values (red) indicate that the robust model outperforms the standard model, while negative values (blue) indicate the opposite. The vertical axis represents the number of latent clusters D, and the horizontal axis represents the contamination level.

model, denoted as ΔARI. For a large proportions of parameter setting, the ΔARI are positive, implying that the robust model outperforms the standard EM model. This improvement is more significant for smaller values of D, where the cluster structure is still sufficiently distinct for the noise component to effectively isolate contaminated observations. However, for certain configurations, such as $D = 5$, the difference between the two methods becomes negligible or slightly negative. This suggests that the benefit of robustness is not uniform and depends on the balance between cluster overlap and contamination.

The estimated noise proportion $\hat{\pi}_0$ exhibits a clear and monotonic increase with respect to the contamination rate across all values of D. Compared to the previous scenario, the estimates appear more stable and closer to the real outlying parameter. Especially when the number of clusters increases, the estimated proportion becomes more accurate and stable, indicating that the noise component aligns well with the underlying contamination level in this setting. The smooth behavior of $\hat{\pi}_0$ suggests that the model is able to effectively distinguish between structured observations and outliers when cluster overlap is large.

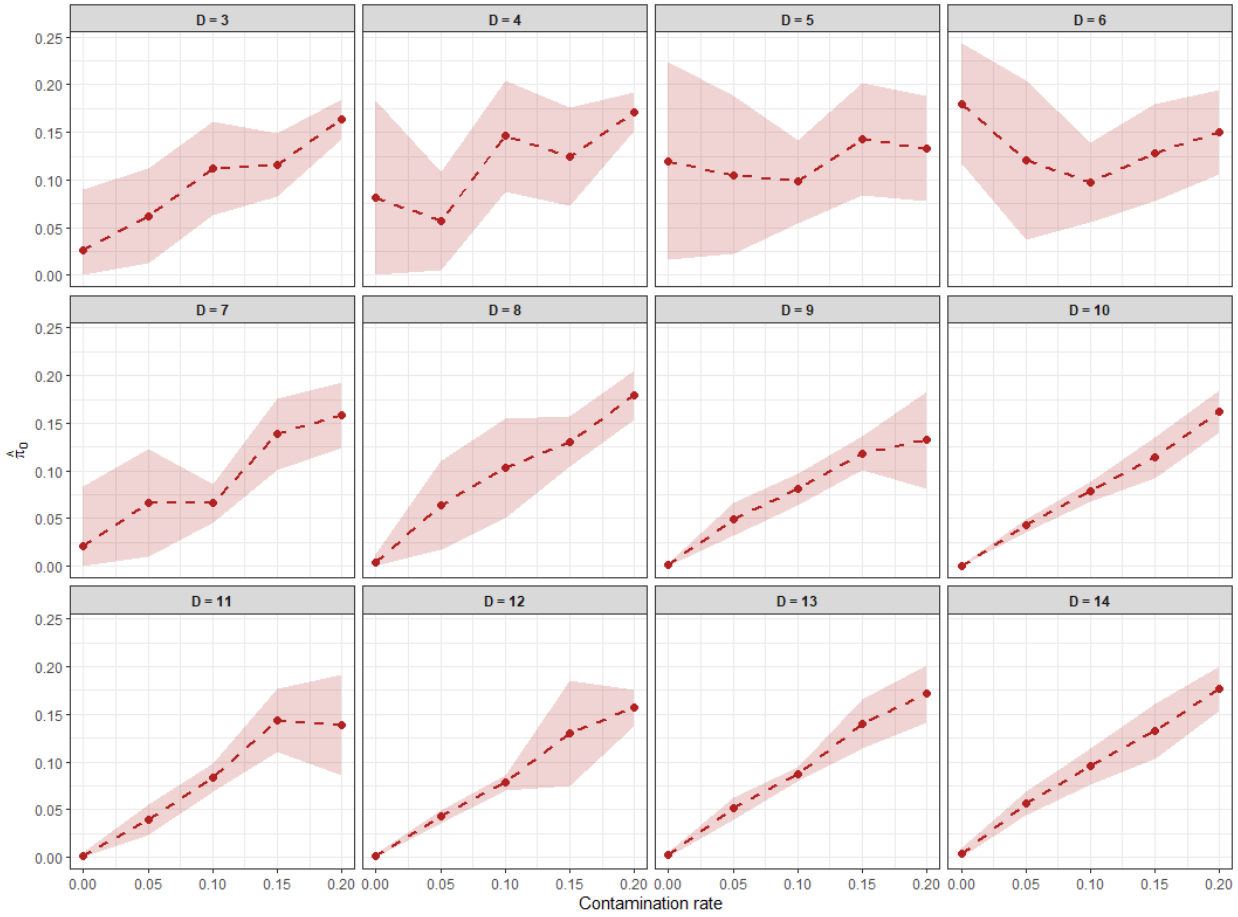


Figure 4.7: Estimated contaminated proportion $\hat{\pi}_0$ in Tier 2. The dashed red lines show the average estimated contamination proportion, while the shaded regions represent variation across simulated values.

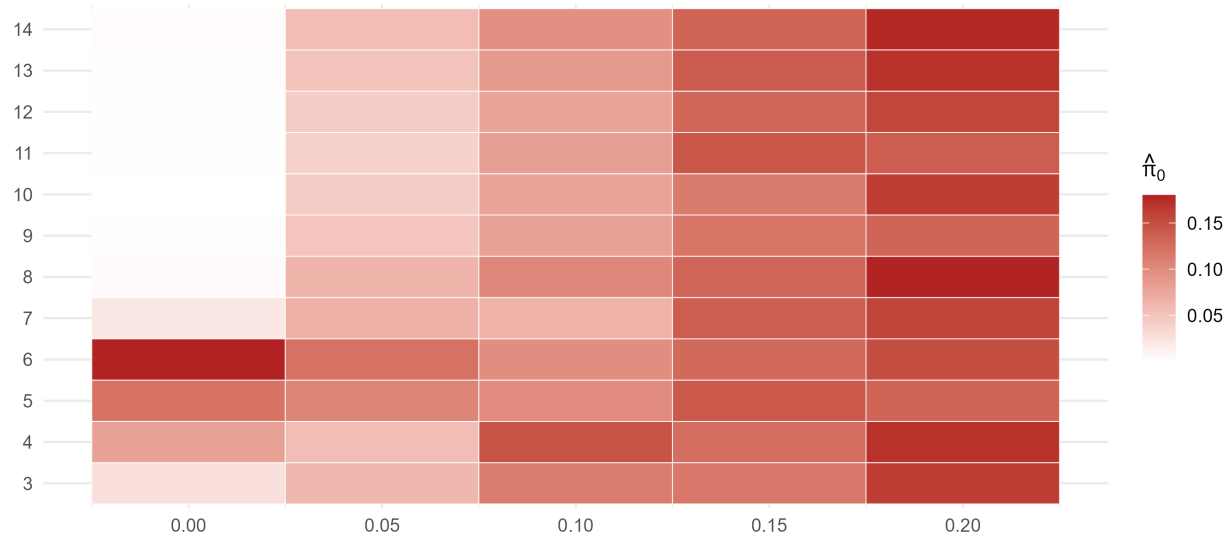


Figure 4.8: Estimated contaminated proportion $\hat{\pi}_0$ heatmap across all simulated runs in Tier 2. Darker red cells indicate larger estimated contaminated proportions. The horizontal axis represents the contamination level, and the vertical axis represents the number of latent clusters D.

4.4.3 Orthogonal structure

The Tier 3 simulation represents the most challenging setting in the simulation study, where cluster structures are defined through orthogonal directions in a high-dimensional space. The cluster identity is concentrated in a subset of coordinates, and this construction enforces a form of orthogonality between cluster parameters. The objective is to investigate the performance of the proposed methods in situations where cluster separation is weak and cannot be captured through marginal dominance. Such a structure is motivated by applications in high-dimensional count data, where signals are often encoded in subtle combinations of features rather than in individual components. A representative example arises in text analysis, where topics are not defined by a few exclusive keywords but rather by distributions over a large vocabulary, and different topics may share substantial overlap (Griffiths & Steyvers, 2004). Similar phenomena appear in genomics (Efron, 2007) and compositional data analysis (Aitchison, 1982), where signals are distributed across many coordinates. In these settings, clustering relies on joint patterns rather than marginal contrasts, making the problem inherently more difficult.

From a mathematical perspective, let $\boldsymbol{\alpha}_d \in \mathbb{R}_+^K$ denote the Dirichlet parameter associated with cluster d. In this level of simulation, the vectors $\{\boldsymbol{\alpha}_d\}_{d=1}^D$ are constructed so that their normalized forms $\boldsymbol{\theta}_d = \frac{\boldsymbol{\alpha}_d}{\|\boldsymbol{\alpha}_d\|_2}$ are ap-

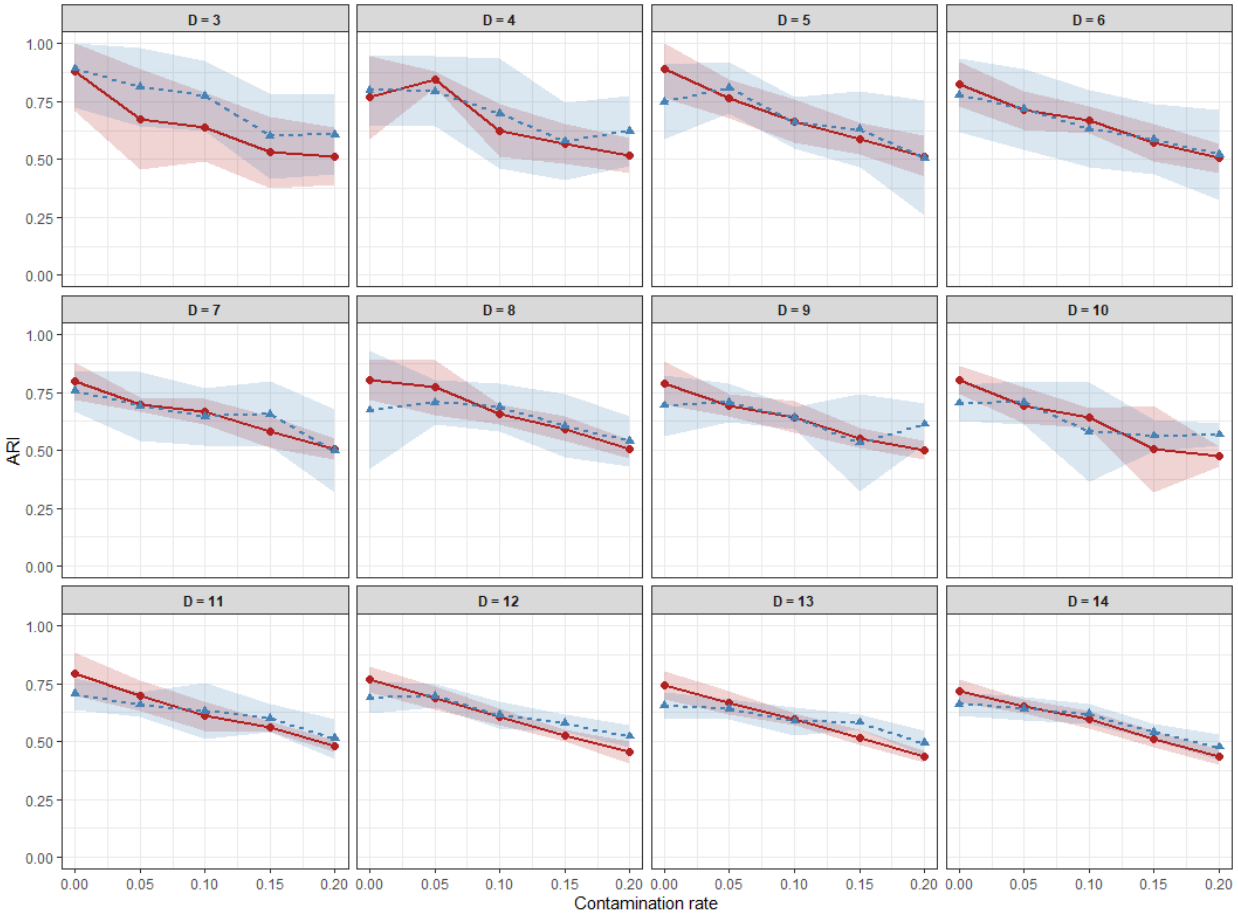


Figure 4.9: The lines are the average of ARI in standard model (red) and in the robust model (blue) in Tier 3. The shaded area describes the variation over the simulated ARI values in the standard model (red) and the robust model (blue). The number of latent clusters varies from $D = 3$ to $D = 14$.

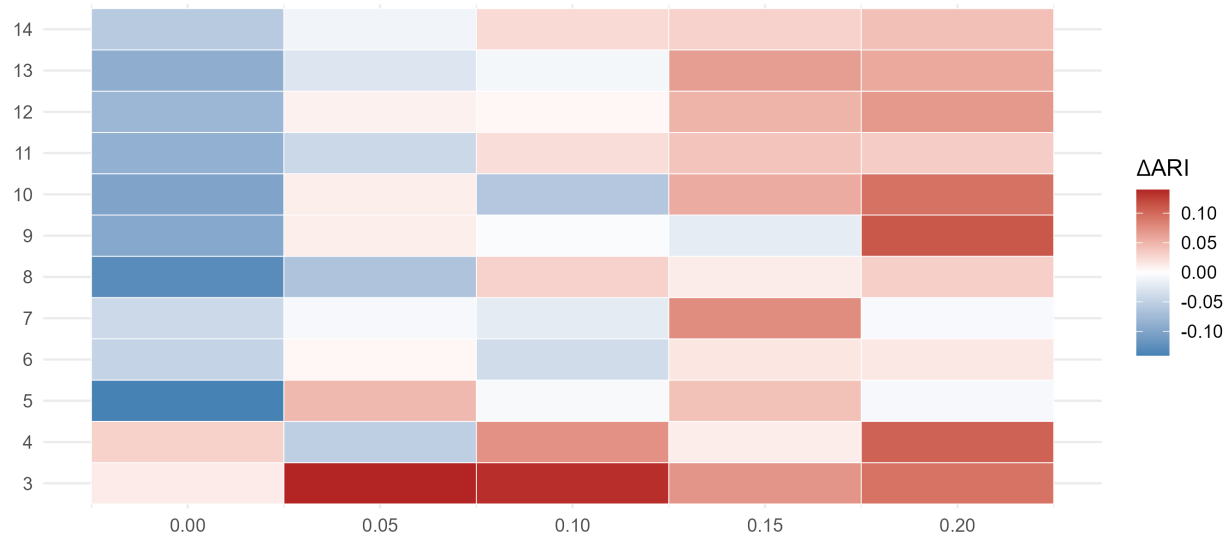


Figure 4.10: The difference between ARI of the robust model and the standard model (ΔARI) in Tier 3. Positive values (red) indicate that the robust model outperforms the standard model, while negative values (blue) indicate the opposite. The vertical axis represents the number of latent clusters D, and the horizontal axis represents the contamination level.

proximately orthogonal, in the sense that $\boldsymbol{\theta}_d'\boldsymbol{\theta}_l \approx 0$ for $d \neq l$. This implies that cluster differences are distributed across coordinates rather than concentrated in any single component. However, orthogonality in this sense does not imply strong separation in the induced probability distributions. Indeed, for the corresponding Dirichlet–Multinomial components f_k the Kullback–Leibler divergence (Csiszar, 1975; Kullback & Leibler, 1951)

$$KL(f_d \| f_l) = \int f_d(x) log \frac{f_d(x)}{f_l(x)} dx \approx 0$$

remains relatively small, reflecting the diffuse nature of the distributions. Consequently, the mixture model operates in a regime of weak identifiability, where different components are difficult to distinguish based on observed data.

The simulation is implemented by constructing cluster-specific Dirichlet parameters that encode orthogonal structure through disjoint support patterns. For each cluster $d = 1, \ldots, D$, a subset of coordinates $S_d \subset \{1, \ldots, K\}$ is assigned such that $S_d \cap S_l = \varnothing$ for $d \neq l$, and $\bigcup_{d=1}^{D} S_d \subseteq \{1, \ldots, K\}$. The parameter vector $\boldsymbol{\alpha}_d \in \mathbb{R}_+^K$ is then defined component-wise by

$$\alpha_{dj} = \begin{cases} a, & j \in S_d, \\ b, & j \notin S_d, \end{cases} \quad \text{with } a > b > 0. \tag{4.3}$$

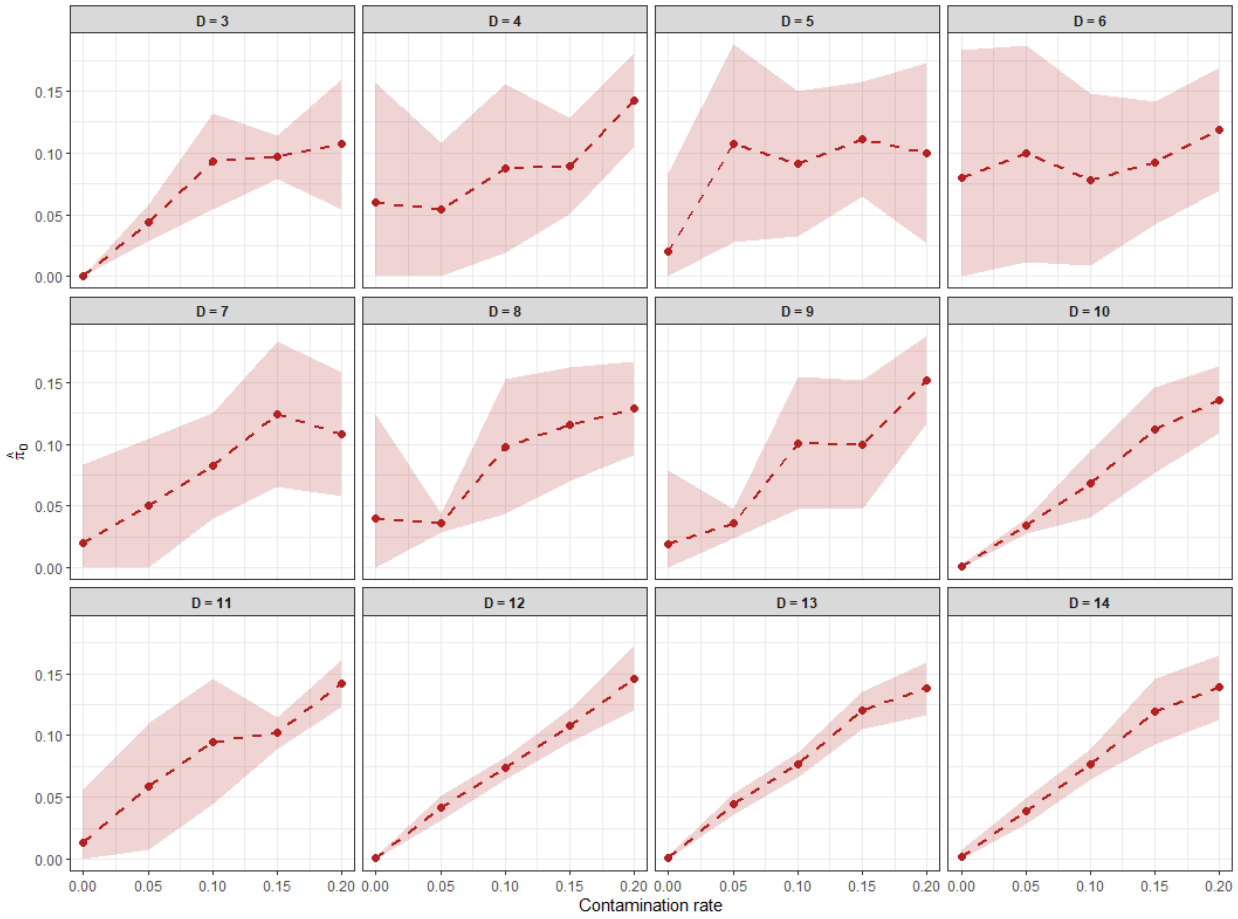


Figure 4.11: Estimated contaminated proportion $\hat{\pi}_0$ in Tier 3. The dashed red lines show the average estimated contamination proportion, while the shaded regions represent variation across simulated values.

This construction ensures that each cluster places relatively higher mass on its assigned coordinates while maintaining a positive baseline across all dimension. Although the supports S_d are disjoint, the presence of the baseline parameter b in equation (4.3) induces overlap between clusters in the resulting Dirichlet-Multinomial distributions. The resulting $\boldsymbol{\alpha}_d$ are distinct but not strongly separated in the probability simplex, reflecting the intended orthogonal yet diffuse structure. Given these parameters, observations are generated according to the procedures described in section 4.1.

The results for this scenario are summarized in Figure 4.9, 4.10, 4.11 and 4.12. Latent structures in this setting do not exhibit clearly interpretable dominant categories, and the differences between latent distributions are more subtle. As a result, substantial overlap between clusters is present even in the absence of contamination. The result in Figure 4.9 shows that, clus-

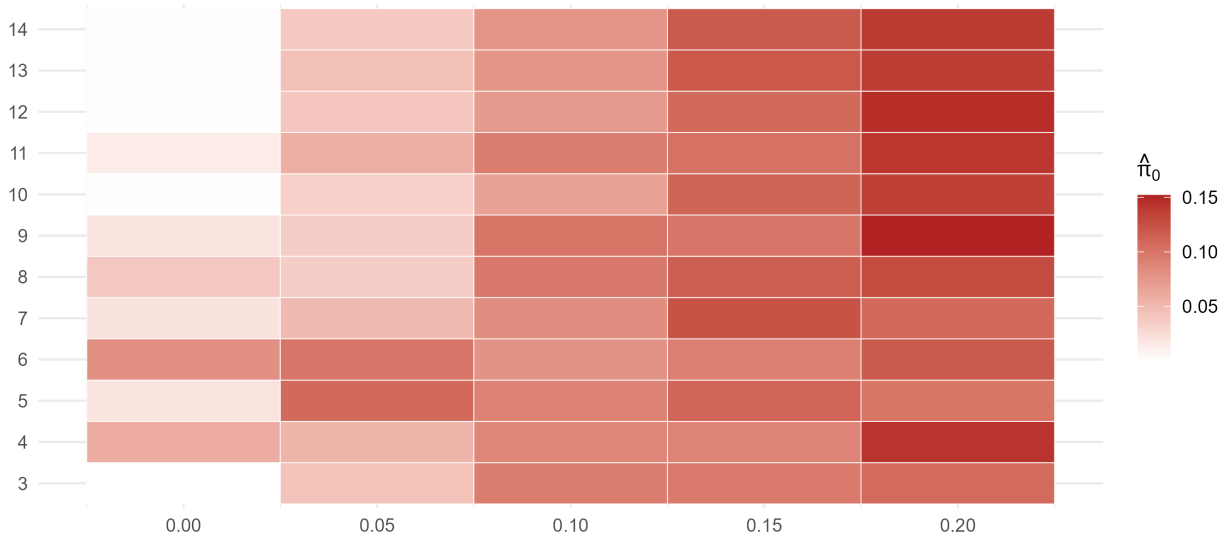


Figure 4.12: Estimated contaminated proportion $\hat{\pi}_0$ heatmap across all simulated runs in Tier 3. Darker red cells indicate larger estimated contaminated proportions. The horizontal axis represents the contamination level, and the vertical axis represents the number of latent clusters D.

tering accuracy is uniformly lower than in the previous scenarios across all numbers of latent clusters and contamination levels. Even when no contamination is present, the ARI does not attain values close to one, particularly for larger dimensions. This is consistent with the theoretical expectation that diffuse and overlapping component distributions lead to weak separation.

As the contamination rate increases, the performance of both methods deteriorates, but the pattern differs across dimensions. The difference between the two methods is less pronounced than in earlier scenarios. Figure 4.10 reveals a heterogeneous pattern that for smaller values of D, the robust model often achieves moderate improvements, particularly at intermediate contamination levels. In contrast, for larger values of D, the advantage becomes inconsistent and may even reverse, with the standard model performing comparably or better at small contamination levels while the robust model deals better with larger contamination. This behavior can be explained by examining the role of the noise component. The estimated noise proportion $\hat{\pi}_0$ increases steadily with the contamination rate. The robust model is more responsive to the presence of outliers. However, $\hat{\pi}_0$ also takes non-negligible values even at low contamination levels, such as when the true number of underlying latent clusters D lying between 3 and 10. This implies that the model interprets part of the intrinsic cluster overlap as noise. Formally, when the component densities $f_d(x)$ are similar across D, it may be assigned to the noise component due to its higher responsibility as computed

in equation (3.9) when $d = 0$.

Overall, the results demonstrate that, in settings where cluster differences are distributed across many dimensions and separation is weak, the benefits of incorporating a noise component could work more robustly when the dimensions and contamination levels are higher at the cost of running time. This highlights the dependence of robust mixture methods on the underlying geometric structure of the data and underscores the challenges of assigning the correct latent clusters in high-dimensional, weakly separated regimes.

Chapter 5

Conclusion

This thesis investigated the performance of Dirichlet-Multinomial mixture models for clustering count data under varying structural complexity and contamination levels. A robust extension incorporating a noise component was evaluated against the standard EM-based estimation procedure. The simulation study was designed to systematically vary both the degree of latent cluster separability and the level of contamination, providing a comprehensive assessment of model behavior across different regimes.

The results demonstrate that the standard Dirichlet-Multinomial mixture model performs well in settings where latent variable distributions are sufficiently distinct and the data are less contaminated. The standard model is able to recover the underlying structure with high accuracy. However, this behavior deteriorates as contamination happens. Since the standard model lacks a mechanism to accommodate atypical observations, such data points are assigned to one of the mixture components, leading to biased parameter estimates and degraded clustering performance. The robust extension addresses this limitation by introducing an additional noise component, which allows observations that are poorly explained by all cluster components to be absorbed by a diffuse baseline distribution. As a result, the influence of outliers on the estimation of cluster-specific parameters is significantly reduced.

These findings highlight the importance of incorporating robustness into mixture models for count data. In practical applications, where data quality cannot be guaranteed, the proposed robust method offers a simple yet effective solution to the problem. At the same time, the robust model introduces additional computational complexity due to the increased number of components and the iterative nature of the estimation procedure. The running time shows similar patterns that the robust model takes longer time to find an optimized solution due to the extra parameters. Although the computational cost remains manageable in the simulation settings considered in this thesis, further optimization may be required for large-scale applications. This

work could also provide some future research directions for future research. From a methodological perspective, one may consider using more flexible distributions to model the probability vector, such as generalized Dirichlet distribution or logistic-normal formulations. Another interesting direction is to conduct the estimation under the framework of density power divergence since maximum likelihood is established based on Kullback-Leibler divergence. Additionally, the Bayesian inference will also be considered to solve such problems.

Bibliography

Aitchison, J. (1982). "The Statistical Analysis of Compositional Data". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 44.2, pp. 139–160. DOI: https://doi.org/10.1111/j.2517-6161.1982.tb01195.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1982.tb01195.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1982.tb01195.x.

Blei, D. M. & Lafferty, J. D. (2005). "Correlated topic models". In: *Proceedings of the 19th International Conference on Neural Information Processing Systems*. NIPS'05. Vancouver, British Columbia, Canada: MIT Press, pp. 147–154.

Blei, D. M., Ng, A. Y. & Jordan, M. I. (2003). "Latent dirichlet allocation". In: *Journal of machine Learning research* 3.Jan, pp. 993–1022.

Bouguila, N. (2008). "Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions". In: *IEEE Transactions on Knowledge and Data Engineering* 20.4, pp. 462–474. DOI: 10.1109/TKDE.2007.190726.

Budig, S., Vogel, C. & Schaarschmidt, F. (2026). "Multiple Comparisons With Overdispersed Multinomial Data: Methods, Properties and Application". In: *Pharmaceutical Statistics* 25.1. e70073 7747353, e70073. DOI: https://doi.org/10.1002/pst.70073. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/pst.70073. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/pst.70073.

Church, K. W. & Gale, W. A. (1995). "Poisson mixtures". In: *Natural Language Engineering* 1.2, pp. 163–190. DOI: 10.1017/S1351324900000139.

Connor, R. J. & Mosimann, J. E. (1969). "Concepts of Independence for Proportions with a Generalization of the Dirichlet Distribution". In: *Journal of the American Statistical Association* 64.325, pp. 194–206. DOI: 10.1080/01621459.1969.10500963. eprint: https://www.tandfonline.com/doi/pdf/10.1080/01621459.1969.10500963. URL: https://www.tandfonline.com/doi/abs/10.1080/01621459.1969.10500963.

Coretto, P. & Hennig, C. (2016). "Robust improper maximum likelihood: tuning, computation, and a comparison with other methods for robust Gaussian clustering". In: *Journal of the American Statistical Association* 111.516, pp. 1648–1659.

Csiszar, I. (1975). "I-Divergence Geometry of Probability Distributions and Minimization Problems". In: *The Annals of Probability* 3.1, pp. 146–158. DOI: 10.1214/aop/1176996454. URL: https://doi.org/10.1214/aop/1176996454.

Dempster, A. P., Laird, N. M. & Rubin, D. B. (1977). "Maximum Likelihood from Incomplete Data Via the EM Algorithm". In: *Journal of the Royal Statistical Society: Series B (Methodological)* 39.1, pp. 1–22. DOI: https://doi.org/10.1111/j.2517-6161.1977.tb01600.x. eprint: https://rss.onlinelibrary.wiley.com/doi/pdf/10.1111/j.2517-6161.1977.tb01600.x. URL: https://rss.onlinelibrary.wiley.com/doi/abs/10.1111/j.2517-6161.1977.tb01600.x.

Dorazio, R. M. & Royle, J. A. (2003). "Mixture Models for Estimating the Size of a Closed Population When Capture Rates Vary among Individuals". In: *Biometrics* 59.2, pp. 351–364. DOI: https://doi.org/10.1111/1541-0420.00042.

Efron, Bradley (2007). "Correlation and large-scale simultaneous significance testing". In: *Journal of the American Statistical Association* 102.477, pp. 93–103.

Elkan, C. (2006). "Clustering documents with an exponential-family approximation of the Dirichlet compound multinomial distribution". In: ICML '06. Pittsburgh, Pennsylvania, USA: Association for Computing Machinery, pp. 289–296. ISBN: 1595933832. DOI: 10.1145/1143844.1143881. URL: https://doi.org/10.1145/1143844.1143881.

Eriksson, A. W., Frisk, M. D. & Nevanlinna, H. R. (1962). "ABO and Rhesus Group Frequencies and their Distribution among Blood Donors in Finland". In: *Acta Genetica et Statistica Medica* 12.3/4, pp. 312–321. ISSN: 03652785, 2571743X. URL: http://www.jstor.org/stable/45103220 (visited on 12/29/2025).

Farine, Damien R & Whitehead, Hal (2015). "Constructing, conducting and interpreting animal social network analysis". In: *Journal of animal ecology* 84.5, pp. 1144–1163.

Fraley, C. & Raftery, A. E. (2002). "Model-based clustering, discriminant analysis, and density estimation". In: *Journal of the American statistical Association* 97.458, pp. 611–631.

Frigyik, A. Bé., Kapila, A. & Gupta, M. R. (2010). "Introduction to the Dirichlet Distribution and Related Processes". In: URL: https://api.semanticscholar.org/CorpusID:8763665.

Griffiths, T. L. & Steyvers, M. (2004). "Finding scientific topics". In: *Proceedings of the National Academy of Sciences* 101.suppl_1, pp. 5228–5235. DOI: 10.1073/pnas.0307752101. eprint: https://www.pnas.org/doi/pdf/10.1073/pnas.0307752101. URL: https://www.pnas.org/doi/abs/10.1073/pnas.0307752101.

Hubert, L. & Arabie, P. (1985). "Comparing partitions". In: *Journal of classification* 2.1, pp. 193–218.

Katz, S. M. (1996). "Distribution of content words and phrases in text and language modelling". In: *Natural Language Engineering* 2.1, pp. 15–59. DOI: 10.1017/S1351324996001246.

Kullback, S. & Leibler, R. A. (1951). "On Information and Sufficiency". In: *The Annals of Mathematical Statistics* 22.1, pp. 79–86. ISSN: 00034851. URL: http://www.jstor.org/stable/2236703 (visited on 05/03/2026).

Lochner, R. H. (1975). "A Generalized Dirichlet Distribution in Bayesian Life Testing". In: *Journal of the Royal Statistical Society. Series B (Methodological)* 37.1, pp. 103–113. ISSN: 00359246. (Visited on 12/29/2025).

McLachlan, G. J. & Peel, D. (2000). *Finite mixture models*. John Wiley & Sons.

Royle, J. A. & Dorazio, R. M. (2008). *Hierarchical modeling and inference in ecology: the analysis of data from populations, metapopulations and communities*. Elsevier.

Rudin, W. (1976). *Principles of Mathematical Analysis*. McGraw-Hill, Inc.

Sivic, J. & Zisserman, A. (2003). "Video Google: A text retrieval approach to object matching in videos". In: *Proceedings ninth IEEE international conference on computer vision*. IEEE, pp. 1470–1477.

Wong, T.T. (1998). "Generalized Dirichlet distribution in Bayesian analysis". In: *Applied Mathematics and Computation* 97.2, pp. 165–181. ISSN: 0096-3003. DOI: https://doi.org/10.1016/S0096-3003(97)10140-0. URL: https://www.sciencedirect.com/science/article/pii/S0096300397101400.

Wu, J. C. F. (1983). "On the Convergence Properties of the EM Algorithm". In: *The Annals of Statistics* 11.1, pp. 95–103. DOI: 10.1214/aos/1176346060. URL: https://doi.org/10.1214/aos/1176346060.

Appendix A

Proofs for section 2.2.2

In the derivation of joint probability density function $f(\boldsymbol{p}, Y)$, consider the one-to-one transformation of

$$\begin{bmatrix} Z_1 \\ Z_2 \\ \vdots \\ Z_K \end{bmatrix} \xrightarrow{g} \begin{bmatrix} \frac{Z_1}{\sum_{l=1}^{K} Z_l} \\ \frac{Z_2}{\sum_{l=1}^{K} Z_l} \\ \vdots \\ Z_K \end{bmatrix} = \begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ Y \end{bmatrix}.$$

We can derive the inverse transformation as

$$\begin{bmatrix} p_1 \\ p_2 \\ \vdots \\ Y \end{bmatrix} \xrightarrow{g^{-1}} \begin{bmatrix} p_1 Y \\ p_2 Y \\ \vdots \\ \left(1 - \sum_{j=1}^{K-1} p_j\right) Y \end{bmatrix}.$$

Based on Theorem of change of variables, we have

$$\begin{aligned} f(\boldsymbol{p}, Y) &= f(g^{-1}(\boldsymbol{p}))|J| \\ &= \prod_{l=1}^{K-1} \frac{\beta^{\alpha_l}}{\Gamma(\alpha_l)} (p_l Y)^{\alpha_l - 1} e^{-\beta p_l Y} \frac{\beta^{\alpha_K}}{\Gamma(\alpha_K)} \left[\left(1 - \sum_{j=1}^{K-1} p_j\right) Y \right]^{\alpha_K - 1} e^{-\beta(1 - \sum_{j=1}^{K-1} p_j) Y^{K-1}} \\ &= \prod_{l=1}^{K} \frac{\beta_l^{\alpha}}{\Gamma(\alpha_l)} p_l^{\alpha_l - 1} Y^{\alpha_0 - 1} e^{-\beta Y} \end{aligned} \tag{A.1}$$

Then we can derive the marginal function for $\boldsymbol{p}$ given the independence relationship

$$\begin{aligned} f(\boldsymbol{p}) &= \frac{f(\boldsymbol{p}, Y)}{f(Y)} \\ &= \frac{\prod_{l=1}^{K} \frac{\beta_l^{\alpha}}{\Gamma(\alpha_l)} p_l^{\alpha_l - 1} Y^{\alpha_0 - 1} e^{-\beta Y}}{\frac{\beta_0^{\alpha}}{\Gamma(\alpha_0)} Y^{\alpha_0 - 1} e^{-\beta Y}} \\ &= \frac{\Gamma(\alpha_0)}{\prod_{l=1}^{K} \Gamma(\alpha_l)} \prod_{l=1}^{K} p_l^{\alpha_l - 1} \end{aligned} \tag{A.2}$$

Cross moments of Dirichlet distribution

$$\begin{aligned} E(\prod_{j=1}^{K} p_j^{\beta_j}) &= \int_{\Delta_K} \prod_{j=1}^{K} p_j^{\beta_j} \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} p_j^{\alpha_j - 1} d\boldsymbol{p} \\ &= \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \int_{\Delta_K} \prod_{j=1}^{K} p_j^{\beta_j + \alpha_j - 1} d\boldsymbol{p} \\ &= \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \frac{\prod_{j=1}^{K} \Gamma(\beta_j + \alpha_j)}{\Gamma(\sum_{j=1}^{K} (\beta_j + \alpha_j))} \int_{\Delta_K} \frac{\Gamma(\sum_{j=1}^{K} (\beta_j + \alpha_j))}{\prod_{j=1}^{K} \Gamma(\beta_j + \alpha_j)} \prod_{j=1}^{K} p_j^{\beta_j + \alpha_j - 1} d\boldsymbol{p} \\ &= \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \frac{\prod_{j=1}^{K} \Gamma(\beta_j + \alpha_j)}{\Gamma(\sum_{j=1}^{K} (\beta_j + \alpha_j))} \end{aligned} \tag{A.3}$$

Covariance struture of Dirichlet distribution

$$\begin{aligned} Cov(p_k, p_l) &= E(p_k p_l) - E(p_k) E(p_l) \\ &= \frac{\Gamma(\alpha_0)}{\prod_{j=1}^{K} \Gamma(\alpha_j)} \frac{\prod_{j=1, j \neq k, l}^{K} \Gamma(\alpha_j) \Gamma(1 + \alpha_k) \Gamma(1 + \alpha_l)}{\Gamma(2 + \alpha_0)} - \frac{\alpha_k}{\alpha_0} \frac{\alpha_l}{\alpha_0} \\ &= \frac{\alpha_k \alpha_l}{\alpha_0 (1 + \alpha_0)} - \frac{\alpha_k \alpha_l}{\alpha_0^2} \\ &= -\frac{\alpha_k \alpha_l}{\alpha_0^2 (1 + \alpha_0)}, \qquad \alpha_0 = \sum_{j=1}^{K} \alpha_j, \end{aligned} \tag{A.4}$$

where $k \neq l$ and $k, l = 1, \ldots, K$.

Appendix B

Pseudo-code of simulation

Algorithm 1 Robust clustering simulation under Dirichlet–Multinomial mixtures.

Require: K_{grid}, c_{grid}, R, n_{per}, D_{mode}, D_{fixed}, tier, truth_params
Ensure: Result table containing ARI, runtime, estimated noise proportion, and outlier capture rate for EM and EM+Noise

Define helper functions for:
    stable log-likelihood computation
    Dirichlet–Multinomial density evaluation
    sampling from Dirichlet and Dirichlet–Multinomial distributions
    evaluation metrics
Define the baseline Dirichlet–Multinomial mixture model:
    E-step for posterior responsibilities
    M-step for updating mixture weights and Dirichlet parameters
    full EM iteration loop
Define the robust Dirichlet–Multinomial mixture model with one fixed noise component:
    E-step over $K + 1$ components
    M-step updating mixture weights and real-cluster parameters
    final labels where $0 =$ noise and $1, \ldots, K =$ clusters
Define truth generators:
    Tier 1: single dominant feature per cluster
    Tier 2: adjacent dominant block
    Tier 3: orthogonal latent directions mapped through softmax
for each $K \in K_{\text{grid}}$ **do**
    Select D according to D_{mode} and the chosen tier
    if Tier 3 configuration is invalid **then**
        skip to the next K
    end if
    for each contamination level $c \in c_{\text{grid}}$ **do**
        for $r = 1$ to R **do**
            Generate clean data from the selected truth tier

Replace $\lfloor c \cdot n \rfloor$ observations with contaminated points
Fit the baseline EM model
Compute baseline ARI and runtime
Fit the robust EM+Noise model
Compute:
ARI on non-noise points
outlier capture rate
estimated noise proportion
runtime
Save results for both methods
end for
end for
end for
Combine all stored rows into one final result table
Save selected results and visualize:
ARI versus contamination
runtime by method
estimated noise proportion versus contamination