UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Contaminated models of reparameterised versions of the Dirichlet-multinomial distribution

by

Ockert van Heerden

Submitted in fulfilment of
the requirements for the degree

Master of Science (Advanced Data Analytics)

in the Faculty of Natural and Agricultural Sciences,
University of Pretoria,
Pretoria

November 2025

© University of Pretoria

Acknowledgements

I wish to express my gratitude to the following individuals and organizations who made this mini-dissertation possible: Dr. Seite L. Makgai, Prof. Andriëtte Bekker, Prof. Antonio Punzo, ma, en pa.

© University of Pretoria

Abstract

The Dirichlet-Multinomial (DM) distribution is often used for the modelling of multivariate count data, which has been applied in diverse areas such as microbiome studies, genetics, and ecological analysis. Despite its wide use, the distribution lacks easily interpretable parameters and the ability to account for outliers. In this study, we propose a novel reconstruction/perspective of the DM distribution: namely, reparameterisation of the DM distribution, which will be utilised to develop contaminated versions. Two reparameterisations are considered: the first in terms of the mode and a parameter referred to as the pseudo-variance and the second in terms of the mean and another pseudo-variance parameter. Such reparameterisations improve interpretability and allow the further construction of contaminated models that are robust to outliers. We consider properties such as the derived probability mass functions and moments for the proposed models. Simulation studies evaluate these models under varying scenarios, comparing estimation accuracy, bias, and computational performance. The relevance of the proposed models is illustrated via a microbiome data application. The developments from this study enhance the flexibility of the DM distribution and reinforce its usefulness for analyzing modern complex datasets in the biological and statistical sciences.

Keywords: contaminated models; Dirichlet-multinomial; microbiome data; outliers; overdispersion; reparameterisation; sensitivity analysis.

Supervisor: Dr. Seite L. Makgai
Co-supervisors: Prof. Andriëtte Bekker

© University of Pretoria

Contents

© University of Pretoria

List of Figures

© University of Pretoria

List of Tables

© University of Pretoria

List of Abbreviations

PMF	probability mass function
PDF	probability density function
AIC	Akaike information criterion
BIC	Bayesian information criterion
MSE	mean-squared-error
D	Dirichlet
M	multinomial
DM	Dirichlet-multinomial
MPV	mode pseudo-variance
PPV	proportional mean pseudo-variance
D-MPV	Dirichlet mode pseudo-variance
D-PPV	Dirichlet proportional mean pseudo-variance
DM-MPV	Dirichlet-multinomial mode pseudo-variance
DM-PPV	Dirichlet-multinomial proportional mean pseudo-variance
DM-CMPV	contaminated Dirichlet-multinomial mode pseudo-variance
DM-CPPV	contaminated Dirichlet-multinomial proportional mean pseudo-variance

© University of Pretoria

Chapter 1

Introduction

The multinomial distribution is valued in the analysis of multivariate count data. Such data can exhibit overdispersion, meaning that the variability observed in the counts exceeds that which is predicted by the classical multinomial distribution (Ng et al. (2011), p. 199). To address this limitation, the Dirichlet-multinomial (DM) distribution was introduced, extending the multinomial distribution by assuming that its probability vector follows a Dirichlet distribution (Mosimann (1962)), with applications ranging from microbiome studies (Subedi et al. (2020)) to genetics (Nowicka and Robinson (2016)), epidemiology (Bartolucci et al. (2021)), and text modelling (Bouguila (2008)).

Despite the added value of the DM distribution, it is not a distribution suitable for modelling all multivariate count datasets. For example, it has a clear negative correlation structure (Ng et al. (2011), Bouguila (2008)). This was addressed by Bouguila (2008) who developed a generalised DM distribution. Another area where the DM distribution is lacking is in the interpretability of its parameters and it also lacks a mechanism that can account for mild outliers.

Briefly, mild outliers can be defined as "observations sampled from some population different or even far from the assumed model" (Otto et al. (2025)), while gross outliers can be defined as "observations that cannot be modelled by a distribution as they are unpredictable" (Otto et al. (2025)). When outliers are mentioned in this study, it always refers to mild outliers.

The traditional parameterisation of the DM distribution ($\underline{\alpha}$) is not intuitive. Each α parameter simultaneously influences both the expected proportions and variance of every variable, making it difficult to directly interpret how changes in a parameter affect distributional behaviour. This lack of interpretability poses a barrier for applied researchers who require parameters with clear, real-world meaning. There is an alternative parameterisation of the DM distribution that makes use of allele probabilities (see Tvedebrink (2022) where this parameterisation was employed in a package that provides useful tools for implementing the DM distribution). This is helpful for interpretability; however, the issue of the DM distribution's inability to account for outliers is still present. This study utilises a technique referred to as contamination to address this problem.

Contamination represents a methodological approach that employs a mixture model to manage and account for outliers (Otto et al. (2025)). Several studies in the literature have demonstrated the application of this approach. For instance, Punzo et al. (2018) used mixtures of contaminated gamma distributions to model rent data, Punzo (2019) used contaminated inverse Gaussian distributions to model income data, Tomarchio et al. (2024) used contaminated Dirichlet distributions to analyse data of labour force participation rates, Tomarchio and Punzo (2020) used compound models of the log-normal distribution and gamma distribution respectively to model insurance losses. In short, the premise of contamination, in the context of the DM distribution for argument's sake, is that one assumes that the data was generated from two DM distributions (Alhaj-Dibo et al. (2008)). These two distributions have the same location parameter, however the second DM distribution has a variance that is greater than that of the first one (Alhaj-Dibo et al. (2008)). In the context of a mixture model, this would mean that when modelling this set of data, the first and second component have

© University of Pretoria

the same location parameters, but the second component has an inflated variance compared to the first one (Tomarchio et al. (2024)). This second component is the one that is used to account for outliers (Alhaj-Dibo et al. (2008)).

Contaminated DM distributions will be proposed in this study and the paper by Tomarchio et al. (2024) will be of great value, since in this paper a contaminated Dirichlet distribution was developed. In particular, the Dirichlet distribution was contaminated using the mode of the distribution, which has the trade off of restricting the parameters of the Dirichlet distribution (Tomarchio et al. (2024)). The reason for this is that the mode of the distribution may be a better indicator of location than the mean (Chacón (2020)). Despite the value of the mode, a mean parameterisation of the DM distribution will also be considered, since it will be more flexible than the mode parameterisation.

This mini-dissertation responds to both issues by exploring reparameterised and contaminated versions of the DM distribution. Reparametrisation seeks to provide more interpretable parameters, while contamination introduces robustness of parameter estimates (Tomarchio et al. (2024)) to outliers. Together, these advances aim to enhance the practical applicability of the DM distribution in data analysis.

1.1 Motivation

This study considers two major perspectives on the DM distribution:

1. Interpretability of parameters:
    - The parameters of the DM distribution do not influence the probability mass function (PMF) independently of another. This makes understanding the influence of a set of parameter values on the PMF not so intuitive.
    - This can be remedied by reparameterising the distribution such that the influence of a parameter on the distribution is more intuitive, but also so that one can change one parameter, say the mean, without influencing the variance, or at the very least have a minor influence on the variance.
2. Robustness to outliers:
    - The DM distribution does not account for outliers or atypical values. This means that outliers can influence parameter estimates when fitting the distribution to data, leading to biased results.
    - To combat this we propose a contaminated version of the DM distribution to account for these outliers. A contaminated version of the Dirichlet distribution was proposed by Tomarchio et al. (2024). This paper will be useful when developing a contaminated DM distribution considering the close relationship between the Dirichlet distribution and DM distribution.

Together, these considerations justify the present study: to develop and investigate reparameterised and contaminated versions of the DM distribution, assessing their performance through simulation and real-world applications.

1.2 Aims and objectives

The aim of this mini-dissertation is to develop and evaluate contaminated DM distributions that will give robust parameter estimates and account for outliers.

The specific objectives are:

1. Review the properties of the DM distribution, highlighting its strengths and limitations.

© University of Pretoria

2. Propose two reparametrisations of the DM distribution.
    - The DM distribution reparameterised in terms of the mode and pseudo-variance parameter. This distribution is referred to as the DM-MPV (DM mode pseudo-variance). The mode parameter in this parameterisation is in fact the mode of the Dirichlet distribution. It is the same parameterisation of the Dirichlet distribution as was proposed by Tomarchio et al. (2024) when a contaminated Dirichlet distribution was developed. The pseudo-variance parameter is a parameter that is closely related to the variance of the distribution and influences it.
    - The DM distribution reparameterised in terms of the mean and pseudo-variance parameter. This distribution is referred to as the DM-PPV (DM proportional mean pseudo-variance). A proportion of the mean is used as a parameter, hence the probability parameter. The pseudo-variance parameter is a parameter that is closely related to the variance of the distribution and influences it.
3. Conduct a simulation study comparing the well-known DM distribution, and the newly proposed DM-MPV distribution and DM-PPV distribution in terms of bias, mean squared error, and information criteria.
4. Extend the reparameterised DM distributions into contaminated models (DM-CMPV model and DM-CPPV model) capable of handling outliers.
5. Use a simulation study to show the value of the newly proposed contaminated models. This will be done using a sensitivity analysis which will work by generating data from a non-contaminated model, after which a few observations that are clearly outliers will be added to the data.
6. Apply the proposed models to real-world microbiome data, comparing their performance.
7. Discuss the findings, contributions, limitations, and directions for future research.

1.3 Outline of the study

This mini-dissertation is structured as follows: Chapter 2 provides the theoretical foundation of the Dirichlet-multinomial distribution, discussing its formulation, properties, computational challenges, and coding strategies. Chapter 3 focuses on reparameterization of the DM (see figure 1.1), which is a crucial step for developing a contaminated model as the parameters of the DM are not intuitive. In this chapter a simulation study will also be performed on the DM and its reparameterised counterparts. Chapter 4 presents the contaminated DM models (see figure 1.2) and uses a simulation study, in particular a sensitivity analysis, to evaluate the performance of the contaminated DM models. Chapter 5 will focus on applying the contaminated DM to real world data (The data used is a subset of the data used by Subedi et al. (2020)). Chapter 6 discusses the findings.

© University of Pretoria

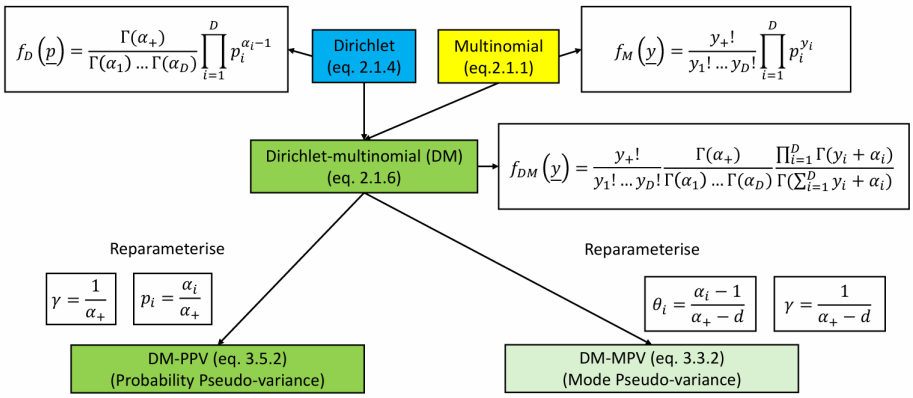


Figure 1.1: Flowchart of the reparameterised distributions

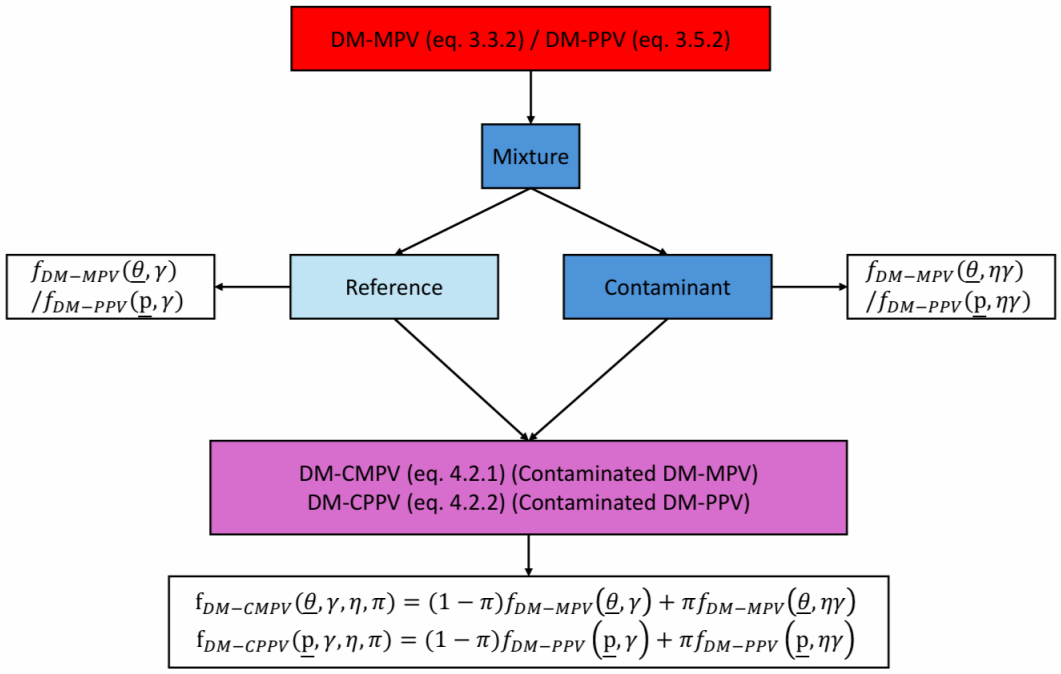


Figure 1.2: Flowchart of the contaminated distributions

© University of Pretoria

Chapter 2

The Dirichlet-multinomial distribution

The Dirichlet-multinomial distribution was developed by Mosimann (1962), and is a natural extension of the multinomial distribution which models multivariate count data. The DM distribution has the ability to model count data with a high variance (Subedi et al. (2020)). The DM distribution is constructed by assuming that the probability parameters of the multinomial distribution follow a Dirichlet distribution (Holmes et al. (2012), Subedi et al. (2020)).

In this chapter, we provide the general framework of the DM distribution. We revisit in section 2.1 the DM distribution, the effects of the parameters on the PMF function are displayed in section 2.2, and in section 2.3 computational aspects of coding the model will be outlined.

2.1 Dirichlet-multinomial distribution

The multinomial distribution is used to model count data. It is the multivariate extension of the binomial distribution (see definition 1 for the PMF), and has parameters that are probabilities. The multinomial is a useful distribution, but cannot adequately model data where the variance is high (Ng et al. (2011), p. 199).

The Dirichlet distribution is used to model compositional data, i.e. data where the sum of the variables for a given observation equals one (see definition (2) for the probability density function (PDF)). For a model with D variables, it has D parameters. Trade offs of the Dirichlet distribution include that it cannot model concave data (Aitchison (1982)) and that all variables are negatively correlated (Ng et al. (2011), p. 39-40).

A relation can be drawn between these two distributions when assuming that the probability parameters of the multinomial distribution follow a Dirichlet distribution, which yields the DM distribution (Mosimann (1962), Holmes et al. (2012), Subedi et al. (2020)). The reason for creating this distribution is mentioned by Subedi et al. (2020), "to account for the over-dispersion in the data".

Definition 1 (Multinomial distribution (Bain (1992), p. 138, Subedi et al. (2020))). *Let $\underline{y} = \{Y_1, ..., Y_D\}$ denote a random vector, where each Y_i is a natural number, then the PMF of the multinomial distribution is given as,*

$$f_M(\underline{y}; \underline{p}) = \frac{y_+!}{y_1!y_2!...y_D!} \prod_{d=1}^{D} p_d^{y_d}, \tag{2.1.1}$$

where $y_+ = \sum_{i=1}^{D} y_i$. The parameters in the multinomial distribution are the probabilities $\underline{p} = \{p_1, ..., p_D\}$. The constraints on these parameters are, $0 < p_d < 1$, where $d \in \{1, ..., D\}$, and crucially, $\sum_{d=1}^{D} p_d = 1$. We denote the random vector $\underline{y} \sim multinomial(y_+, \underline{p})$.

© University of Pretoria

Proposition 1 (Moments of the multinomial distribution (Ng et al. (2011), p. 203)). *Given a multinomial distribution of dimension D, the expected count of the d^{th} variable, variance of the d^{th} variable, and the covariance between the i^{th} and j^{th} variables are given in eq. (2.1.2).*

$$\begin{aligned} E(P_d) &= y_+ p_d, \\ Var(P_d) &= y_+ p_d(1-p_d), \\ Cov(P_i, P_j) &= -y_+ p_i p_j, \end{aligned} \tag{2.1.2}$$

where $d, i, j \in \{1, ..., D\}$; $i \neq j$, and $y_+ = \sum_{i=1}^{D} y_i$.

Definition 2 (Dirichlet distribution (Ng et al. (2011), p. 38)). *Let $\underline{P} = (P_1, ..., P_D)$ denote a random vector on the Ω_D unit simplex where,*

$$\Omega_{D-1} = \{(p_1, ..., p_D), 0 < p_i < 1, \sum_{i=1}^{D} p_i = 1\}, \tag{2.1.3}$$

*then the PDF of the **Dirichlet** distribution is given as,*

$$f_D(\underline{p}; \underline{\alpha}) = \frac{\Gamma(\alpha_+)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} p_d^{\alpha_d - 1}, \tag{2.1.4}$$

where $\sum_{i=1}^{D} p_i = 1$, $\alpha_d > 0$ where $d \in \{1, ..., D\}$, and $\alpha_+ = \sum_{i=1}^{D} \alpha_i$. We denote the random vector $\underline{P} \sim Dirichlet(\underline{\alpha})$.

Proposition 2 (Moments of the Dirichlet distribution (Ng et al. (2011), p. 39)). *Given a Dirichlet distribution of dimension D, the expected proportion of the d^{th} variable, variance of the d^{th} variable, and the covariance between the i^{th} and j^{th} variables are given in eq. (2.1.5).*

$$\begin{aligned} E(P_d) &= \frac{\alpha_d}{\alpha_+}, \\ Var(P_d) &= \frac{\alpha_d}{\alpha_+}\left(1 - \frac{\alpha_d}{\alpha_+}\right)\frac{1}{1+\alpha_+}, \\ Cov(P_i, P_j) &= -\frac{\alpha_i}{\alpha_+}\frac{\alpha_j}{\alpha_+}\frac{1}{1+\alpha_+}, \end{aligned} \tag{2.1.5}$$

where $d, i, j \in \{1, ..., D\}$; $i \neq j$, $\alpha_+ = \sum_{i=1}^{D} \alpha_i$.

The DM distribution is developed by assuming that the probability parameter in the multinomial distribution follows a Dirichlet distribution (Holmes et al. (2012), Subedi et al. (2020)). Considering proposition 2, it is clear that when $\alpha_+ \to \infty$, that the variance of the Dirichlet tends towards zero. This is important when considering the DM distribution. If the Dirichlet has variance zero, it becomes degenerate. In the context of the DM distribution, if $\alpha_+ \to \infty$, then the DM converges to the multinomial distribution.

Definition 3 (Dirichlet-multinomial distribution (Ng et al. (2011) p. 200, Holmes et al. (2012), Subedi et al. (2020))). *Suppose that a D-dimensional vector $\underline{Y} = (Y_1, ..., Y_D)$, where each Y_i is a natural number, is said to have a multinomial distribution where the parameters of the multinomial distribution*

© University of Pretoria

are said to follow a Dirichlet distribution, i.e. $\underline{Y}|\underline{p} \sim multinomial(y_+, \underline{p})$, *where* $\underline{P} \sim Dirichlet(\underline{\alpha})$. *The PMF of the Dirichlet-multinomial distribution is then constructed as follows,*

$$\begin{aligned} f_{DM}(\underline{y}, \underline{\alpha}) &= \int_{\underline{p}} f_M(\underline{y}, \underline{p}) f_D(\underline{p}, \underline{\alpha}) d\underline{p} \\ &= \int_{\underline{p}} \left[\frac{y_+!}{y_1! y_2! ... y_D!} \prod_{d=1}^{D} p_d^{y_d} \right] \left[\frac{\Gamma(\alpha_+)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} p_d^{\alpha_d - 1} \right] d\underline{p} \\ &= \int_{\underline{p}} \frac{y_+!}{y_1! y_2! ... y_D!} \frac{\Gamma(\alpha_+)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \prod_{d=1}^{D} p_d^{y_d + \alpha_d - 1} d\underline{p} \\ &= \frac{y_+!}{y_1! y_2! ... y_D!} \frac{\Gamma(\alpha_+)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \int_{\underline{p}} \prod_{d=1}^{D} p_d^{y_d + \alpha_d - 1} d\underline{p} \\ &= \frac{y_+!}{y_1! y_2! ... y_D!} \frac{\Gamma(\alpha_+)}{\prod_{d=1}^{D} \Gamma(\alpha_d)} \frac{\prod_{d=1}^{D} \Gamma(y_d + \alpha_d)}{\Gamma(y_+ + \alpha_+)}, \end{aligned} \tag{2.1.6}$$

where $y_+ = \sum_{i=1}^{D} y_i$ *and* $\alpha_+ = \sum_{i=1}^{D} \alpha_i$. *The integral in line 4 of the expression is identified as a Dirichlet distribution with parameters* $Y_1 + \alpha_1, ..., Y_D + \alpha_D$, *where each* $\alpha_i > 0$.

Remark 1 (Beta-binomial distribution (Ishii and Hayakawa (1960), Ng et al. (2011), p. 201)). *Considering eq. (2.1.6) in definition 3, it is clear that if* $D = 2$, *the Dirichlet-multinomial reduces to the beta-binomial. For* $D = 2$, *the PMF of the Dirichlet-multinomial distribution becomes,*

$$f_{DM}(\underline{y}; \underline{\alpha}) = \frac{(y_1 + y_2)!}{y_1! y_2!} \frac{\Gamma(\alpha_1 + \alpha_2)}{\Gamma(\alpha_1)\Gamma(\alpha_2)} \frac{\Gamma(y_1 + \alpha_1)\Gamma(y_2 + \alpha_2)}{\Gamma(y_1 + y_2 + \alpha_1 + \alpha_2)}, \tag{2.1.7}$$

which is the PMF of the beta-binomial with parameters α_1 *and* α_2.

Remark 2 (Mixtures of the Dirichlet-multinomial distribution (Anderlucci and Viroli (2020))). *A mixture model of the Dirichlet-multinomial distribution with* k *components will have a PMF as follows,*

$$f(\underline{y}; \underline{\alpha}_1, ..., \underline{\alpha}_k, \underline{\pi}) = \sum_{i=1}^{k} \pi_i f_{DM}(\underline{y}|\underline{\alpha}_i), \tag{2.1.8}$$

where $\sum_{i=1}^{k} \pi_i = 1$ *and* f_{DM} *is the PMF of the Dirichlet-multinomial distribution.*

Remark 2, which refers to a mixture model of DM distributions, will be useful in chapter 4 when contamination is introduced.

The moments of the multinomial distribution are given in proposition 1 and the moments of the Dirichlet distribution are given in proposition 2. The moments of these two distributions are used to derive the moments of the DM distribution (Ng et al. (2011), p. 204). Proposition 3 derives the expected count of the DM distribution (Ng et al. (2011), p. 204), proposition 4 derives the variance of the DM distribution (Ng et al. (2011), p.204), and proposition 5 derives the covariance of the DM distribution (Ng et al. (2011), p. 204).

Proposition 3 (Expected count of the Dirichlet-multinomial distribution (Ng et al. (2011), p. 204)). *If the variables* $Y_1, ..., Y_D$ *follow a Dirichlet-multinomial distribution, then the mean of the variables is obtained as follows.*

$$\begin{aligned} E[Y_d] &= E\{E[Y_d|P_d]\} \\ &= E[y_+ P_d] \\ &= y_+ \frac{\alpha_i}{\alpha_+}, \end{aligned} \tag{2.1.9}$$

where $d \in \{1, ..., D\}$, $y_+ = \sum_{i=1}^{D} y_i$, *and* $\alpha_+ = \sum_{i=1}^{D} \alpha_i$.

© University of Pretoria

Proposition 4 (Variance of the Dirichlet-multinomial (Ng et al. (2011), p. 204)). *If the variables $Y_1, ..., Y_D$ follow a Dirchlet-multinomial distribution, then the variance of the d^{th} variable is obtained as follows.*

$$\begin{aligned}
Var[Y_d] &= E\{Var[Y_d|P_d]\} + Var\{E[Y_d|P_d]\} \\
&= E[y_+P_d(1-P_d)] + Var[y_+P_d] \\
&= y_+E[P_d - P_d^2] + y_+^2\frac{\alpha_d(\alpha_+ - \alpha_d)}{\alpha_+^2(1+\alpha_+)} \\
&= y_+E[P_d] - y_+E[P_d^2] + y_+^2\frac{\alpha_d(\alpha_+ - \alpha_d)}{\alpha_+^2(1+\alpha_+)} \\
&= y_+\frac{\alpha_d}{\alpha_+} - y_+\frac{\alpha_d(1+\alpha_d)}{\alpha_+(1+\alpha_+)} + \frac{\alpha_d(\alpha_+ - \alpha_d)}{\alpha_+^2(1+\alpha_+)} \\
&= \frac{y_+\alpha_d\alpha_+^2 - y_+\alpha_d^2\alpha_+ + y_+^2\alpha_d\alpha_+ - y_+^2\alpha_d^2}{\alpha_+^2(1+\alpha_+)} \\
&= y_+\alpha_d\frac{(y_+ + \alpha_+)(\alpha_+ - \alpha_d)}{\alpha_+^2(1+\alpha_+)} \\
&= y_+\frac{\alpha_d}{\alpha_+}\left(1 - \frac{\alpha_d}{\alpha_+}\right)\frac{y_+ + \alpha_+}{1+\alpha_+},
\end{aligned} \tag{2.1.10}$$

where $d \in \{1, ..., D\}$, $y_+ = \sum_{i=1}^D y_i$, and $\alpha_+ = \sum_{i=1}^D \alpha_i$.

Proposition 5 (Covariance of the Dirichlet-multinomial (Ng et al. (2011), p. 204)). *If the variables $Y_1, ..., Y_D$ follow a Dirichlet-multinomial distribution, then the covariance of the i^{th} and j^{th} variables is obtained as follows.*

$$\begin{aligned}
Cov[y_i, y_j] &= E\{E(y_iy_j|p_ip_j)\} - E[y_i]E[y_j] \\
&= E\{Cov(y_i, y_j|p_i, p_j) + E(y_i|p_i)E(y_j|p_j)\} - E[y_i]E[y_j] \\
&= E(-y_+p_ip_j + y_+^2p_ip_j) - E[y_i]E[y_j] \\
&= y_+(y_+ - 1)E[p_ip_j] - E[y_i]E[y_j] \\
&= y_+(y_+ - 1)\{Cov(p_i, p_j) + E[p_i]E[p_j]\} - E[y_i]E[y_j] \\
&= y_+(y_+ - 1)\left[-\frac{\alpha_i}{\alpha_+}\frac{\alpha_j}{\alpha_+}\frac{1}{1+\alpha_+} + \frac{\alpha_i}{\alpha_+}\frac{\alpha_j}{\alpha_+}\right] - y_+^2\frac{\alpha_i}{\alpha_+}\frac{\alpha_j}{\alpha_+} \\
&= -y_+\frac{y_+ + \alpha_+}{1+\alpha_+}\frac{\alpha_i}{\alpha_+}\frac{\alpha_j}{\alpha_+},
\end{aligned} \tag{2.1.11}$$

where $i \in \{1, ..., D\}$; $j \in \{1, ..., D\}$, $y_+ = \sum_{i=1}^D y_i$, and $\alpha_+ = \sum_{i=1}^D \alpha_i$.

The parameters $\alpha_1, ..., \alpha_D$ are flexible, however their effect on the distribution is not intuitively understood. The mean and variance of the DM distribution are of interest, since they give a greater understanding of the effect the parameter values has on the distribution. Considering the mean of the DM distribution in eq. (2.1.9), its clear that the value of the mean of say, Y_d, is related to the value of the d^{th} parameter; however, it is also dependent on all the other parameters. Considering the variable Y_d and eq. (2.1.10), it is clear that the d^{th} parameter influences the variance, but so do all the other parameters. This creates the need for a DM distribution with a different set of parameters where the influence of the parameters on the mean and variance of one variable is more decisive.

2.2 Influence of the parameters on the distribution

The variance of the DM distribution seen in eq. (2.1.10), is written as it is in the final line to point out the similarity between it and the variance of the multinomial. This was pointed out in the original

© University of Pretoria

paper by Mosimann (1962) on the DM distribution. This relation is used by Mosimann (1962) to estimate the parameters through methods of moments estimation (Ng et al. (2011), p. 210-212). This relation is given in the eq. (2.2.1).

$$\begin{aligned} Var_{DM}[y_d] &= y_+ \frac{\alpha_d}{\alpha_+}\left(1-\frac{\alpha_d}{\alpha_+}\right)\frac{y_+ + \alpha_+}{1+\alpha_+} \\ &= y_+ p_d(1-p_d)\frac{y_+ + \alpha_+}{1+\alpha_+} \\ &= Var_M[y_d] \times \frac{y_+ + \alpha_+}{1+\alpha_+} \\ &= Var_M[y_d] \times c, \end{aligned} \tag{2.2.1}$$

where $d \in \{1, ..., D\}$.

Considering eq. (2.2.1), the quantity c serves as an inflation of the variance of the multinomial distribution. This quantity, referred to as c (Ng et al. (2011), p. 210), is bounded as follows $1 < c < y_+$, and as a result allows the variance of the DM to span from $Var_M[y_d]$ to $y_+ Var_M[y_d]$, thereby addressing overdispersion. This is where the DM distribution outshines the multinomial distribution, since the variance of the distribution can be changed without affecting the mean. This is not the case with the multinomial distribution where the variance cannot be modified independently of the mean.

The influence of α_+ on the quantity c is important to consider, since the quantity c serves as the inflation value of the variance. The quantity c increases as α_+ decreases and vice versa. This means that the variance of a given variable increases as α_+ decreases and vice versa. The general influence of the parameters on the distribution are displayed in the figures 2.1a, 2.1b, 2.1c, and 2.1d.

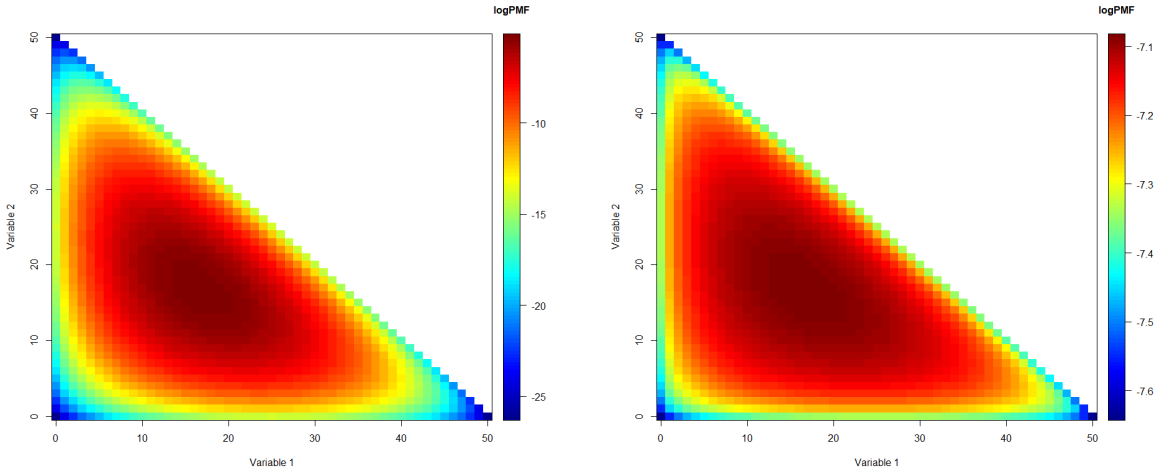
In figure 2.1a, it is clear that with parameter values equal to each other and values close to 0, the PMF has higher probabilities around the bounds of the distribution. Figure 2.1b displays the effect of the proportions of parameter values on the mean of the distribution. Figures 2.1d and 2.1c show the effect of an increased sum of the parameter values on the variance. It is important to note that an increased variance does not imply a flattened curve as it would with other distributions, say the normal. This is because the mode does not always exist as observations can crowd around the edges.

It should be noted that for the Dirichlet distribution, when all parameters are equal to one, the distribution becomes uniform (Huelsenbeck and Andolfatto (2007)). This is also the case for the Dirichlet-multinomial, where for all parameters equal to one ($\underline{\alpha} = \underline{1}$), the PMF reduces to a constant. Using eq. (2.2.2) given by Bain (1992) p. 111, the PMF reduces as seen in eq. (2.2.3).

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1). \tag{2.2.2}$$

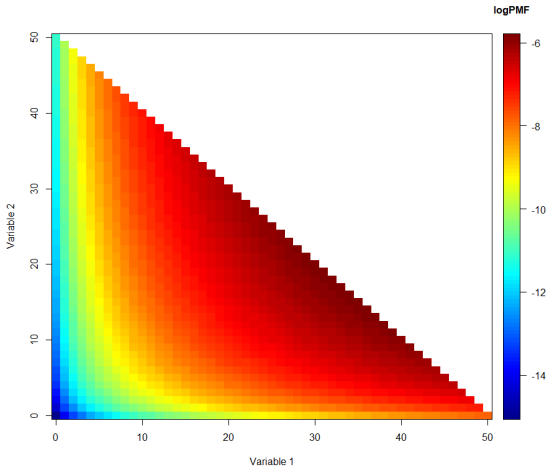
$$\begin{aligned} f_{DM}(\underline{y}, \underline{\alpha}) &= \frac{y_+!}{y_1!y_2!...y_D!}\frac{\Gamma(D)}{\prod_{d=1}^{D}\Gamma(1)}\frac{\prod_{d=1}^{D}\Gamma(y_d+1)}{\Gamma(y_+ + D)} \\ &= \frac{y_+!}{y_1!y_2!...y_D!}\Gamma(D)\frac{y_1!y_2!...y_D!}{\Gamma(y_+ + D)} \\ &= y_+!\frac{\Gamma(D)}{\Gamma(y_+ + D)} \\ &= y_+!\frac{(D-1)!}{(y_+ + D - 1)!} \\ &= \frac{(D-1)!}{(y_+ + D - 1) \times (y_+ + D - 2) \times ... \times (y_+ + 1)} \end{aligned} \tag{2.2.3}$$

© University of Pretoria

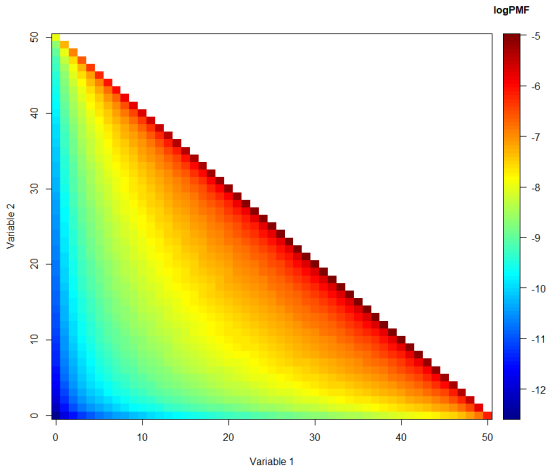


(a) $\underline{\alpha} = (5, 5, 5)$

(b) $\underline{\alpha} = (1.1, 1.1, 1.1)$



(c) $\underline{\alpha} = (3, 2, 1)$



(d) $\underline{\alpha} = (2, 1.5, 0.5)$

Figure 2.1: Plots of the DM distribution (eq. (2.1.6)) for selected values of $\underline{\alpha}$ where $y_+ = 50$.

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

2.3 Computational aspects of the DM distribution

To determine the parameter estimates of the DM distribution, one can use the method of moments estimates. There are a few different method of moments estimates and they are described by Ng et al. (2011), p.208-212; however, the method of maximum likelihood will be used over the method of moments. From eq. (2.1.6), the log-likelihood function of the DM distribution is given in eq. (2.3.1). Thus,

$$\begin{aligned} l(\underline{\alpha}) &= \sum_{i=1}^{n} log f(\underline{y}_i, \underline{\alpha}) \\ &= \sum_{i=1}^{n} log\left(\frac{y_{i+}!}{y_{i1}!y_{i2}!...y_{iD}!}\right) + \sum_{i=1}^{n} log\left(\frac{\Gamma(\alpha_+)}{\prod_{d=1}^{D}\Gamma(\alpha_d)}\right) + \sum_{i=1}^{n} log\left(\frac{\prod_{d=1}^{D}\Gamma(y_{id}+\alpha_d)}{\Gamma(y_{i+}+\alpha_+)}\right) \\ &= \sum_{i=1}^{n} log(y_{i+}!) - \sum_{i=1}^{n} log(y_{i1}!) - ... - \sum_{i=1}^{n} log(y_{iD}!) \\ &+ n \times log\Gamma(\alpha_+) - n \times log\Gamma(\alpha_1) - ... - n \times log\Gamma(\alpha_D) \\ &+ \sum_{i=1}^{n} log\Gamma(y_{i1}+\alpha_1) + ... + \sum_{i=1}^{n} log\Gamma(y_{iD}+\alpha_D) - \sum_{i=1}^{n} log\Gamma(y_{i+}+\alpha_+). \end{aligned} \tag{2.3.1}$$

Noting the log-likelihood function in eq. (2.3.1), it should be considered how this log-likelihood function will be calculated when attempting to find the maximum likelihood estimates of the parameters. This is important to consider since the modelling of data using the distribution will be done the programming language R (R Core Team (2024)), and this programming language has its limitations. For example, in the code provided by Subedi et al. (2020), where a finite mixture of DM distributions was fitted, there is a comment that the gamma function will not work for values over 170. This is because for too large an output value, R will give the output of Inf. The same is true for large values of the factorial in R. This means that a workaround is needed, which is explained below.

It is important to remember that for an integer Z,

$$\begin{aligned} log(Z!) &= \sum_{i=1}^{Z} log(i), \\ i.e. \quad Z! &= exp\left(\sum_{i=1}^{Z} log(i)\right). \end{aligned} \tag{2.3.2}$$

This relation will be used to calculate any factorial values that are too large for R.

For a value α, the value of the gamma function, according to (Bain (1992), p. 111), can be written as in eq. (2.3.3).

$$\Gamma(\alpha) = (\alpha - 1)\Gamma(\alpha - 1). \tag{2.3.3}$$

Eq. (2.3.3) and eq. (2.3.4) can be used to rewrite the gamma function as in eq. (2.3.5).

$$\begin{aligned} \alpha_f &= floor(\alpha), \\ \alpha_\delta &= \alpha - \alpha_f, \end{aligned} \tag{2.3.4}$$

where α_f refers to the value of α being rounded down. Thus,

$$log(\Gamma(\alpha)) = log(\Gamma(\alpha_\delta)) + \sum_{i=1}^{\alpha_f} log(i + \alpha_\delta). \tag{2.3.5}$$

© University of Pretoria

The reason for rewriting the gamma function as in eq. (2.3.5) and the factorial as in eq. 2.3.2, is that this will allow for the computation of the log of the gamma function and also the factorial of the observations at very large values, rather than simply giving an output of Inf in R. This will be very useful, however for larger inputs into either function, the output value will take longer to compute. This is unfortunately something that must be accepted when using the factorial in the context of the DM distribution, since the input values will be the observations themselves. With respect to the gamma function in the context of the DM distribution, these inputs will be the parameter values. This opens up the possibility to put a cap on these values (see eq. (2.2.1).

The quantity c in eq. (2.2.1), referred to by Mosimann (1962) and Ng et al. (2011), p. 210., is bounded below by 1 and bounded above by y_+. The larger the value of α_+, the closer this quantity will move to the value of 1. There are, however, diminishing returns when the value of α_+ is increased. Therefore, depending on the data and the values of y_+ present in the data, it maybe sensible to place a cap on the value of α_+ such that one can save substantial time on the computation, while still also getting reliable results.

Regarding the parameters of the DM distribution, there is another challenge that needs to be addressed other than whether the log-likelihood function can be calculated. This is that the parameters are bounded - the parameters must be positive. When not using a built in function, this creates a problem, as one must insure that the parameters stay positive in the estimation process. This can be done with if-statements, however it would be prudent to use a different approach. In a DM distribution regression model, Subedi et al. (2020) and Bartolucci et al. (2021) modelled the parameters using an exponential, which ensures that the parameters will be positive. The parameters are then modelled as given in eq. (2.3.6), while their back transformation is given in eq. (2.3.7).

$$\alpha_i = exp(\beta_i). \tag{2.3.6}$$

$$\beta_i = log(\alpha_i). \tag{2.3.7}$$

Using this approach, the estimation of the parameters becomes simpler, as it means that one does not have to worry about the parameters exceeding their bounds. This approach will also be used with the reparameterised distributions and the contaminated models that appear later in this document.

2.4 Conclusion

The DM distribution is a useful model for count data with a high variance, but it has limitations. The interpretation of the parameters are not intuitive, which leads to the necessity of an adaptation of the DM distribution that will address this problem. This will be done in chapter 3 where the parameters of the DM distribution will be altered.

© University of Pretoria

Chapter 3

Two reparameterised versions of the Dirichlet-multinomial distribution

This chapter focuses on two reparameterised versions of the DM distribution to overcome the interpretability challenges of the conventional parameterisation in terms of the parameters $\underline{\alpha} = (\alpha_1, ..., \alpha_D)$ of the DM distribution. These two reparameterised versions are,

- The **Mode-Pseudo-Variance (MPV)** parameterisation, based on the mode of the Dirichlet distribution and a dispersion parameter γ. This is the same parameterisation utilised for the Dirichlet by Tomarchio et al. (2024).
- The **Proportional-mean-Pseudo-Variance (PPV)** parameterisation, based on the mean of the Dirichlet distribution and a dispersion parameter γ.

This chapter proceeds by motivating the need for a reparameterised version of the DM distribution in section 3.1, defining the proposed MPV Dirichlet form in section 3.2, defining the proposed MPV parmeterisation of the DM distribution in section 3.3, defining the proposed PPV Dirichlet form in section 3.4, defining the proposed PPV Dirichlet-multinomial form in section 3.5, and then performing a simulation study in section 3.6.

3.1 Motivation for reparameterisation

Before reparameterising the DM distribution, the influence of the parameters must be understood. There are three key points with regard to this.

- The mode of the Dirichlet distribution only exists if all the parameters, $\underline{\alpha} = (\alpha_1, ..., \alpha_D)$, are greater than 1 (Tomarchio et al. (2024)).
- The variance of a variable in the DM distribution (eq. 2.1.10) increases as the sum of the parameters, α_+, decreases and vice versa, assuming the ratio of α_i to α_+ remains constant.
- The mean of a given variable that is DM distributed is the proportion of α_i to α_+ (eq. (2.1.9)).

Given the influence of the parameters on the shape of the distribution as outlined above, it is clear that one cannot compare the influence of two different sets of parameters for the DM distribution immediately. A few calculations may need to be done first. It would, therefore, be prudent to consider an alternative parameterisation of the DM distribution where the location and the variance of the DM distribution can be gleamed immediately when confronted with a set of parameters.

© University of Pretoria

3.2 Dirichlet distribution with MPV parameterisation

The MPV parameterisation of the Dirichlet distribution was developed by Tomarchio et al. (2024). This was done with the goal in mind of developing a contaminated Dirichlet distribution that can address outliers present in real world data. The reason for the use of the mode is simple, as Tomarchio et al. (2024) states that it "may be a may be a more informative and meaningful measure of central tendency for data that originates from distributions that are skewed and have heavy tails", which is also a sentiment expressed by Chacón (2020). This modal representation of the Dirichlet distribution is defined in definition 4.

Definition 4 (Dirichlet distribution with MPV parameterisation (Tomarchio et al. (2024))). *Let* $\underline{P} = (P_1, ..., P_D)$ *denote a random vector on the* Ω_D *unit simplex where,*

$$\Omega_D = \{(P_1, ..., P_D), 0 < P_i < 1, \sum_{i=1}^{D} P_i = 1\}, \tag{3.2.1}$$

then the PDF of the ***mode-pseudo-variance reparameterised Dirichlet*** *distribution (D-MPV) is given by*

$$\begin{aligned} f_{D-MPV}(\underline{p}; \underline{\theta}, \gamma) &= \frac{\Gamma(D + \frac{1}{\gamma})}{\prod_{j=1}^{D} \Gamma(1 + \frac{\theta_j}{\gamma})} \prod_{j=1}^{D} p_j^{\frac{\theta_j}{\gamma}} \\ &= \frac{\Gamma(D + \frac{1}{\gamma})}{\Gamma\left(1 + \frac{1 - \sum_{i=1}^{D-1} \theta_i}{\gamma}\right) \prod_{j=1}^{D-1} \Gamma(1 + \frac{\theta_j}{\gamma})} p_j^{\frac{1 - \sum_{i=1}^{D-1} \theta_i}{\gamma}} \prod_{j=1}^{D-1} p_j^{\frac{\theta_j}{\gamma}}, \end{aligned} \tag{3.2.2}$$

where each $\theta_i > 0$, $\sum_{i=1}^{D} \theta_i = 1$, *denotes the modal proportions,* $\gamma > 0$ *is the pseudo-variance parameter. Note that though there are* D *modal proportion parameters, in practice there will be only* $D - 1$*, since they are all positive and sum to one, implying that one of them need not be estimated, since its value can be calculated using all the other modal proportion parameters. We denote the random vector* $\underline{P} = (P_1, ..., P_D) \sim D - MPV(\underline{\theta}, \gamma)$.

Proposition 6 (Relationship between the classical and MPV parameterisation of the Dirichlet distribution (Tomarchio et al. (2024))). *If* $\underline{P} \sim D-MPV(\underline{\theta}, \gamma)$*, then the equivalent standard Dirichlet parameters are,*

$$\begin{aligned} \alpha_j &= 1 + \frac{\theta_j}{\gamma}, \quad j \in \{1, ..., D\}, \\ \alpha_+ &= \frac{1 + \gamma D}{\gamma}. \end{aligned} \tag{3.2.3}$$

Proof. When setting the parameters θ_j *and* γ *equal to the eq. (3.2.4) below,*

$$\begin{aligned} \theta_j &= \frac{\alpha_j - 1}{\alpha_+ - D}, \quad j \in \{1, ..., D-1\}, \\ \gamma &= \frac{1}{\alpha_+ - D}, \end{aligned} \tag{3.2.4}$$

and substituting them into the PDF of the Dirichlet distribution (see eq. (3.2.2)), one will obtain eq. (2.1.4). Note that $\alpha_+ = \sum_{i=1}^{D} \alpha_i$.

The parameter γ defined in eq. (3.2.4) is not referred to by Tomarchio et al. (2024) as the "pseudo-variance" parameter; however, this is certainly a suitable name for it given that Tomarchio et al. (2024) states that the choice of γ (see eq. (3.2.4)), is "so that γ is approximately related to the variability of the $d - 1$ variables".

© University of Pretoria

With this MPV parameterisation of the Dirichlet, it should be noted that the Dirichlet does lose some flexibility as the parameters are restricted to be greater than one ($\underline{\alpha} > \underline{1}$). This is because as Tomarchio et al. (2024) state, the mode exists only if all parameters are greater than 1. Tomarchio et al. (2024) states that this makes the D-MPV distribution a subclass of the Dirichlet distribution, which is a considerable trade off.

3.3 Dirichlet-multinomial distribution with the MPV parameterisation

The DM distribution is reparameterised using the same relations as the Dirichlet under the MPV parameterisation (see section 3.2). The new parameters of the distribution are now the mode of the Dirichlet distribution, which is the distribution of the probability parameters of the multinomial distribution when considering the DM distribution (Holmes et al. (2012), Subedi et al. (2020)), and the pseudo-variance parameter. Crucially, the pseudo-variance parameter retains its most important characteristic, which is that it will increase the variance when it is increased and vice versa.

Definition 5 (Dirichlet-multinomial distribution with MPV parameterisation). *Let,*

$$\begin{aligned} \underline{Y}|\underline{p} &\sim multinomial(y_+, \underline{p}), \\ \underline{P} &\sim D - MPV(\underline{\theta}, \gamma), \end{aligned} \tag{3.3.1}$$

*then the random vector $\underline{Y} = (Y_1, ..., Y_D)$ is said to follow a **Mode-Pseudo-Variance Dirichlet-Multinomial** (DM-MPV) distribution if its PMF is given in eq. (3.3.2).*

$$\begin{aligned} f_{DM-MPV}(\underline{y}; \underline{\theta}, \gamma) &= \frac{y_+}{y_1!y_2!...y_D!} \frac{\Gamma(D + \frac{1}{\gamma})}{\prod_{d=1}^{D} \Gamma(1 + \frac{\theta_d}{\gamma})} \frac{\prod_{d=1}^{D} \Gamma(y_d + 1 + \frac{\theta_d}{\gamma})}{\Gamma(y_+ + D + \frac{1}{\gamma})} \\ &= \frac{y_+}{y_1!y_2!...y_D!} \frac{\Gamma(D + \frac{1}{\gamma})}{\Gamma\left(1 + \frac{1 - \sum_{i=1}^{D-1} \theta_i}{\gamma}\right) \prod_{d=1}^{D-1} \Gamma(1 + \frac{\theta_d}{\gamma})} \\ &\times \frac{\Gamma\left(y_d + 1 + \frac{1 - \sum_{i=1}^{D-1} \theta_i}{\gamma}\right) \prod_{d=1}^{D-1} \Gamma(y_d + 1 + \frac{\theta_d}{\gamma})}{\Gamma(y_+ + D + \frac{1}{\gamma})}, \end{aligned} \tag{3.3.2}$$

where each $\theta_i > 0$, $\sum_{i=1}^{D} \theta_i = 1$, is the modal proportions and $\gamma > 0$ is the pseudo-variance. Note that $y_+ = \sum_{i=1}^{D} y_i$. Note that though there are D modal proportion parameters, in practice there will be only $D - 1$, since they are all positive and sum to one, implying that one of them need not be estimated, since its value can be calculated using all the other modal proportion parameters. We denote the random vector $\underline{Y} = (Y_1, ..., Y_D) \sim DM - MPV(\underline{\theta}, \gamma)$.

Proposition 7 (Moments of the DM-MPV distribution). *From proposition 3, proposition 4, and proposition 5 it follows that the expected count for the d^{th} variable of the DM-MPV distribution, variance for the d^{th} variable of the DM-MPV distribution, and covariance between variables i and j of the DM-MPV distribution are given in eq. (3.6.2).*

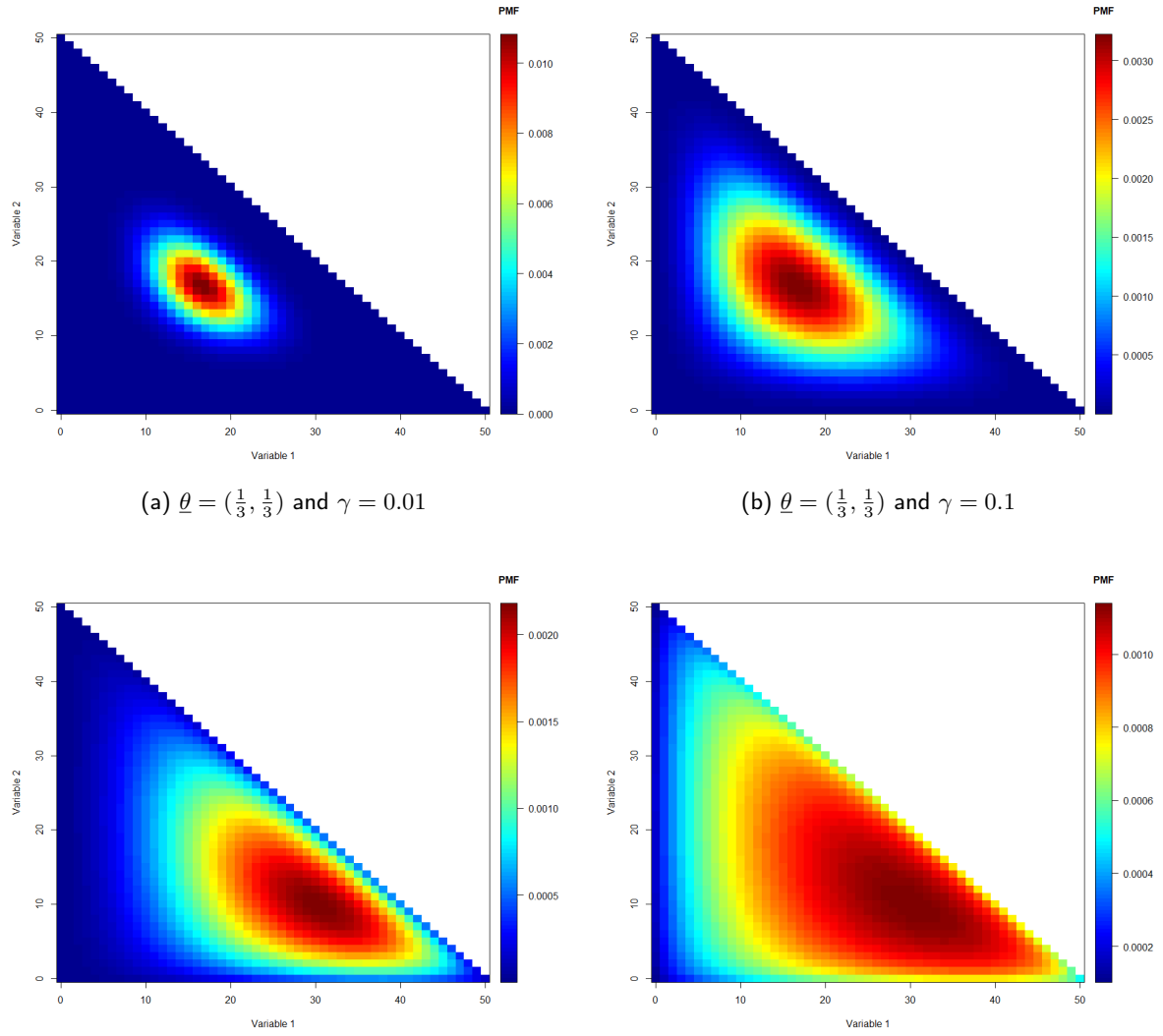
$$\begin{aligned} E[Y_d] &= y_+ \frac{1 + \frac{\theta_d}{\gamma}}{D + \frac{1}{\gamma}}, && d \in \{1, ..., D\}, \\ Var[Y_d] &= y_+ \frac{1 + \frac{\theta_d}{\gamma}}{D + \frac{1}{\gamma}} \left(1 - \frac{1 + \frac{\theta_d}{\gamma}}{D + \frac{1}{\gamma}}\right) \frac{y_+ + D + \frac{1}{\gamma}}{1 + D + \frac{1}{\gamma}}, && d \in \{1, ..., D\}, \\ Cov[Y_i, Y_j] &= -\frac{y_+(y_+ + D + \frac{1}{\gamma})}{1 + D + \frac{1}{\gamma}} \frac{1 + \frac{\theta_i}{\gamma}}{D + \frac{1}{\gamma}} \frac{1 + \frac{\theta_j}{\gamma}}{D + \frac{1}{\gamma}}, && i, j \in \{1, ..., D\},\ i \neq j, \end{aligned} \tag{3.3.3}$$

© University of Pretoria

where $y_+ = \sum_{i=1}^{D} y_i$.

Examples of the impact of the parameters on the PMF are given in figures 3.1a, 3.1b, 3.1c, and 3.1d. In all these figures, it is clear that the mode parameters correspond roughly to the mode proportions of the PMF. These figures also display the influence of the pseudo-variance parameter on the spread of the distribution. In particular, it is clear that in figures 3.1a and 3.1b, where the mode parameters are equal, for an increased value of the pseudo-variance, the spread of the PMF is greater. The same is true for figures 3.1c and 3.1d.

© University of Pretoria



(a) $\underline{\theta} = (\frac{1}{3}, \frac{1}{3})$ and $\gamma = 0.01$

(b) $\underline{\theta} = (\frac{1}{3}, \frac{1}{3})$ and $\gamma = 0.1$

(c) $\underline{\theta} = (0.6, 0.2)$ and $\gamma = 0.25$

(d) $\underline{\theta} = (0.6, 0.2)$ and $\gamma = 1$

Figure 3.1: Plots of the DM-MPV distribution (eq. (3.3.2)) with selected values of $\underline{\theta}$ and γ where $y_+ = 50$.

3.4 Dirichlet distribution with the PPV parameterisation

Considering the drawback of the MPV parameterisation, which is that it restricts the original parameters of the Dirichlet distribution and DM distribution, which in turn limits the flexibility of these distributions, another parameterisation is considered. This parameterisation is referred to as the PPV parameterisation, which uses as its parameters the mean of the Dirichlet distribution (mean proportions of the DM distribution) and again a pseudo-variance parameter.

First, the parameterisation of the Dirichlet distribution is defined in terms of its allele probabilities. This parameterisation is applied from a parameterisation of the DM distribution which can be found in the documentation of an R package on the DM distribution (Tvedebrink (2022)). This allele parameterisations solves the interpretability problem somewhat; however, it poses a problem for contaminating the DM distribution that will be proposed in the next chapter. Therefore, the PPV parameterisation is considered.

Definition 6 (Dirichlet distribution using allele probabilities (Tvedebrink (2022))). *Let* $\underline{P} = (P_1, ..., P_D)$

© University of Pretoria

denote a random vector on the Ω_D unit simplex where,

$$\Omega_D = \{(P_1, ..., P_D), 0 < P_i < 1, \sum_{i=1}^{D} P_i = 1\}, \tag{3.4.1}$$

*then the PDF of the **allele Dirichlet** distribution (D-MPV) is given by,*

$$f_{D-allele}(\underline{p}; \underline{\pi}, \gamma) = \frac{\Gamma(\frac{1-\gamma}{\gamma})}{\prod_{j=1}^{D} \Gamma(\gamma\pi_j)} \prod_{j=1}^{D} p_j^{\gamma\pi_j} \tag{3.4.2}$$

where each $\pi_i > 0$, and $0 < \gamma < 1$. We denote the random vector $\underline{P} = (P_1, ..., P_D) \sim D - allele(\underline{\pi}, \gamma)$.

Proposition 8 (Relationship between the classical and the allele parameterisation (Tvedebrink (2022))). *The original parameters of the Dirichlet distribution can be rewritten in terms of the allele parameterisation (see eq. (3.4.3)).*

$$\begin{aligned} \alpha_d &= \pi_d\gamma, \qquad d \in \{1, ..., D\}, \\ \alpha_+ &= \frac{1-\gamma}{\gamma}. \end{aligned} \tag{3.4.3}$$

Proof. When setting the parameters π_j and γ equal to the equations below,

$$\begin{aligned} \pi_j &= \alpha_j(1 + \alpha_+), \quad j \in \{1, ..., D\}, \\ \gamma &= \frac{1}{1 + \alpha_+}, \end{aligned} \tag{3.4.4}$$

and substituting them into eq. (3.4.2), one will obtain the eq. (2.1.4). Note that $\alpha_+ = \sum_{i=1}^{D} \alpha_i$.

Although this parameterisation solves the interpretability problem, it has the downside of having more restrictive bounds on the value of the parameter γ. This is a concern, since the ultimate goal of this study is to develop a contaminated DM model. When contaminating a distribution, which will be discussed in the next chapter, one will have to inflate the variance by a factor of η. The problem with the bounds of the pseudo-variance parameter, $0 < \gamma < 1$, is that when it is inflated, say $\eta\gamma$, there will need to be bounds on the inflation parameter as well, which makes the computation more cumbersome. Therefore, an alternative parameterisation will be proposed in this study - the PPV parameterisation (see definition 7 and proposition 9).

Definition 7 (Dirichlet distribution with PPV parameterisation). *Let $\underline{P} = (P_1, ..., P_D)$ denote a random vector on the Ω_D unit simplex where,*

$$\Omega_D = \{(P_1, ..., P_D), 0 < P_i < 1, \sum_{i=1}^{D} P_i = 1\}, \tag{3.4.5}$$

*then the PDF of the **proportional-mean-pseudo-variance reparameterised Dirichlet** distribution (D-PPV) is given by*

$$\begin{aligned} f_{D-PPV}(\underline{p}; \underline{\mu}, \gamma) &= \frac{\Gamma(\frac{1}{\gamma})}{\prod_{j=1}^{D} \Gamma(\frac{\mu_j}{\gamma})} \prod_{j=1}^{D} p_j^{\frac{\mu_j}{\gamma}} \\ &= \frac{\Gamma(\frac{1}{\gamma})}{\Gamma\left(\frac{1-\sum_{i=1}^{D-1}\mu_i}{\gamma}\right) \prod_{j=1}^{D-1} \Gamma(\frac{\mu_j}{\gamma})} p_j^{\frac{1-\sum_{i=1}^{D-1}\mu_i}{\gamma}} \prod_{j=1}^{D-1} p_j^{\frac{\mu_j}{\gamma}}, \end{aligned} \tag{3.4.6}$$

© University of Pretoria

where each $\mu_i > 0$, $\sum_{i=1}^{D} \mu_i = 1$, *denotes the mean proportions,* $\gamma > 0$ *is the pseudo-variance parameter. Note that though there are* D *mean proportion parameters, in practice there will be only* $D-1$, *since they are all positive and sum to one, implying that one of them need not be estimated, since its value can be calculated using all the other mean proportion parameters. We denote the random vector* $\underline{P} = (P_1, ..., P_D) \sim D - PPV(\underline{\mu}, \gamma)$.

Proposition 9 (Relationship between the classical and the PPV parameterisation). *The original parameters of the Dirichlet can be rewritten in terms of the PPV parameterisation (see eq. (3.4.7)).*

$$\begin{aligned} \alpha_d &= \frac{\mu_d}{\gamma}, \quad d \in \{1, ..., D\}, \\ \alpha_+ &= \frac{1}{\gamma}. \end{aligned} \tag{3.4.7}$$

Proof. When setting the parameters μ_j *and* γ *equal to the eq. (3.4.8),*

$$\begin{aligned} \mu_j &= \frac{\alpha_j}{\alpha_+}, \\ \gamma &= \frac{1}{\alpha_+},, \end{aligned} \tag{3.4.8}$$

and substituting them into eq. (3.4.6), one will obtain eq. (2.1.4). Note that $\alpha_+ = \sum_{i=1}^{D} \alpha_i$.

3.5 Dirichlet-multinomial distribution with the PPV parameterisation

The DM distribution will now be reparameterised in terms of the PPV parameterisation. This parameterisation utilised is given in proposition 9, which is the parameterisation utilised for the D-PPV. The DM distribution with the PPV parameterisation is given in definition 8.

Definition 8 (Dirichlet-multinomial with PPV parameterisation). *Let*

$$\begin{aligned} \underline{Y}|\underline{p} &\sim multinomial(y_+, \underline{p}) \\ \underline{P} &\sim D - PPV(\underline{\mu}, \gamma), \end{aligned} \tag{3.5.1}$$

then the random vector $\underline{Y} = (Y_1, ..., Y_D)$ *is said to follow a* ***Probability-Pseudo-Variance Dirichlet-Multinomial*** *(DM-PPV) distribution if its PMF is given by,*

$$\begin{aligned} f_{DM-PPV}(\underline{y}; \underline{p}, \gamma) &= \frac{y_+}{y_1!y_2!...y_D!} \frac{\Gamma(\frac{1}{\gamma})}{\prod_{d=1}^{D} \Gamma(\frac{p_d}{\gamma})} \frac{\prod_{d=1}^{D} \Gamma(y_d + \frac{p_d}{\gamma})}{\Gamma(y_+ + \frac{1}{\gamma})} \\ &= \frac{y_+}{y_1!y_2!...y_D!} \frac{\Gamma(\frac{1}{\gamma})}{\Gamma\left(\frac{1-\sum_{i=1}^{D-1} p_i}{\gamma}\right) \prod_{d=1}^{D-1} \Gamma(\frac{p_d}{\gamma})} \\ &\times \frac{\Gamma\left(y_D + \frac{1-\sum_{i=1}^{D-1} p_i}{\gamma}\right) \prod_{d=1}^{D-1} \Gamma(y_d + \frac{p_d}{\gamma})}{\Gamma(y_+ + \frac{1}{\gamma})} \end{aligned} \tag{3.5.2}$$

where each $p_i > 0$, $\sum_{i=1}^{D} p_i = 1$, *is the mean proportions and* $\gamma > 0$ *is the pseudo-variance. Note that* $y_+ = \sum_{i=1}^{D} y_i$. *Note that though there are* D *mean proportion parameters, in practice there will be only* $D-1$, *since they are all positive and sum to one, implying that one of them need not be estimated, since its value can be calculated using all the other mean proportion parameters. We denote the random vector* $\underline{Y} = (Y_1, ..., Y_D) \sim DM - PPV(\underline{p}, \gamma)$.

© University of Pretoria

Proposition 10 (Moments of the DM-PPV distribution). *From proposition 3, proposition 4, and proposition 5 it follows that the expected count for the d^{th} variable of the DM-PPV distribution, variance for the d^{th} variable of the DM-PPV distribution, and covariance between variables i and j of the DM-PPV distribution are given in eq. (3.5.3).*

$$\begin{aligned} E[Y_d] &= y_+ p_d, \\ Var[Y_d] &= y_+ p_d(1-p_d)\frac{y_+ + \frac{1}{\gamma}}{1+\frac{1}{\gamma}}, \\ Cov[Y_i, Y_j] &= -\frac{y_+(y_+ + \frac{1}{\gamma})}{1+\frac{1}{\gamma}} p_i p_j, \end{aligned} \tag{3.5.3}$$

where $d, i, j \in \{1, ..., D\}$; $i \neq j$, and $y_+ = \sum_{i=1}^{D} y_i$.

Examples of the impact of the parameters on the PMF are given in figures 3.2a, 3.2b, 3.2c, and 3.2d. These figures display the influence of the pseudo-variance parameter on the spread of the distribution. In particular, it is clear that in figures 3.2a and 3.2b, where the mean parameters are equal, for an increased value of the pseudo-variance, the spread of the PMF is greater. Figures 3.2c and 3.2d display the added flexibility of the PPV parameterisation over the MPV parameterisation, from which it is clear that the mode does not always exist.

3.6 Simulation study of the three different parameterisations of the Dirichlet-multinomial distribution

A parameter recovery of the three different parameterisations of the DM distributions will be completed in this section, similar to the parameter recovery by Otto et al. (2025). One hundred datasets were generated from a DM-MPV distribution with dimension $D = 3$ and parameters $\underline{\theta} = (\frac{1}{3}, \frac{1}{3})$ and $\gamma = 0.1353373$, since it is a subclass of the DM distribution and the DM-PPV distribution (the data were generated using the *simPop* function in the package by Tvedebrink (2022)). These parameters would equate to $\underline{\alpha} = (3.462982, 3.462982, 3.462982)$ for a DM distribution and $\underline{p} = (\frac{1}{3}, \frac{1}{3})$ and $\gamma = 0.09625615$ for the mean parameterisation. The DM distribution (eq. (2.1.6)), DM-MPV distribution (eq. (3.3.2)), and the DM-PPV distribution (eq. (3.5.2)) will be fitted to each dataset and the mean-squared-error (MSE) and bias of the parameter estimates will be calculated using eq. (3.6.1),

$$\begin{aligned} Bias(\hat{\kappa}) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\kappa} - \kappa) \\ MSE(\hat{\kappa}) &= \frac{1}{n}\sum_{i=1}^{n}(\hat{\kappa} - \kappa)^2 \end{aligned} \tag{3.6.1}$$

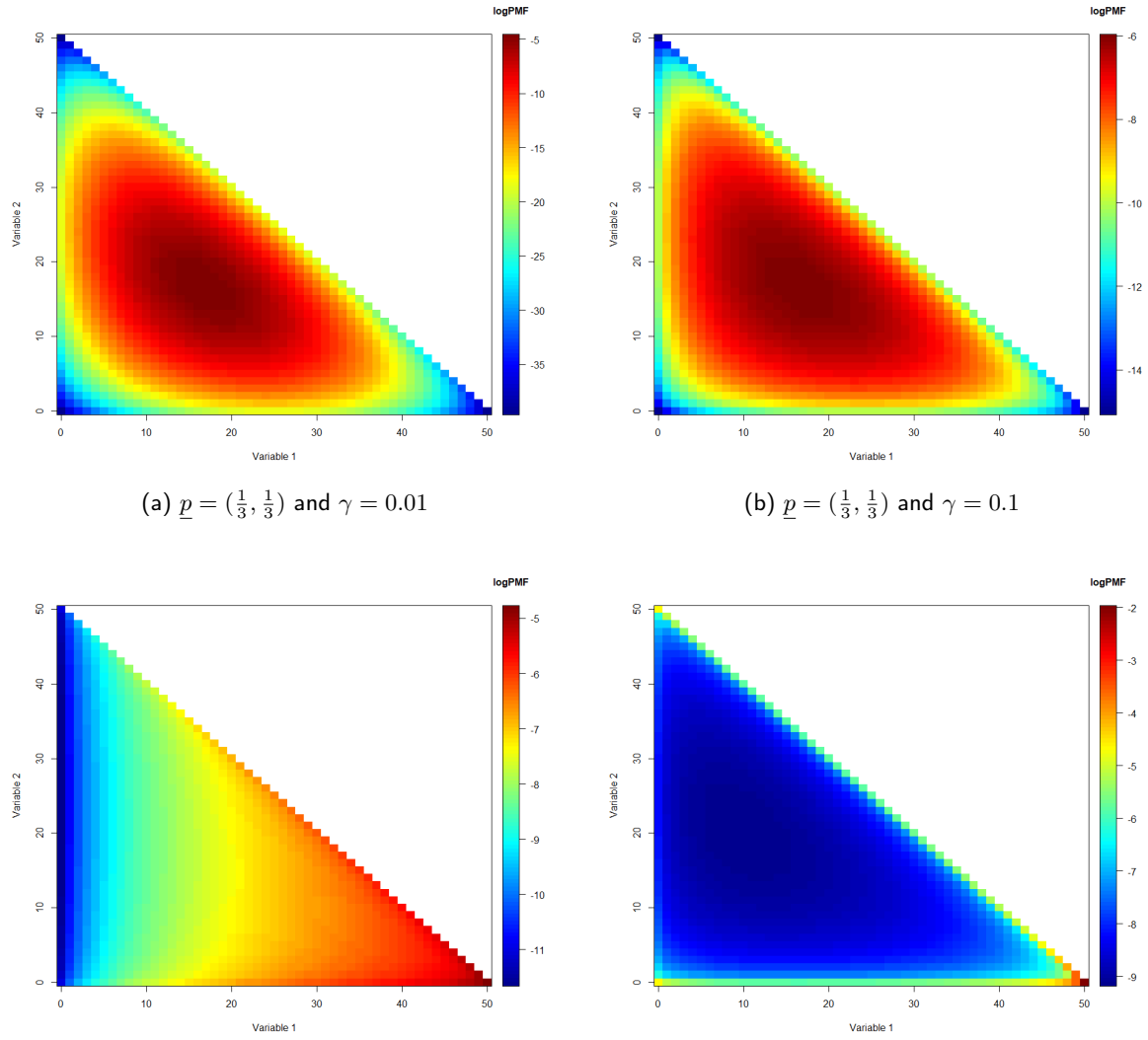
where n is the number of simulations and $\hat{\kappa}$ is the parameter in question.

A visual representation of the data is given in figure 3.3. This figure is a heatmap of two of the 100 generated datasets. From this it is clear that data have a mode.

3.6.1 Log-likelihood and evaluation criteria

The log-likelihood function of the DM-MPV distribution is given in eq. (3.6.2) and the log-likelihood function of the DM-PPV distribution is given in eq. (3.6.3). Both of these equations contain the

© University of Pretoria



(a) $\underline{p} = (\frac{1}{3}, \frac{1}{3})$ and $\gamma = 0.01$

(b) $\underline{p} = (\frac{1}{3}, \frac{1}{3})$ and $\gamma = 0.1$

(c) $\underline{p} = (0.6, 0.2)$ and $\gamma = 0.25$

(d) $\underline{p} = (0.6, 0.2)$ and $\gamma = 1$

Figure 3.2: Plots of the logarithm of the DM-PPV distribution (eq. (3.5.2)) with selected values of $\underline{p}$ and γ where $y_+ = 50$.

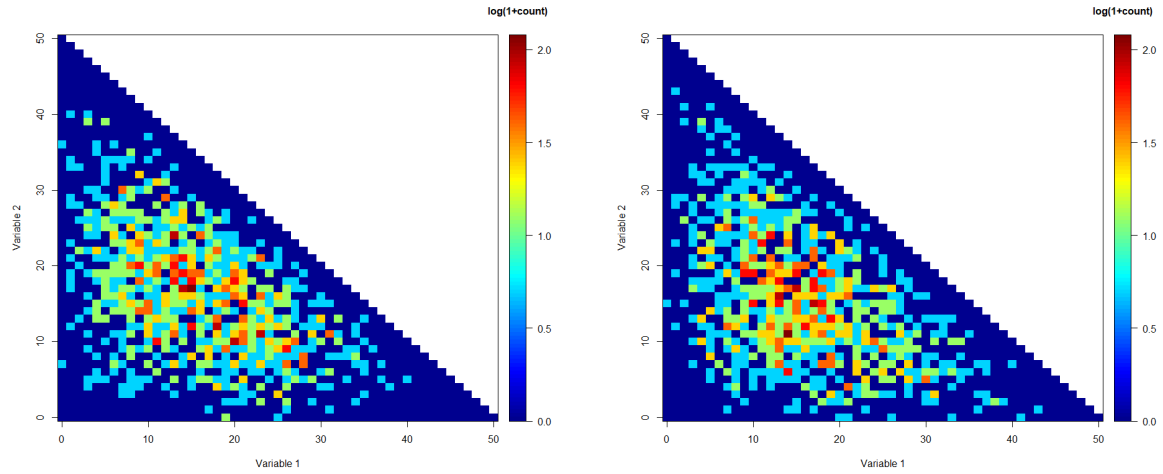


Figure 3.3: Heatmap of two generated datasets from the DM-MPV distribution (eq. (3.3.2)) in the simulation study.

© University of Pretoria

quantity $\sum_{i=1}^{n} log(\frac{y_+}{y_1!y_2!...y_D!})$, which will be constant over each iteration. It will be wise when coding this function to calculate this quantity once independently and then add it after calculating the rest of the equation to speed up computation.

From eq. (3.2.2) it follows that,

$$\begin{aligned} l(\underline{\kappa}) = &\sum_{i=1}^{n} log\left(\frac{y_+}{y_1!y_2!...y_D!}\right) \\ &+ \sum_{i=1}^{n} log\left(\frac{\Gamma(D+\frac{1}{\gamma})}{\Gamma\left(1+\frac{1-\sum_{i=1}^{D-1}\theta_i}{\gamma}\right)\prod_{d=1}^{D}\Gamma(1+\frac{\theta_d}{\gamma})}\right) \\ &+ \sum_{i=1}^{n} log\left(\frac{\Gamma\left(y_d+1+\frac{1-\sum_{i=1}^{D-1}\theta_i}{\gamma}\right)\prod_{d=1}^{D-1}\Gamma(y_d+1+\frac{\theta_d}{\gamma})}{\Gamma(y_++D+\frac{1}{\gamma})}\right), \end{aligned} \tag{3.6.2}$$

where $\underline{\kappa} = (\underline{\theta}, \gamma)$.

From eq. (3.4.6) it follows that,

$$\begin{aligned} l(\underline{\kappa}) = &\sum_{i=1}^{n} log\left(\frac{y_+}{y_1!y_2!...y_D!}\right) \\ &+ \sum_{i=1}^{n} log\left(\frac{\Gamma(\frac{1}{\gamma})}{\Gamma\left(\frac{1-\sum_{i=1}^{D-1}p_i}{\gamma}\right)\prod_{d=1}^{D-1}\Gamma(\frac{p_d}{\gamma})}\right) \\ &+ \sum_{i=1}^{n} log\left(\frac{\Gamma\left(y_d+\frac{1-\sum_{i=1}^{D-1}p_i}{\gamma}\right)\prod_{d=1}^{D-1}\Gamma(y_d+\frac{p_d}{\gamma})}{\Gamma(y_++\frac{1}{\gamma})}\right), \end{aligned} \tag{3.6.3}$$

where $\underline{\kappa} = (\underline{p}, \gamma)$.

To evaluate the performance of the models presented, the Akaike information criterion (AIC, eq. (3.6.4), Akaike (1974), Subedi et al. (2020)) and the Bayesian information criterion (BIC, eq. (3.6.5), Schwarz (1978), Subedi et al. (2020)) will be employed. For both of these criterion, a lower value indicates a better fit. Note that the quantity ψ refers to the number of parameters in the model and $\hat{\underline{\kappa}}$ refers to the vector of parameter estimates.

$$AIC(\hat{\underline{\kappa}}) = -2l(\hat{\underline{\kappa}}) + 2\psi. \tag{3.6.4}$$

$$BIC(\hat{\underline{\kappa}}) = -2l(\hat{\underline{\kappa}}) + \psi \log n. \tag{3.6.5}$$

3.6.2 Modelling the parameters

Similar to the parameters of the DM distribution, these reparameterised versions have parameters that are bounded. Both versions have a pseudo-variance parameter that needs to be positive. The pseudo-variance parameter will be dealt with in the same way that the parameters of the DM distribution are dealt with. The mode parameters of the DM-MPV distribution and the probability parameters of the DM-PPV distribution have more complex bounds. Both these parameter vectors must be bound such that every element of the vector must be positive, and also sum to one. To ensure that these bounds are met, the logistic function is used. These parameters are modelled using a regression, similar to

© University of Pretoria

the regression of the DM distribution by Subedi et al. (2020) and Bartolucci et al. (2021) discussed in section 2.3, and are given in proposition 3.6.2.

Proposition 11. *The mode parameters for the DM-MPV distribution will be coded as follows. Assuming D variables, there will be $D-1$ mode parameters, say $\beta_1, ..., \beta_{D-1}$, which are modelled as follows.*

$$\begin{aligned} \theta_j &= \frac{exp(\beta_j)}{1+\sum_{j=1}^{D-1} exp(\beta_j)}, \quad for\ j < D, \\ \theta_D &= \frac{1}{1+\sum_{j=1}^{D-1} exp(\beta_j)}. \end{aligned} \tag{3.6.6}$$

The back transformations for these mode parameters are as follows.

$$\beta_j = log(\theta_j) + |log(\theta_D)| \tag{3.6.7}$$

The mean parameters will for the DM-PPV distribution be coded as follows. Assuming D variables, there will be $D-1$ mode parameters, say $\beta_1, ..., \beta_{D-1}$, which are modelled as follows.

$$\begin{aligned} p_j &= \frac{exp(\beta_j)}{1+\sum_{j=1}^{D-1} exp(\beta_j)}, \quad for\ j < D, \\ p_D &= \frac{1}{1+\sum_{j=1}^{D-1} exp(\beta_j)}. \end{aligned} \tag{3.6.8}$$

The back transformations for these mean parameters are as follows.

$$p_j = log(p_j) + |log(p_D)| \tag{3.6.9}$$

The pseudo-variance parameter for both the DM-MPV distribution and DM-PPV distribution version will be modelled as in eq. (3.6.10).

$$\gamma = exp(\beta_D). \tag{3.6.10}$$

The back transformations for the pseudo-variance parameter is as follows.

$$\beta_D = log(\gamma). \tag{3.6.11}$$

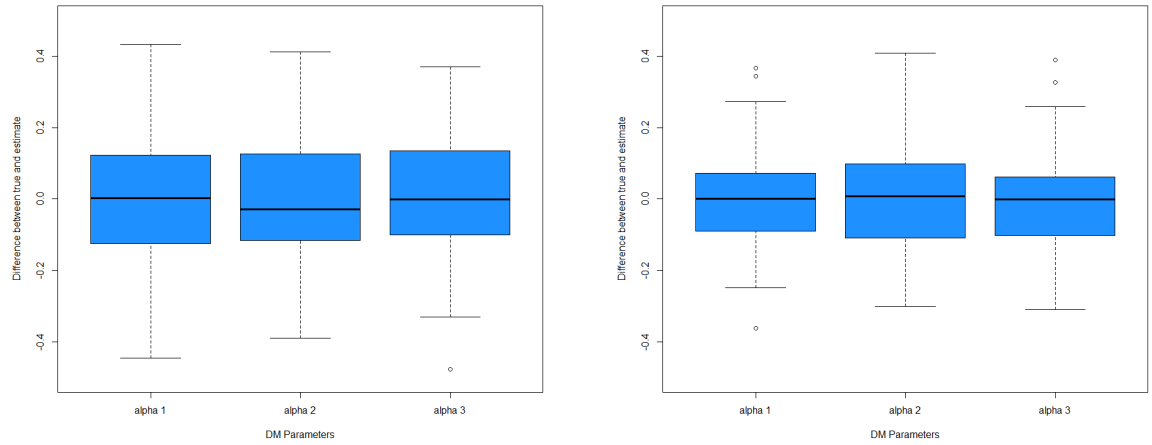
3.6.3 Results from the simulation study

The bias and MSE of all three parameterisations considered decreases as the sample size increases. This is evident in figure 3.4 for the DM distribution, figure 3.5 for the DM-MPV distribution, and figure 3.6 for the DM-PPV distribution. It should be noted that the DM-MPV distribution and DM-PPV distribution have a lower bias than the DM distribution at every sample size. This is mainly down to the bounds of the DM-MPV distribution and DM-PPV distribution. Consider the mode parameters in the DM-MPV distribution or the mean parameters in the DM-PPV distribution. These parameters are bound such that they must sum to 1. This means that there is a limit to the bias and MSE for these parameters. It should be noted that the pseudo-variance parameter has a lower bias and MSE than any of the parameters of the DM distribution. This is down to the low variance of the simulated data. It is assumed that if the variance of the generated data were increased, the bias and MSE of the pseudo-variance parameter would increase.

3.7 Conclusion

The reparameterised distributions have parameters which are easier to understand and interpret. These reparameterised distributions also result in parameter estimates that have a lower bias and MSE, but this appears to be down to the bounds of these reparameterised parameters. In the next chapter, these reparameterised distributions are used to develop contaminated DM models with the goal of addressing outliers and providing robust parameter estimates.

© University of Pretoria



(a) 500 observations

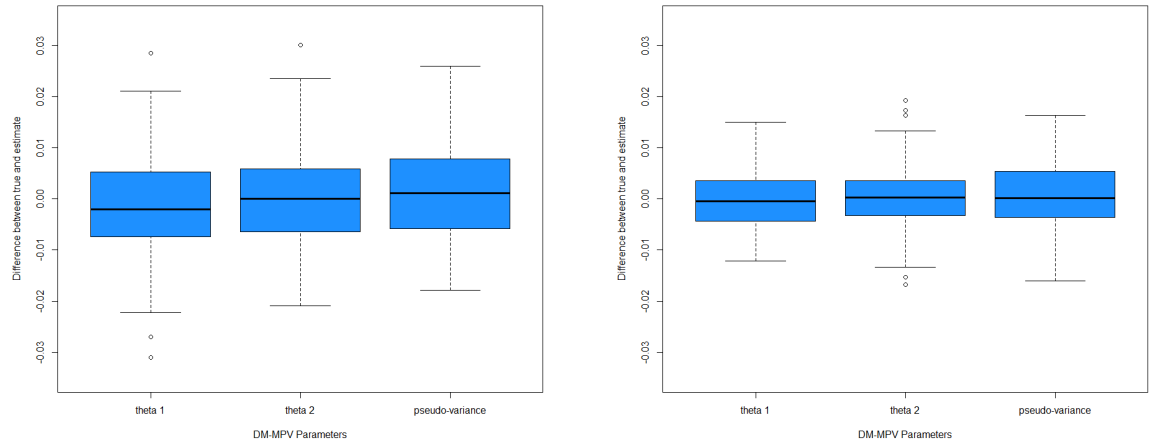
(b) 1000 observations

Figure 3.4: Boxplots of the difference between true and estimated parameters of the DM distribution (eq. (2.1.6)).

Table 3.1: Results for the DM distribution (eq. (2.1.6)) fitted to the simulated data.

Sample size		Mean	Median	Lower CI (90%)	Upper CI (90%)	Bias	MSE
100	α_1	3.51824	3.45217	2.85332	4.25359	-0.05526	0.23699
	α_2	3.46567	3.36356	2.81825	4.26188	-0.00269	0.21095
	α_3	3.51192	3.45232	2.82335	4.40004	-0.04894	0.20751
	$l(\underline{\kappa})$	-667.51717	-667.84215	-678.77862	-655.00544	-	-
	AIC	1341.03433	1341.68429	1312.80032	1362.14949	-	-
	BIC	1348.84984	1349.49980	1320.61583	1369.96500	-	-
200	α_1	3.44027	3.44444	2.87380	3.90759	0.02271	0.08951
	α_2	3.51907	3.47270	2.99663	4.13639	-0.05608	0.11038
	α_3	3.48961	3.48500	2.94343	4.03137	-0.02663	0.08785
	$l(\underline{\kappa})$	-1335.16035	-1335.64866	-1352.98375	-1319.10517	-	-
	AIC	2676.32069	2677.29733	2640.98094	2711.26522	-	-
	BIC	2686.21565	2687.19228	2650.87590	2721.16017	-	-
500	α_1	3.45747	3.46458	3.11527	3.76761	0.00551	0.03990
	α_2	3.46628	3.43364	3.19175	3.77465	-0.00330	0.03296
	α_3	3.47268	3.46206	3.18158	3.75034	-0.00970	0.02969
	$l(\underline{\kappa})$	-3341.31463	-3342.09798	-3365.46586	-3317.90517	-	-
	AIC	6688.62927	6690.19596	6639.99287	6736.02504	-	-
	BIC	6701.27309	6702.83978	6652.63669	6748.66887	-	-
1000	α_1	3.45837	3.46362	3.25436	3.64297	0.00461	0.01590
	α_2	3.46201	3.46998	3.25450	3.65878	0.00097	0.01868
	α_3	3.45874	3.46270	3.26727	3.68444	0.00424	0.01710
	$l(\underline{\kappa})$	-6681.46630	-6681.54529	-6718.75055	-6649.83079	-	-
	AIC	13368.93260	13369.09058	13294.58980	13440.51926	-	-
	BIC	13383.65587	13383.81385	13309.31307	13455.24253	-	-

© University of Pretoria



(a) 500 Observations

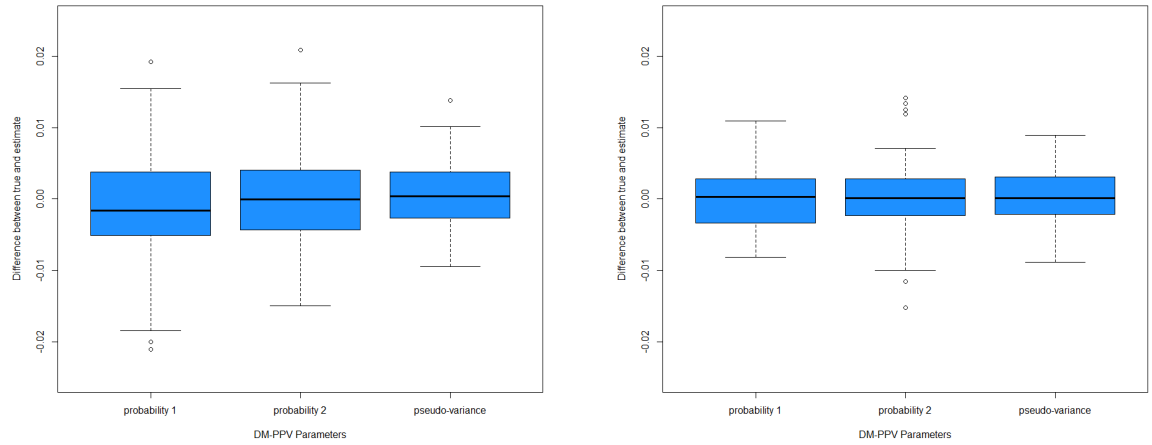
(b) 1000 Observations

Figure 3.5: Boxplots of the difference between true and estimated parameters of the DM-MPV distribution (eq. (3.3.2)).

Table 3.2: Results for the DM-MPV distribution (eq. (3.3.2)) fitted to the simulated data.

Sample size		Mean	Median	Lower CI (90%)	Upper CI (90%)	Bias	MSE
100	θ_1	0.33574	0.33352	0.29862	0.36830	-0.00241	0.00055
	θ_2	0.32911	0.32916	0.29197	0.35891	0.00422	0.00050
	γ	0.13738	0.13681	0.10159	0.17445	-0.00205	0.00051
	$l(\underline{\kappa})$	-667.51476	-667.83828	-678.77863	-655.00544	-	-
	AIC	1341.02951	1341.67657	1312.80032	1362.14949	-	-
	BIC	1348.84502	1349.49208	1320.61583	1369.96500	-	-
200	θ_1	0.32787	0.32690	0.30006	0.35231	0.00546	0.00025
	θ_2	0.33774	0.33813	0.31486	0.35759	-0.00441	0.00022
	γ	0.13579	0.13603	0.11086	0.16452	-0.00046	0.00023
	$l(\underline{\kappa})$	-1335.15294	-1335.64548	-1352.98741	-1319.10516	-	-
	AIC	2676.30588	2677.29097	2640.98094	2711.29154	-	-
	BIC	2686.20083	2687.18592	2650.87589	2721.18649	-	-
500	θ_1	0.33203	0.33136	0.31321	0.35317	0.00130	0.00012
	θ_2	0.33368	0.33332	0.31732	0.35158	-0.00035	0.00010
	γ	0.13656	0.13651	0.12132	0.15125	-0.00122	0.00009
	$l(\underline{\kappa})$	-3341.28992	-3342.09423	-3365.39049	-3317.91197	-	-
	AIC	6688.57983	6690.18846	6639.82577	6736.06098	-	-
	BIC	6701.22366	6702.83229	6652.46960	6748.70481	-	-
1000	θ_1	0.33314	0.33284	0.32361	0.34233	0.00019	0.00003
	θ_2	0.33355	0.33358	0.32400	0.34542	-0.00022	0.00004
	γ	0.13616	0.13547	0.12566	0.14777	-0.00082	0.00004
	$l(\underline{\kappa})$	-6681.40639	-6681.49779	-6718.99819	-6649.90165	-	-
	AIC	13368.81279	13368.99557	13294.86157	13440.52062	-	-
	BIC	13383.53605	13383.71884	13309.58484	13455.24388	-	-

© University of Pretoria



(a) 500 Observations

(b) 1000 Observations

Figure 3.6: Boxplots of the difference between true and estimated parameters of the DM-PPV distribution (eq. (3.5.2)).

Table 3.3: Results for the DM-PPV distribution (eq. (3.5.2)) fitted to the simulated data.

Sample size		Mean	Median	Lower CI (90%)	Upper CI (90%)	Bias	MSE
100	p_1	0.33513	0.33349	0.30778	0.36069	-0.00179	0.00029
	p_2	0.33007	0.33048	0.30389	0.35200	0.00326	0.00025
	γ	0.09679	0.09700	0.07784	0.11456	-0.00053	0.00013
	$l(\underline{\kappa})$	-667.52586	-667.83828	-678.77862	-655.00544	-	-
	AIC	1341.05173	1341.67657	1312.80032	1362.15574	-	-
	BIC	1348.86724	1349.49208	1320.61583	1369.97125	-	-
200	p_1	0.32924	0.32951	0.30887	0.34695	0.00410	0.00012
	p_2	0.33660	0.33655	0.32035	0.35189	-0.00326	0.00011
	γ	0.09621	0.09657	0.08321	0.11028	0.00005	0.00006
	$l(\underline{\kappa})$	-1335.15073	-1335.65505	-1352.98375	-1319.10516	-	-
	AIC	2676.30146	2677.31010	2640.98096	2711.26516	-	-
	BIC	2686.19642	2687.20505	2650.87591	2721.16011	-	-
500	p_1	0.33232	0.33168	0.31803	0.34720	0.00101	0.00006
	p_2	0.33357	0.33325	0.32204	0.34603	-0.00024	0.00005
	γ	0.09678	0.09662	0.08934	0.10374	-0.00053	0.00002
	$l(\underline{\kappa})$	-3341.31049	-3342.09816	-3365.38317	-3317.91418	-	-
	AIC	6688.62098	6690.19632	6639.88821	6736.05840	-	-
	BIC	6701.26481	6702.84015	6652.53203	6748.70223	-	-
1000	p_1	0.33329	0.33360	0.32677	0.34037	0.00004	0.00002
	p_2	0.33339	0.33347	0.32508	0.34026	-0.00006	0.00002
	γ	0.09666	0.09636	0.09102	0.10266	-0.00040	0.00001
	$l(\underline{\kappa})$	-6681.40128	-6681.49228	-6718.75350	-6649.86202	-	-
	AIC	13368.80255	13368.98456	13294.62924	13440.49995	-	-
	BIC	13383.52582	13383.70783	13309.35251	13455.22322	-	-

© University of Pretoria

Chapter 4

Contaminated Dirichlet-multinomial distributions

Contamination is a methodological approach where a given distribution can be made to account for outliers (Tomarchio et al. (2024)) through a specific application of mixture models. Contaminated models have been applied to many distributions (see Punzo et al. (2018), Punzo (2019), Tomarchio et al. (2024), Tomarchio and Punzo (2020) for examples). In this chapter, the contamination technique will be applied to the DM-MPV distribution and DM-PPV distribution. This is part of the justification for constructing reparameterised versions of the DM distribution in the previous chapter.

In the remainder of the chapter the following will be done: in section 4.1 the technique of contamination will be defined, in section 4.2 the contaminated versions of the DM-MPV distribution and DM-PPV distribution will be proposed, in section 4.3 an algorithm for selecting initial values of the DM-CMPV model and DM-CPPV model will be given, and in section 4.4 a sensitivity analysis will be performed to display the value of the contaminated models.

4.1 Contaminated distributions

In contaminating a distribution, a mixture model is employed where there are two components (Otto et al. (2025)). The first component is the reference component, while the second component is the contaminant (Otto et al. (2025)). The contaminant component has the location parameter values equal to that of the reference component, but has an inflated variance or pseudo-variance parameter (Otto et al. (2025)). This yields a mixture model where the location parameters are the same for both the components, while the variance or pseudo-variance differs (Otto et al. (2025)). A contaminated model is defined more formally in definition 9.

Definition 9 (Contaminated distribution (Otto et al. (2025))). *Consider a distribution with PDF, $f(\underline{y}, \underline{\theta}, \gamma)$, where the parameter $\underline{\theta}$ is a location parameter (ideally the mode) and γ is a parameter that is closely related to the variance of the distribution (ideally the variance). A contaminated version of the distribution with PDF or PMF, $f(\underline{y}, \underline{\theta}, \gamma)$, can be constructed as follows.*

$$\begin{aligned} f(\underline{y}; \underline{\theta}, \gamma, \eta, \pi_c) &= (1-\pi_c) f(\underline{y}; \underline{\theta}, \gamma) \quad + \quad \pi_c f(\underline{y}; \underline{\theta}, \eta\gamma) \\ &= (1-\pi_c) * reference \quad + \pi_c * \ contaminant \end{aligned} \tag{4.1.1}$$

The parameter η is an inflation parameter that needs to be greater than one ($\eta > 1$). The parameter π_c is the mixing parameter associated with the outliers or atypical observations and it is assumed that $0 < \pi_c < 1$.

© University of Pretoria

The reference distribution referred to in definition 9 is the target distribution referred to by Davies and Gather (1993), which can broadly be defined as the distribution from which the non-outlier observations are generated (Davies and Gather (1993)).

4.2 Contaminated Dirichlet-multinomial distributions

A contaminated version of the Dirichlet distribution was developed by Tomarchio et al. (2024), where the MPV parameterisation was utilised for the Dirichlet distribution. Considering the close relationship between the Dirichlet distribution and the DM distribution, the contaminated Dirichlet model under the MPV parameterisation was utilised as an inspiration to develop the contaminated DM-MPV model (DM-CMPV, definition 10 follows from definition 9). However, knowing the decreased flexibility of the DM-MPV distribution, contamination was also applied to the DM-PPV distribution (DM-CPPV, definition 11 follows from definition 9).

Definition 10 (Contaminated Dirichlet-multinomial with MPV parameterisation)**.** *Suppose the random vector* $\underline{Y} = (Y_1, ..., Y_D)$ *follows a DM-MPV distribution with mode parameters* $\underline{\theta} = (\theta_1, ..., \theta_D)$ *and pseudo-variance* γ*. A contaminated version of the DM-MPV distribution can be constructed as follows (eq. (4.2.1)).*

$$\begin{aligned} f(\underline{y}; \underline{\theta}, \gamma, \eta, \pi_c) = (1-\pi_c)\frac{y_+}{y_1!y_2!...y_D!}\frac{\Gamma(d+\frac{1}{\gamma})}{\prod_{d=1}^{D}\Gamma(1+\frac{\theta_d}{\gamma})}\frac{\prod_{d=1}^{D}\Gamma(y_d+1+\frac{\theta_d}{\gamma})}{\Gamma(\sum_{d=1}^{D}y_d+1+\frac{\theta_d}{\gamma})} \\ + \pi_c\frac{y_+}{y_1!y_2!...y_D!}\frac{\Gamma(d+\frac{1}{\eta\gamma})}{\prod_{d=1}^{D}\Gamma(1+\frac{\theta_d}{\eta\gamma})}\frac{\prod_{d=1}^{D}\Gamma(y_d+1+\frac{\theta_d}{\eta\gamma})}{\Gamma(\sum_{d=1}^{D}y_d+1+\frac{\theta_d}{\eta\gamma})}, \end{aligned} \tag{4.2.1}$$

where $\sum_{i=1}^{D}\theta_i = 1$, $\underline{\theta} > \underline{0}$, $\gamma > 0$, $\eta > 1$, *and* $0 < \pi_c < 1$*. Note that* $y_+ = \sum_{i=1}^{D} y_i$*. This contaminated model has two added parameters,* η *which is the inflation parameter and* π_c *which is the proportion of data that are outliers. This model is referred to as the* ***contaminated Dirichlet-multinomial mode-pseudo-variance*** *(DM-CMPV) parameterisation. Note that since the modal proportion parameters are positive and sum to one, this means that one of the parameters need not be estimated, implying that the model has only* $D-1$ *modal proportion parameters that need to be estimated.*

Definition 11 (Contaminated Dirichlet-multinomial with PPV parameterisation)**.** *Suppose the random vector* $\underline{Y} = (Y_1, ..., Y_D)$ *follows a DM-PPV distribution with proportional mean parameters* $\underline{p} = (p_1, ..., p_D)$ *and pseudo-variance* γ*. A contaminated version of the DM-PPV distribution can be constructed as follows (eq. (4.2.2)).*

$$\begin{aligned} f(\underline{y}; \underline{p}, \gamma, \eta, \pi_c) = (1-\pi_c)\frac{y_+}{y_1!y_2!...y_D!}\frac{\Gamma(\frac{1}{\gamma})}{\prod_{d=1}^{D}\Gamma(\frac{p_d}{\gamma})}\frac{\prod_{d=1}^{D}\Gamma(y_d+\frac{p_d}{\gamma})}{\Gamma(\sum_{d=1}^{D}y_d+\frac{p_d}{\gamma})} \\ + \pi_c\frac{y_+}{y_1!y_2!...y_D!}\frac{\Gamma(\frac{1}{\eta\gamma})}{\prod_{d=1}^{D}\Gamma(\frac{p_d}{\eta\gamma})}\frac{\prod_{d=1}^{D}\Gamma(y_d+\frac{p_d}{\eta\gamma})}{\Gamma(\sum_{d=1}^{D}y_d+\frac{p_d}{\eta\gamma})}, \end{aligned} \tag{4.2.2}$$

where $\sum_{i=1}^{D}p_i = 1$, $\underline{p} > \underline{0}$, $\gamma > 0$, $\eta > 1$, *and* $0 < \pi_c < 1$*. Note that* $y_+ = \sum_{i=1}^{D} y_i$*. This contaminated model has two added parameters,* η *which is the inflation parameter and* π_c *which is the proportion of data that are outliers. This model is referred to as the* ***contaminated Dirichlet-multinomial proportional-mean-pseudo-variance*** *(DM-CPPV) parameterisation. Note that since the mean proportion parameters are positive and sum to one, this means that one of the parameters need not be estimated, implying that the model has only* $D-1$ *mean proportion parameters that need to be estimated.*

© University of Pretoria

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

For both the DM-CMPV and DM-CPPV, if the inflation parameter has value 1 ($\eta = 1$) and/or the mixing parameter has value 0 ($\pi_c = 0$), then the DM-CMPV becomes the DM-MPV and the DM-CPPV becomes the DM-PPV.

4.3 Algorithm for coding the contaminated Dirichlet-multinomial distributions

Care needs to be taken when applying the contaminated DM models to data as the parameters are bounded. The bounds of the parameters of the reparameterised DM distributions have been discussed in the previous chapter and a method for modelling these parameters such that these bounds are maintained has been given in subsection 3.6.2. Now it remains to present a method for modelling the parameters unique to contamination such that their bounds are met. The bounds for the inflation parameter are $\eta > 1$ and for the mixing proportion they are $0 < \pi_c < 1$. As mentioned in subsection 3.6.2, the parameters are modelled as though they are a regression similar to a regression of the DM distribution by Subedi et al. (2020) and Bartolucci et al. (2021). This is also true for the parameters in the contamination. The contamination mixing proportion parameter is modelled using a logistic function, while the inflation parameter is modelled using an exponential. This is shown in proposition 12.

Proposition 12 (Modelling the inflation parameter and mixing parameter of the contaminated models)**.** *The inflation (eq. (4.3.1)) and mixing proportion (eq. (4.3.2)) parameters are modelled as follows. Eq. (4.3.1) and eq. (4.3.2) are utilised for both the mode and the mean parameterisation.*

$$\eta = 1 + exp(\beta_{D+1}), \tag{4.3.1}$$

$$\pi_c = \frac{exp(\beta_{D+2})}{1 + exp(\beta_{D+2})}. \tag{4.3.2}$$

The back transformation for eq. (4.3.1) is given in eq. (4.3.3) and the back transformation for eq. (4.3.2) is given in eq. (4.3.4).

$$\beta_{D+1} = log(\eta - 1), \tag{4.3.3}$$

$$\beta_{D+2} = log\left(\frac{\pi_c}{1 - \pi_c}\right). \tag{4.3.4}$$

Though the contaminated models are mixture models, the expectation-maximisation algorithm is not implemented. Fitting this model without the expectation-maximisation algorithm appeared to work better than using the expectation-maximisation algorithm, therefore it was not used.

Algorithm 1 (Initialisation for the DM-CMPV model)**.** *To fit the DM-CMPV model, the following algorithm can be followed.*

- *Fit the DM-MPV distribution. Take as initial values for the mode parameters the output of the pam function in R (available in the package by Maechler et al. (2013)), as was done by Subedi et al. (2020).*
- *Take the parameters from the estimated DM-MPV distribution and implement them as initial values for the DM-CMPV model. Do, however, for the pseudo-variance, decrease the value slightly.*
- *An initial values for the inflation parameter (η) must be a fairly large value.*

© University of Pretoria

- *As initial values for the mixing parameter (π_c) a fairly size-able value needs to be employed, say 0.25.*

Algorithm 2 (Initialisation for the DM-CPPV model)**.** *In the DM-CMPV model, the following algorithm can be followed.*

- *Fit the DM-PPV distribution. Take as initial values the proportional means of the variables.*
- *Take the parameters from the estimated DM-PPV distribution and implement them as initial values for the DM-CPPV model. Do, however, for the pseudo-variance, decrease the value slightly.*
- *An initial value for the inflation parameter (η) must be a value that is fairly small. Bear in mind that for too large an initial value of the inflation parameter, the DM-PPV distribution could result in a bathtub shaped distribution, which has in experience resulted in a fit that ignores the contamination component entirely ($\pi_c = 0$).*
- *As initial values for the mixing parameter (π_c) a fairly size-able value needs to be employed, say 0.25.*

Initialisation advice for fitting the DM-CMPV and DM-CPPV are given in algorithms 1 and 2. The reason for not choosing initial values such that the contaminated models converge to the reference distributions, is because this did not work reliably. It should be noted that the pseudo-variance and inflation parameters of the two parameterisations are not directly comparable. It has also been experienced that the DM-CPPV model is a more sensitive model than the DM-CMPV model. It is thought that this may be a result of its flexibility. More care needs to be taken when fitting the DM-CPPV model than the DM-CMPV model, since one can with too high an initial value of the pseudo-variance parameter result in a distribution that has a bathtub shape.

4.4 Sensitivity analysis

Similar to the paper by Otto et al. (2025), where a contaminated beta-binomial model was considered (see eq. (2.1.7) for the PMF of the beta-binomial distribution), a sensitivity analysis is performed to test whether the DM-CMPV model and DM-CPPV model can provide robust estimates of the mode/mean and variance with respect to outliers.

This will be done by generating data from a DM-MPV distribution, since it is a subclass of the rest of the distributions and models that will be tested, with mode $(\frac{1}{3}, \frac{1}{3})$ and a variance of $\gamma = 0.006739947$ (the data were generated using the *simPop* function in the package by Tvedebrink (2022)). For the DM-PPV distribution this corresponds to mean parameters $\underline{p} = (\frac{1}{3}, \frac{1}{3})$ and pseudo-variance parameter $\gamma = 0.006606367$.

Thereafter, observations that represent outliers will be added to the data. In total, 100 such datasets will be generated. Two different sensitivity analyses will be performed. In subsection 4.4.1 one point that is a clear outlier will be used to test the ability of the contaminated models to provide robust location parameter estimates, while for the second analysis in subsection 4.4.2, outliers will be scattered around the mode of the data to test the ability of the contaminated models to provide robust pseudo-variance parameter estimates.

4.4.1 Analysis 1

In the first analysis, a point is chosen that is clearly not contained within the span of the generated data. This point is then be added to the data multiple times such that it accounts for 20% of the data. In the end there are 100 datasets with 500 observations each, where 400 of these observations are generated from the DM-MPV distribution and 100 observations are the point chosen as an outlier.

© University of Pretoria

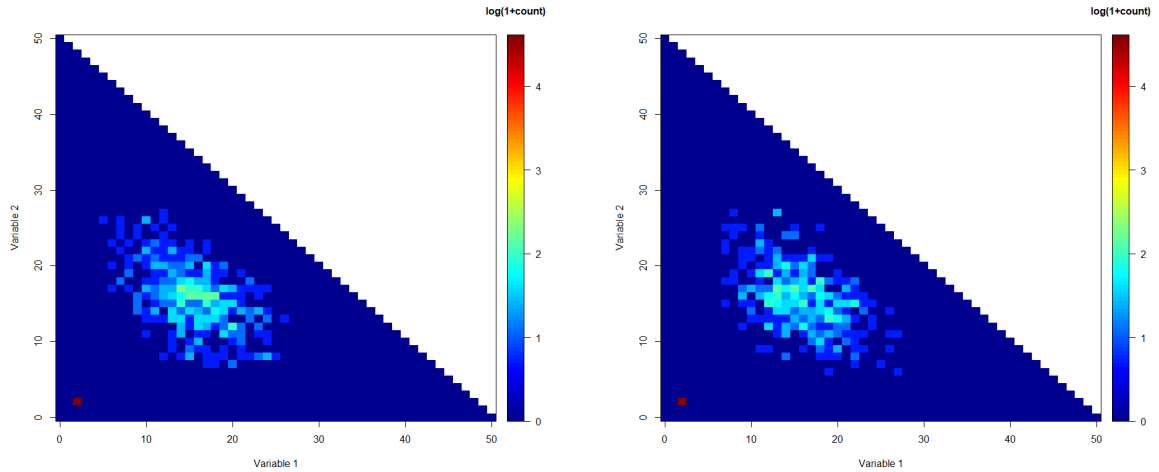


Figure 4.1: Heatmap of two generated datasets from the DM-MPV distribution (3.3.2) for the sensitivity analysis number 1.

The DM-MPV distribution (eq. (3.3.2)), DM-PPV distribution (eq. (3.5.2)), DM-CMPV model (eq. (4.2.1)), and DM-CPPV model (eq. (4.2.2)) are fitted to this data to see how these models account for this outlier case as well as the ability of the contaminated models to provide robust estimates of the location parameters.

A heatmap of two of the generated datasets is given in figure 4.1 to show what the data looks like. It is clear that the data were generated from a DM-MPV distribution and a point contained in the set of possible values for this DM-MPV distribution was chosen as being a reasonable representation of an outlier and added to the data. This sensitivity analysis shows that the contaminated DM models are capable of accounting for this outlier.

The table 4.1 displays the results and figure 4.2 displays the difference between the true and estimated location parameters. When comparing the DM-MPV distribution with the DM-CMPV model and the DM-PPV distribution with the DM-CPPV model, the contaminated models have substantially lower AIC and BIC values, which indicates that they are better models of the data. Considering the values for the location parameter of the DM-MPV distribution, it is clear that the outliers had a significant effect on these parameter estimates, while for the contaminated models, the DM-CMPV model and DM-CPPV model, it did not. This displays the robustness of the location parameter estimates of contaminated models in the presence of outliers.

Considering the value of the pseudo-variance parameter of the DM-MPV in table 4.1, this indicates that the DM-MPV distribution tended towards uniformity. The same applies for the contaminant part of the DM-CMPV. This is because for very large values of γ in the DM-MPV distribution, when translated to the traditional parameterisation (see (eq. 3.2.3) in chapter 3), results in parameters $\underline{\alpha} \to \underline{1}$. Considering eq. (2.2.3) in chapter 2, this implies a uniform distribution. This indicates that the mode parameters of the DM-MPV distribution become irrelevant and that the DM-MPV distribution can absolutely not account for the outliers. Considering figure 4.2, this is the reason that the DM-MPV distribution is not included since its mode parameters are all the same.

The true positive rate and false positive rate for identifying outliers was calculated for the contaminated models after every simulation. For both contaminated models the true positive rate was 1 in every instance. The average false positive rate for the DM-CMPV was 0.18205 and for the DM-CPPV was 0.129975. The DM-CPPV therefore performed better on this front.

© University of Pretoria

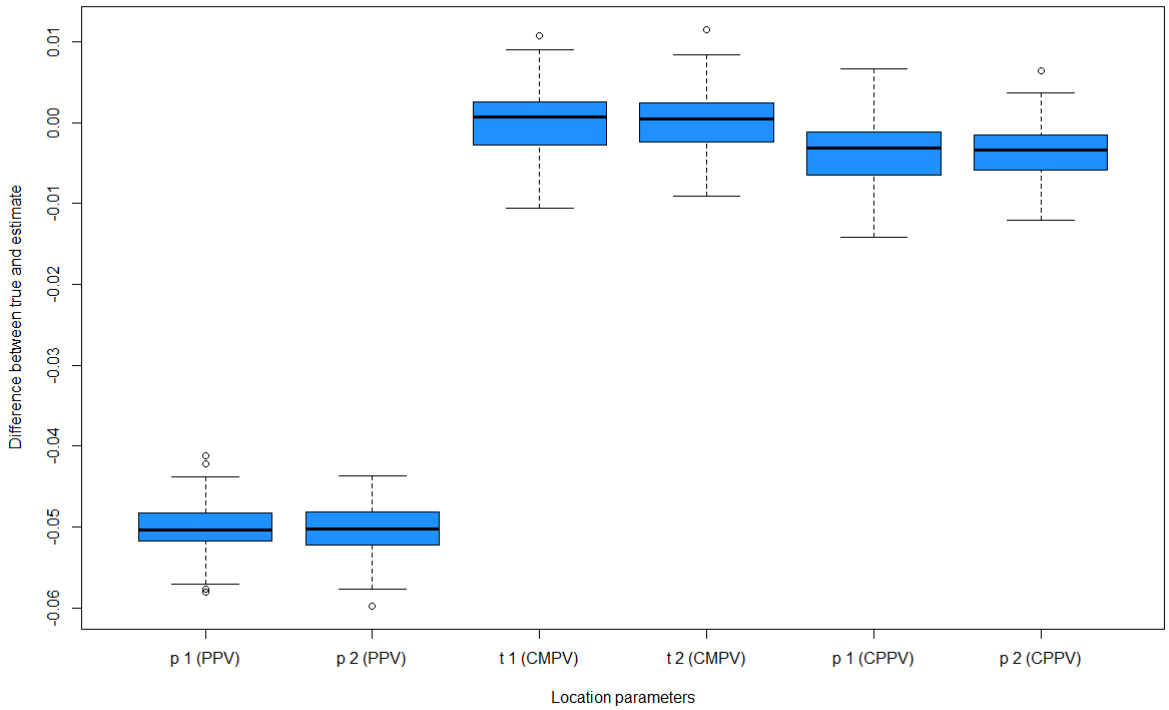


Figure 4.2: Boxplots of the difference between true and estimated values of the location parameters for the DM-PPV, DM-CMPV, and DM-CPPV in analysis 1.

4.4.2 Analysis 2

Another sensitivity analysis is performed. The same concept as outlined in the first two paragraphs of section 4.4 will be used. The difference will be that the outliers will be scattered around the generated data, rather than one single point being chosen to represent the outliers (see figure 4.3). In the end there are 100 datasets with 500 observations each, where 450 of these observations are generated from the DM-MPV distribution and 50 observations are the points scattered around as outliers. Therefore, 10% of the data are outliers. This tests the robustness of the pseudo-variance parameter estimates of the contaminated models.

From table 4.2, it is clear that the pseudo-variance parameter estimates of the contaminated models (DM-CMPV and DM-CPPV) are lower than those of the reparameterised distributions (DM-MPV and DM-PPV). This can also be seen in figure 4.4. The contaminated models therefore provide more robust estimates of the pseudo-variance parameter. It must be noted that the true value of the pseudo-variance of the generated data for both the contaminated models is not contained within the 90% confidence interval. Both contaminated models appear to estimate a value of the pseudo-variance that is below its true value. It is also of note that the different contaminated models were effectively equally capable of accounting for these outliers as their BIC values are very close to each other.

As with the previous sensitivity analysis, the true positive and false positive rates of identifying outliers was calculated for both contaminated models after they were fitted to every dataset. Again, for both contaminated models the true positive rate was 1 in every instance. The average false positive rate for both contaminated models was roughly 0.47. Both models struggled to identify outliers correctly.

© University of Pretoria

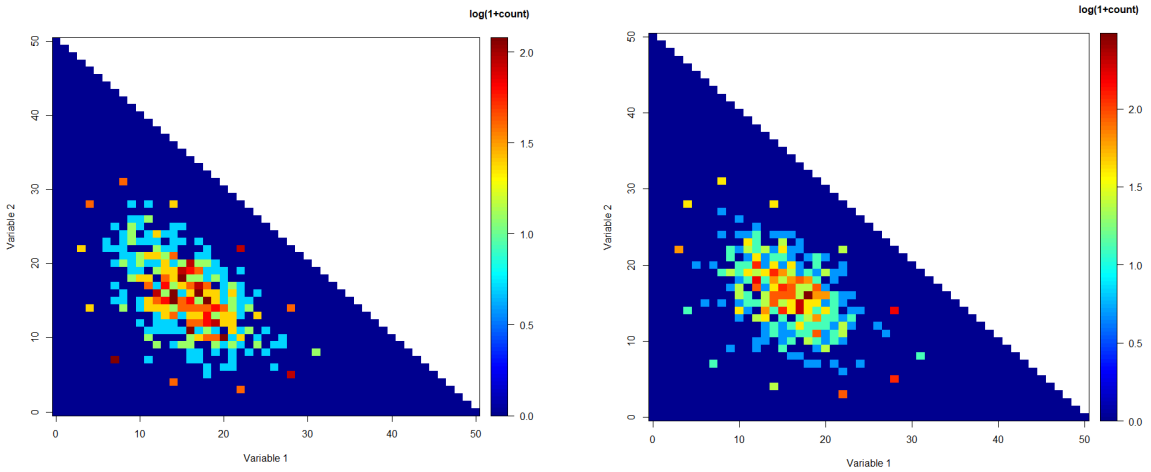


Figure 4.3: Heatmap of two generated datasets from the DM-MPV distribution (eq. (3.3.2) for the sensitivity analysis number 2.

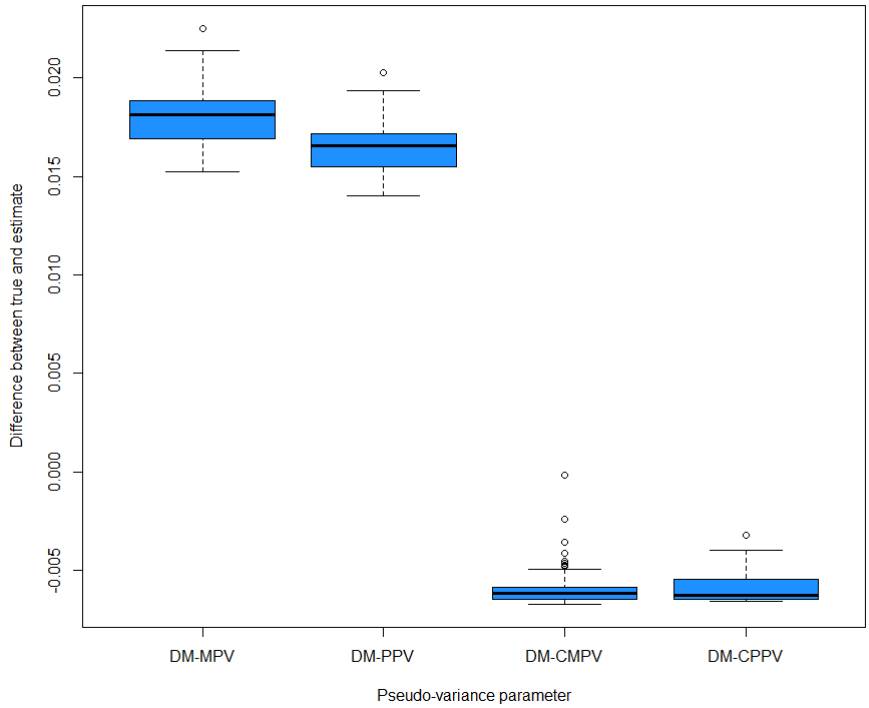


Figure 4.4: Boxplots of the difference between true and estimated value of the pseudo-variance parameter for the DM-MPV, DM-PPV, DM-CMPV, and DM-CPPV in analysis 2.

© University of Pretoria

Table 4.1: Results for the different reparameterised distributions and contaminated models fitted to the simulated data.

		Mean Estimate	Median Estimate	Lower CI (90%)	Upper CI (90%)
DM-MPV	θ_1	0.33333	0.33333	0.33333	0.33333
	θ_2	0.33333	0.33333	0.33333	0.33333
	γ	Inf	Inf	Inf	Inf
	$l(\underline{\theta})$	-3594.96109	-3594.96109	-3594.96109	-3594.96109
	AIC	7195.92217	7195.92217	7195.92217	7195.92217
	BIC	7208.56600	7208.56600	7208.56600	7208.56600
DM-CMPV	θ_1	0.33341	0.33410	0.32561	0.34009
	θ_2	0.33341	0.33377	0.32575	0.34055
	γ	0.00245	0.00255	0.00047	0.00447
	η	3391749.61629	3351713.53796	3280127.31250	3606819.32777
	π_c	0.27595	0.27539	0.26686	0.28526
	$l(\underline{\theta})$	-3076.76994	-3076.92782	-3106.09590	-3048.74429
	AIC	6163.53988	6163.85564	6101.41975	6218.38644
	BIC	6184.61292	6184.92868	6122.49279	6239.45948
DM-PPV	p_1	0.28308	0.28301	0.27655	0.28742
	p_2	0.28304	0.28313	0.27778	0.28773
	γ	0.11505	0.11510	0.11186	0.11844
	$l(\underline{\theta})$	-3366.54778	-3366.48220	-3378.69966	-3357.46453
	AIC	6739.09557	6738.96440	6719.12742	6761.36661
	BIC	6751.73939	6751.60823	6731.77124	6774.01043
DM-CPPV	p_1	0.32953	0.33025	0.32213	0.33565
	p_2	0.32958	0.33001	0.32253	0.33579
	γ	0.00365	0.00356	0.00134	0.00582
	η	177.91965	142.92257	84.02662	309.37920
	π_c	0.24772	0.24724	0.23941	0.25648
	$l(\underline{\theta})$	-3059.37528	-3059.50935	-3089.53521	-3027.98373
	AIC	6128.75056	6129.01870	6065.55827	6184.06895
	BIC	**6149.82360**	**6150.09174**	**6086.63131**	**6205.14199**

4.5 Conclusion

The contaminated DM models aid in parameter estimation as they are less influenced by the presence of outliers than the reparameterised distributions. They also provided better models of the simulated data than the uncontaminated reparameterised distributions, as seen in subsection 4.4.1 and subsection 4.4.2. In the next chapter, these contaminated models will be applied to real world data.

© University of Pretoria

Table 4.2: Results for the different reparameterised distributions and contaminated models fitted to the simulated data.

		Mean Estimate	Median Estimate	Lower CI (90%)	Upper CI (90%)
DM-MPV	θ_1	0.33345	0.33378	0.32544	0.34005
	θ_2	0.33272	0.33262	0.32580	0.33993
	γ	0.02474	0.02488	0.02224	0.02715
	$l(\underline{\theta})$	-2918.90421	-2920.18970	-2943.07743	-2898.63801
	AIC	5843.80843	5846.37941	5801.93569	5886.04654
	BIC	5856.45225	5859.02323	5814.57952	5898.69036
DM-CMPV	θ_1	0.33372	0.33412	0.32689	0.33933
	θ_2	0.33324	0.33271	0.32676	0.34044
	γ	0.00080	0.00059	0.00005	0.00210
	η	486.58120	165.53085	47.88418	2166.64335
	π_c	0.31114	0.31238	0.25567	0.37311
	$l(\underline{\theta})$	-2879.82315	-2879.50476	-2911.34568	-2854.47088
	AIC	5769.64631	5769.00952	5712.42049	5825.92891
	BIC	5790.71935	5790.08256	5733.49353	5847.00195
DM-PPV	p_1	0.33343	0.33373	0.32597	0.33961
	p_2	0.33277	0.33265	0.32631	0.33945
	γ	0.02301	0.02315	0.02081	0.02531
	$l(\underline{\theta})$	-2918.90757	-2920.18960	-2943.07741	-2898.63804
	AIC	5843.81514	5846.37920	5801.93568	5886.04718
	BIC	5856.45896	5859.02303	5814.57951	5898.69101
DM-CPPV	p_1	0.33363	0.33410	0.32732	0.33866
	p_2	0.33321	0.33279	0.32689	0.34003
	γ	0.00066	0.00033	0.00003	0.00162
	η	566.34666	221.21614	41.11442	2624.32501
	π_c	0.31638	0.31590	0.26918	0.36491
	$l(\underline{\theta})$	-2879.30671	-2879.49706	-2907.56202	-2854.55291
	AIC	5768.61341	5768.99412	5711.96368	5824.93938
	BIC	**5789.68645**	**5790.06716**	**5733.03672**	**5846.01242**

© University of Pretoria

Chapter 5

Microbiome dataset analysis

The reparameterised distributions introduced in chapter 3 and the contaminated models introduced in chapter 4 will now be applied to real world data. This is an important step to determine the value of these distributions and models. In section 5.1 the real world data that were used to test the proposed models (eq. (3.3.2), (3.5.2), (4.2.1), and (4.2.2)) is introduced and in section 5.2 the results of the proposed models when fitted to this data is given.

5.1 Human microbiome dataset analysis

In the paper of Subedi et al. (2020), finite mixtures of DM distribution regression models were used to model bacteria counts from human microbiomes from a paper by Nakatsu et al. (2015). The data included samples from healthy subjects (61 observations) and carcinoma subjects (52 observations), among others. The dataset contains counts from many different bacteria, however this will be reduced down to 6 as Subedi et al. (2020) had done in their analysis, for the same reason as Subedi et al. (2020) did, which is to display the value of the proposed models. The variables will be the counts of the 5 bacteria *Firmicutes*, *Proteobacteria*, *Bacteroidetes*, *Fusobacteria*, *Actinobacteria*, and the last variable will be the sum of the rest of the bacterial counts.

5.2 Results from the data application

Table 5.1: Results from fitting the reparameterised distributions to the microbiome data

	Healthy		Carcinoma	
	DM-MPV	DM-PPV	DM-MPV	DM-PPV
θ_1/p_1	0.64343	0.42933	0.80634	0.51876
θ_2/p_2	0.29336	0.26122	0.05238	0.12273
θ_3/p_3	0.06318	0.14551	0.14123	0.17053
θ_4/p_4	0.00000	0.03825	0.00001	0.07948
θ_5/p_5	0.00000	0.05829	0.00000	0.04439
γ	0.16708	0.17305	0.14617	0.12476
$l(\underline{p}, \gamma)$	-2304.46108	-2138.83574	-1952.29898	-1897.52500
AIC	4620.92216	4289.67147	3916.59797	3807.05000
BIC	4633.58740	4302.33672	3928.30543	3818.75746

Tables 5.1 and 5.2 show that the contaminated models outperform the reparameterised distributions in terms of AIC and BIC for both datasets considered. Table 5.1 shows that the DM-PPV distribution outperforms the DM-MPV distribution in terms of AIC and BIC for both datasets considered. Table

© University of Pretoria

Table 5.2: Results from fitting the contaminated models to the microbiome data

	Healthy		Carcinoma	
	DM-CMPV	DM-CPPV	DM-CMPV	DM-CPPV
θ_1/p_1	0.71185	0.48865	0.77158	0.56790
θ_2/p_2	0.16271	0.23182	0.06423	0.10559
θ_3/p_3	0.12541	0.14503	0.16414	0.17118
θ_4/p_4	0.00000	0.03247	0.00001	0.06741
θ_5/p_5	0.00000	0.04581	0.00000	0.03537
γ	0.05662	0.09619	0.05777	0.07005
η	48.80782	3.79774	14.67190	4.10648
π_c	0.26455	0.21487	0.24170	0.17906
$l(\underline{\theta}, \gamma)$	-2251.36417	-2124.64233	-1913.35174	-1883.84558
AIC	4518.72834	**4265.28467**	3842.70349	**3783.69115**
BIC	4535.61533	**4282.17166**	3858.31344	**3799.30110**

5.2 shows that the DM-CPPV model outperforms the DM-CMPV model in terms of AIC and BIC for both datasets considered. It must be noted that the DM-PPV distribution is a more flexible model than the DM-MPV distribution, which explains why for both datasets and both the reparameterised distributions and contaminated models the mean parameterisation outperform the mode parameterisation (see tables 5.1 and 5.2 where for both the healthy subset and cancerous subset the BIC is lower for the mean parameterisations than for the mode parameterisations). The DM-PPV distribution will therefore supply more value when applied to real world data than the DM-MPV distribution, however this does not mean that the DM-MPV distribution is without value.

After fitting the data, the probabilities that indicate a given observation being an outlier was calculated (Tomarchio et al. (2024), see eq. 5.2.1). This was done to see how many of the supposed outliers in the data were identified. Results for this are given in table 5.3, showing that there are many potential outliers in the data.

$$\begin{aligned} P_{DM-CMPV}[\underline{y}_i \; is \; outlier|\hat{\underline{\theta}}, \hat{\gamma}, \hat{\eta}, \hat{\pi}_c] &= \frac{\pi_c f_{DM-MPV}(\underline{y}_i; \hat{\underline{\theta}}, \hat{\eta}\hat{\gamma})}{f_{DM-CMPV}(\underline{y}_i; \hat{\underline{\theta}}, \hat{\gamma}, \hat{\eta}, \hat{\pi}_c)}, \\ P_{DM-CPPV}[\underline{y}_i \; is \; outlier|\hat{\underline{p}}, \hat{\gamma}, \hat{\eta}, \hat{\pi}_c] &= \frac{\pi_c f_{DM-PPV}(\underline{y}_i; \hat{\underline{p}}, \hat{\eta}\hat{\gamma})}{f_{DM-CPPV}(\underline{y}_i; \hat{\underline{p}}, \hat{\gamma}, \hat{\eta}, \hat{\pi}_c)}. \end{aligned} \tag{5.2.1}$$

Table 5.3: Number (percentage of total observations) of possible outliers

	Healthy	Carcinoma
DM-CMPV	19 (0.31)	16 (0.31)
DM-CPPV	26 (0.43)	19 (0.37)

In conclusion, the contaminated DM models are capable of accounting for outliers and providing robust estimates of the real world data. It is also clear how the altered parameterisation improves understanding of the estimates as there is now an immediate understanding of the location as well as the degree of dispersion that these parameter estimates refer to. This indicates the value of these models.

© University of Pretoria

Chapter 6

Final thoughts

In this study, we investigated extensions of the Dirichlet–multinomial (DM) distribution firstly through two reparameterised versions and then extending to two contaminated models. Each chapter built sequentially toward the ultimate aim of obtaining interpretable, robust DM models capable of handling mild outliers.

Chapter 2 presented the theoretical foundations of the DM distribution, beginning with the definition of the multinomial and Dirichlet distributions and demonstrating how the DM is constructed. The PMF, its parameter interpretation, and computational challenges were discussed in this chapter. This chapter provided the baseline model against which later reparameterised versions were compared.

Chapter 3 motivated the need for reparameterisation by identifying limitations in the interpretability of the classical parameterisation. Two alternative formulations were presented:

- The DM-MPV distribution (eq. 3.3.2) based on the modal Dirichlet distribution by Tomarchio et al. (2024).
- The DM-PPV distribution (eq. 3.5.2) using mean proportions and a pseudo-variance parameter.

A simulation study quantified bias, MSE, and information criteria, demonstrating that the new parameterisations often improve interpretability while yielding competitive or superior estimation performance.

Chapter 4 extended the reparameterised DM distributions to contaminated versions, allowing for robustness against mild outliers. Algorithms for implementation were provided, followed by a two-part sensitivity analysis. Both studies confirmed that contaminated models outperform non-contaminated models when outliers are present, yielding more stable parameter estimates and improved AIC and BIC values.

In Chapter 5 the contaminated models were applied to real microbiome count data (Subedi et al. (2020)). Results showed that the DM-CPPV model consistently achieved the lowest AIC and BIC values, suggesting that proportional mean parameters are more flexible and better suited to real-world compositional count structures. The probabilities of observations being outliers were calculated and this identified numerous potential outliers, illustrating the practical value of the contaminated DM framework.

The overarching aim of this mini-dissertation is to propose, develop, and evaluate contaminated Dirichlet–multinomial distributions that provide interpretable, robust parameter estimates and account for mild outliers.

The main contributions of this mini-dissertation are:

- Reparameterised DM distributions with improved interpretability.
- Development of contaminated DM models.

© University of Pretoria

- Simulation evidence demonstrating robustness.
- Successful application to microbiome data.

Overall, this mini-dissertation advances both theoretical and applied work on DM models, positioning contaminated reparameterised DM distributions as valuable tools for modern count data analysis. Future research could consider mixture models of this work, as well as contamination of other related count data distributions such as the generalised Dirtichlet-multinomial distribution by Bouguila (2008). An investigation of the capability of the contaminated models in addressing overdispersion would be valuable.

A summary of the distributions used to construct the models proposed in this study as well as the proposed models is given in table 6.1. This table includes differentiating factors between these models.

Table 6.1: Summary of the models and their capabilities

Model	Probability parameter	Robust to mild outliers	Restricts flexibility
multinomial	Fixed	No	Not Applicable
DM	Variable	No	No
DM-MPV	Variable	No	Yes
DM-PPV	Variable	No	No
DM-CMPV	Variable	Yes	Yes
DM-CPPV	Variable	Yes	No

The code for used in this mini-dissertation is available through the link (Code).

© University of Pretoria

Bibliography

Aitchison, J. (1982). The statistical analysis of compositional data. *Journal of the Royal Statistical Society: Series B (Methodological) 44*(2), 139–160.

Akaike, H. (1974). A new look at statistical model identification. *IEEE Transactions on Automatic Control 19*(6), 716–723.

Alhaj-Dibo, M., D. Maquin, and J. Ragot (2008). Data reconciliation: A robust approach using a contaminated distribution. *Control Engineering Practice 16*(2), 159–170.

Anderlucci, L. and C. Viroli (2020). Mixtures of Dirichlet-multinomial distriutions for supervised and unsupervised classification of short text data. *Advances in Data Analysis and Classification 14*(4), 759–770.

Bain, L. J. (1992). *Max Engelhardt*. Brooks/Cole.

Bartolucci, F., F. Pennoni, and A. Mira (2021). A multivariate statistical approach to predict COVID-19 count data with epidemiological interpretation and uncertainty quantification. *Statistics in Medicine 40*(24), 5351–5372.

Bouguila, N. (2008). Clustering of count data using generalized Dirichlet multinomial distributions. *IEEE Transactions on Knowledge and Data Engineering 20*(4), 462–474.

Chacón, J. E. (2020). The modal age of statistics. *International Review 88*(1), 122–141.

Davies, L. and U. Gather (1993). The identification of multiple outliers. *Journal of the American Statistical Association 88*(423), 782–792.

Holmes, I., K. Harris, and C. Quince (2012). Dirichlet multinomial mixtures: generative models for microbial metagenomics. *PloS One 7*(2), e30126.

Huelsenbeck, J. P. and P. Andolfatto (2007). Inference of population structure under Dirichlet process model. *Genetics 175*(4), 1787–1802.

Ishii, G. and R. Hayakawa (1960). On the compound binomial distribution. *Annals of the Institute of Statistical Mathematics 12*(1), 69–80.

Maechler, M., P. Rousseeuw, A. Struyf, M. Hubert, K. Hornik, M. Studer, P. Roudier, J. Gonzalez, K. Kozlowski, E. Schubert, and K. Murphy (2013). Package 'cluster'. *Dosegljivo na*, 980.

Mosimann, J. E. (1962). On the compound multinomial distribution, the multivariate β-distribution, and correlations among proportions. *Biometrika 49*(1/2), 65–82.

Nakatsu, G., X. Li, H. Zhou, J. Sheng, S. H. Wong, W. K. Wu, S. D. Ng, H. Tsoi, Y. Dong, N. Zhang, Y. He, Q. Kang, L. Cao, K. Wang, J. Zhang, Q. Liang, J. Yu, and J. J. Sung (2015). Gut mucosal microbiome across stages of colorectal carcinogenesis. *Nature communications 6*(1), article 8727.

© University of Pretoria

Ng, K. W., G.-L. Tian, and M.-L. Tang (2011). *Dirichlet and Related Distributions: Theory, methods and Applications*. John Wiley Sons.

Nowicka, M. and M. D. Robinson (2016). Drimseq: a Dirichlet-multinomial framework for multivariate count outcomes in genomics. *F1000Research 5*, article 1356.

Otto, A. F., J. T. Ferreira, A. Bekker, S. D. Tomarchio, and C. Tortura (2025). Modeling bounded count environmental data using a contaminated beta-binomial regression model. *arXiv preprint arXiv:2504.13665*.

Otto, A. F., J. T. Ferreira, S. D. Tomarchio, A. Bekker, and A. Punzo (2025). A contaminated regression model for count health data. *Statistical Methods in Medical Research 34*(2), 369–389.

Punzo, A. (2019). A new look at the inverse Gaussian distribution with applications to insurance and economic data. *Journal of Applied Statistics 46*(7), 1260–1287.

Punzo, A., A. Mazza, and A. Maruotti (2018). Fitting insurance and economic data with outliers: a flexible approach based on finite mixtures of contaminated gamma distributions. *Journal of Applied Statistics 45*(14), 2563–2584.

R Core Team (2024). *R: A Language and Environment for Statistical Computing*. Vienna, Austria: R Foundation for Statistical Computing. `https://www.R-project.org/`.

Schwarz, G. (1978). Estimating the dimension of a model. *Institute of Mathematical Statistics 6*(2), 461–464.

Subedi, S., D. Neish, S. Bak, and Z. Feng (2020). Cluster analysis of microbiome data by using mixtures of Dirichlet–multinomial regression models. *Journal of the Royal Statistical Society Series C: Applied Statistics 69*(5), 1163–1187.

Tomarchio, S. D. and A. Punzo (2020). Dichotomous unimodal compound models: application to the distribution of insurance losses. *Journal of Applied Statistics 47*(13-15), 2328–2353.

Tomarchio, S. D., A. Punzo, J. T. Ferreira, and A. Bekker (2024, July). A new look at the Dirichlet distribution: Robustness, clustering, and both together. *Journal of Classification 42*, 31–53.

Tvedebrink, T. (2022). Package 'dirmult'.

© University of Pretoria

Artificial Intelligence Declaration

Artificial intelligence tools, including advanced language models (e.g., ChatGPT), were utilised in the preparation of this mini-dissertation. These tools were employed solely in support of the author through grammatical amelioration and clarity, formatting recommendations, content structure and organisation, and where appropriate, as a search engine to source relevant literature and facilitate comprehension of the research topic.

All intellectual and scholarly contributions—including the development of research questions, methodological design, data analysis, interpretation of findings, and formulation of final arguments—are entirely the author's own work. No generative AI outputs were used to produce novel research results, and all content was critically reviewed and verified by the author to ensure accuracy, academic integrity, and alignment with the research objectives.

© University of Pretoria