SORT 44 (1) January-June 2020, 99-126 DOI: 10.2436/20.8080.02.96

Modelling count data using the logratio-normal-multinomial distribution

M Comas-Cufí*, J.A. Martín-Fernández, G. Mateu-Figueras and J. Palarea-Albaladejo

Abstract

The logratio-normal-multinomial distribution is a count data model resulting from compounding a multinomial distribution for the counts with a multivariate logratio-normal distribution for the multinomial event probabilities. However, the logratio-normal-multinomial probability mass function does not admit a closed form expression and, consequently, numerical approximation is required for parameter estimation. In this work, different estimation approaches are introduced and evaluated. We concluded that estimation based on a quasi-Monte Carlo Expectation-Maximisation algorithm provides the best overall results. Building on this, the performances of the Dirichlet-multinomial and logratio-normal-multinomial models are compared through a number of examples using simulated and real count data.

MSC: 62-07, 62F10, 62F86, 62P10, 62P25.

Keywords: Count data, Compound probability distribution, Dirichlet Multinomial, Logratio coordinates, Monte Carlo method, Simplex.

1 Introduction

A compound distribution of a random vector is the probability distribution resulting from assuming that its parameters are themselves random variables (Mosimann, 1962). This type of distribution plays an important role in mixture models (Lindsay, 1995) and Bayesian statistics, among others (Robbins, 1964, 1980). Practical applications are found in diverse areas such as genetics, microbiome studies, document classification and economics (Blei and Lafferty, 2007; Bouguila, 2008; Layton, and Siikamäki, 2009; Holmes, Harris and Quince, 2012; Silverman et al., 2018; Grantham et al., 2019).

Two classical distributions to model multivariate count data are the multinomial distribution and the multivariate Poisson distribution. Whilst in the first case the total number of counts per observation is a parameter, in the second it is not and it depends on the magnitude of the Poisson rates. In the literature, the multivariate Poisson distri-

* Department of Computer Science, Applied Mathematics and Statistics, University of Girona, Campus Montilivi (P4), E-17003 Girona. Email: mcomas@imae.udg.edu

Received: October 2018
Accepted: April 2020

butions has been compounded with Gamma distributions (Nelson, 1985) and with the multivariate log-normal distribution (Aitchison and Ho, 1989). In this article, we focus on distributions which are compounded with the multinomial distribution.

The Dirichlet-multinomial (DM) compound distribution (also called multivariate Pólya-Eggenberger distribution) is the most commonly used for modelling and analysing multivariate count data when they depend on a total number of trials and, unlike the ordinary multinomial distribution, some data overdispersion is present (Chapter 40, Johnson, Kotz and Balakrishnan, 1997). Let $\mathbf{X}$ be a random vector of counts. The DM distribution results from compounding a multinomial $\mathcal{M}(\mathbf{x};n,\boldsymbol{\pi}=\mathbf{p})$ for the measured vector of counts $\mathbf{X}$, with parameters n and $\boldsymbol{\pi}$ being the total number of trials and the vector of probabilities for the range of possible discrete outcomes respectively, and a Dirichlet $\mathcal{D}(\mathbf{p};\boldsymbol{\alpha})$ for the probabilities $\mathbf{p}$, with a parameter $\boldsymbol{\alpha}$. The probability mass function (pmf) of a DM distribution is $\mathcal{DM}(\mathbf{x};n,\boldsymbol{\alpha}) = \Pr(X=\mathbf{x};n,\boldsymbol{\alpha}) = \int_{\mathcal{S}^D} \mathcal{D}(\mathbf{p};\boldsymbol{\alpha})\,\mathcal{M}(\mathbf{x};n,\mathbf{p})\,\mathrm{d}\mathbf{p}$, where $\mathcal{S}^D$ refers to the unit simplex. The unit simplex is the sample space of random vectors $\mathbf{p}$ of length D consisting of strictly positive components adding up to one (Aitchison, 1986), i.e.

$$\mathcal{S}^D = \left\{ \mathbf{p} = (p_1,\ldots,p_D) \in \mathbb{R}^D \;\middle|\; p_k > 0 \text{ and } \sum_{k=1}^{D} p_k = 1 \right\}.$$

The closed form expression for the DM pmf is

$$\mathcal{DM}(\mathbf{x};n,\boldsymbol{\alpha}) = \frac{n!\,\Gamma\left(\sum_{k=1}^{D}\alpha_k\right)}{\Gamma\left(n+\sum_{k=1}^{D}\alpha_k\right)} \prod_{k=1}^{D} \frac{\Gamma(x_k+\alpha_k)}{x_k!\,\Gamma(\alpha_k)},$$

where $\Gamma(\cdot)$ is the well-known gamma function. The DM distribution is defined on the sample space of random count vectors. That is, the $\{n,D\}$-*simplex lattice*

$$\mathcal{S}^{(n,D)} = \left\{ \mathbf{x} = (x_1,\ldots,x_D) \;\middle|\; x_k \in \{0,1,\ldots,n\} \text{ and } \sum_{k=1}^{D} x_k = n \right\}, \tag{1}$$

consisting of random vectors of counts with components in the non-negative integer domain and sum equal to n (Scheffé, 1958). As stressed in Aitchison (1986) and Comas-Cufí, Martín-Fernández and Mateu-Figueras (2016), the Dirichlet distribution imposes a very strong independence structure: any pair of ratios between different components of $\mathbf{p}$ are assumed to be statistically independent. This heavily restricts its potential for data modelling when the analysis is based in ratios between parts, as it is the case of compositional data analysis (Comas-Cufí et al., 2016). Some generalisations of the Dirichlet have been proposed to overcome this difficulty with limited success (Connor and Mosimann, 1969; Minka, 2004; Ongaro and Migliorati, 2013).

In the 1980's John Aitchison introduced the compositional approach to model and analyse multivariate random vectors defined on the simplex (Aitchison, 1986). A number of methodological and practical contributions have been recently published in differ-

ent areas such as statistics Comas-Cufí, Martín-Fernández and Mateu-Figueras (2019), waste management (Edjabou et al., 2017), health (Chastin et al., 2015) and animal science (Palarea-Albaladejo et al., 2017).

We focus here on the logratio-normal-multinomial distribution resulting from compounding a multinomial distribution for the vector of counts $\mathbf{X}$ with a logratio-normal distribution for the corresponding vector of multinomial probabilities $\mathbf{p}$. The logratio-normal distribution was introduced in Mateu-Figueras, Pawlowsky-Glahn and Egozcue (2013) to model compositions. Also known as the normal distribution on the simplex (we denote it by $\mathcal{N}_{\mathcal{S}^D}$), it is defined using the ordinary multivariate normal probability density function (pdf) over a vector of orthonormal logratio coordinates (Egozcue et al., 2003) as follows:

$$\mathcal{N}_{\mathcal{S}^D}(\mathbf{p};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{h};\boldsymbol{\mu},\boldsymbol{\Sigma}) = (2\pi)^{-\frac{D-1}{2}}|\boldsymbol{\Sigma}|^{-\frac{1}{2}}\exp\left(-\frac{1}{2}(\mathbf{h}-\boldsymbol{\mu})'\boldsymbol{\Sigma}^{-1}(\mathbf{h}-\boldsymbol{\mu})\right), \quad (2)$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are the usual expectation and covariance parameters, and $\mathbf{h} = (h_1,\ldots,h_{D-1})$ are orthonormal logratio coordinates of a composition $\mathbf{p}$ defined on $\mathcal{S}^D$ with respect to a predefined orthonormal basis of the simplex, see (Egozcue et al., 2003). Although the logratio-normal is a reparametrisation of the logistic-normal distribution (Aitchison and Shen, 1980), its definition avoids using the logistic transformation in order to focus on the appropriate reference measure (see Mateu-Figueras et al. (2013) for details). In our developments, we will use so-called isometric logratio (ilr) coordinates obtained from a particular choice of orthonormal basis as introduced in Egozcue et al. (2003). Namely, $\mathbf{h} = \text{ilr}(\mathbf{p})$ with elements

$$h_i = \sqrt{\frac{i}{i+1}}\ln\frac{\sqrt[i]{\prod_{j=1}^{i}p_j}}{p_{i+1}}, i = 1,\ldots,D-1. \quad (3)$$

Note that the composition $\mathbf{p}$ associated with orthonormal logratio coordinates $\mathbf{h}$ is obtained by inverse transformation $\mathbf{p} = \text{ilr}^{-1}(\mathbf{h})$. Importantly, using this particular ilr representation does not imply a lack of generality, since the results are invariant under change of orthonormal basis. This is because different orthonormal logratio coordinate systems are orthogonal rotations from one to another. In Billheimer, Guttorp and Fagan (2001) the multivariate logistic-normal-multinomial distribution was defined for modelling multinomial counts by compounding the multinomial distribution with the additive-logistic-normal distribution (Chapter 6, Aitchison (1986)). Practical applications of this model can be found in (Xia et al., 2013; Silverman et al., 2018) for microbiome data, or Hughes, Munkvold and Samita (1998) where the additive-logistic-normal was combined with the binomial distribution to model two-part compositions. In the following, we refer to the distribution obtained by composing the logratio-normal and the multinomial distribution as the logratio-normal-multinomial distribution (referred to as LNM in the following). From a probabilistic point of view, the logratio-normal-multinomial and the logistic-normal-multinomial models define the same law of prob-

ability. Nevertheless, we have decided to call it logratio-normal-multinomial in order to emphasize that, instead of using the logistic transformation, we use the orthonormal logratio coordinates together with the reference measure compatible with the algebraic-geometric structure of the simplex and with the compositional approach introduced by (Aitchison, 1986).

Markov Chain Monte Carlo (MCMC) methods have been used so far for parameter estimation with these models. For the case of multivariate logistic-normal-multinomial distribution see (Billheimer et al., 2001) and (Xia et al., 2013). Quasi-Monte Carlo methods (QMC) are well-known tools used to approximate high-dimensional integrals (Morokoff and Caflisch, 1995; Leobacher and Pillichshammer, 2014). They deviate from standard Monte Carlo in the type of sampling procedure used to approximate the integral. While classic Monte Carlo uses pseudo-random samples, QMC methods use quasi-random samples or low-discrepancy sequences. QMC methods have been successfully used in different parameter estimation scenarios (Drmota and Tichy, 1997; Pan and Thompson, 2007; Kuo et al., 2008) and have shown an improvement of the efficiency when embedded in an Expectation-Maximisation (EM) algorithm (Jank, 2005).

Building on these results, we propose more efficient tools to estimate the parameters of a LNM distribution. Their performance in modelling count data is compared with the DM distribution.

The work is organised as follows. In Section 2, some basic definitions are formally introduced. In Section 3, we derive the E and the M steps of an EM scheme for parameter estimation. We propose to combine QMC integration with the EM algorithm to estimate the parameters of the LNM distribution. Section 4 illustrates the use of DM and LNM distributions in three different examples based on simulated and real data. Lastly, Section 5 concludes with some final remarks.

All data analyses discussed in this work were conducted using the R statistical programming environment (R Development Core Team, 2015). Computer routines implementing the methods and the data sets can be obtained at https://github.com/mcomas/SORT-normal-multinomial.

2 Basic definitions and properties

The simplex $\mathcal{S}^D$ has an Euclidean vector space structure of dimension $D-1$ with its own basic operations (perturbation and powering), an inner product and a distance (so-called Aitchison distance) (Egozcue et al., 2003). According to this algebraic-geometric characterisation, compositions can be mapped onto the ordinary real space using logratio coordinates. The logratio-normal distribution is a model which is closed under the main operations in the simplex $\mathcal{S}^D$ (Mateu-Figueras et al., 2013). Also, it is a flexible distribution because it can model compositions whose components have different forms of dependence. Importantly, the density function (2) is defined with respect to what is called the Aitchison measure on the simplex, a probability measure different

from the ordinary Lebesgue measure on real space (Pawlowsky-Glahn and Egozcue, 2001; Mateu-Figueras, Pawlowsky-Glahn and Egozcue, 2011). The Aitchison measure is a natural measure on the simplex, compatible with its vector space structure and absolutely continuous with respect to the Lebesgue measure in the space of logratio coordinates (Mateu-Figueras et al., 2013).

As said above, the LNM is the distribution resulting from compounding the multinomial distribution $\mathcal{M}(\mathbf{x};n,\boldsymbol{\pi}=\mathbf{p})$ with the logratio-normal distribution $\mathcal{N}_{\mathcal{S}^D}(\mathbf{p};\boldsymbol{\mu},\boldsymbol{\Sigma})$. a random vector of counts $\mathbf{X}$ generated from a LNM distribution with parameters n, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ is obtained in two steps:

- Firstly, a random composition $\mathbf{p}$ is generated using the logratio-normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$.
- Secondly, a count random vector $\mathbf{X}$ is generated using the multinomial distribution with parameters n and $\boldsymbol{\pi}=\mathbf{p}$.

The pmf of a LNM distributions, expressed in terms of orthonormal logratio coordinates, is

$$\mathcal{LNM}(\mathbf{x};n,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \Pr(X=\mathbf{x};n,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \int_{\mathcal{S}^D} \mathcal{N}_{\mathcal{S}^D}(\mathbf{p};\boldsymbol{\mu},\boldsymbol{\Sigma})\mathcal{M}(\mathbf{x};n,\mathbf{p})\,\mathrm{d}_{\mathcal{A}}\mathbf{p} \tag{4}$$

$$= \int_{\mathbb{R}^{D-1}} \mathcal{N}(\mathbf{h};\boldsymbol{\mu},\boldsymbol{\Sigma})\,\frac{n!}{\prod_{k=1}^{D} x_k!}\,\prod_{k=1}^{D} \mathrm{ilr}_k^{-1}(\mathbf{h})^{x_k}\,\mathrm{d}\mathbf{h}, \tag{5}$$

where $\mathrm{ilr}_k^{-1}(\mathbf{h})$ stands for the k-th component of the composition $\mathbf{p}=\mathrm{ilr}^{-1}(\mathbf{h})$. Note that expression (4) is written with respect to the Aitchison measure, while expression (5) is written with respect to the Lebesgue measure in the logratio coordinate space. The LNM distribution is defined on $\mathcal{S}^{(n,D)}$ (Eq. 1).

Note that definition of the LNM distribution is similar to the definition of the DM distribution. The difference is that in the former the composition $\mathbf{p}$ is modelled by a normal distribution in terms of ilr-coordinates, instead of using a Dirichlet distribution.

Using the pmf given in (4) or (5), the following properties can be easily derived.

Property 1 *For a fixed $\boldsymbol{x}$ we have*

$$\lim_{\|\boldsymbol{\Sigma}\|\to 0} \mathcal{LNM}(\boldsymbol{x};n,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{M}\left(\boldsymbol{x};n,\boldsymbol{\pi}=\mathrm{ilr}^{-1}(\boldsymbol{\mu})\right).^{1}$$

1. $\lim_{\|\boldsymbol{\Sigma}\|\to 0}$ stands for any sequence of covariance matrices such that their highest eigenvalue goes to 0.

Proof. See Appendix A. ■

Property 2 *Let* $\boldsymbol{x} = (x_1, \ldots, x_D)$ *and* $x_1 + \cdots + x_D = n$. *If* $\lim_{n\to\infty} \frac{x_i}{n} = \pi_i$ *and* $\pi_i > 0$ *for* $1 \leq i \leq D$, *then*

$$\lim_{n\to\infty} n^{D-1} \cdot \mathcal{LNM}(\boldsymbol{x}; n, \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}_{\mathcal{S}^D}(\boldsymbol{\pi}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{1}{\sqrt{D}} \frac{1}{\pi_1 \ldots \pi_D}$$

Proof. See Appendix A. ■

Hence, depending on the parameters n and $\boldsymbol{\Sigma}$, the LNM distribution can be approximated by either a multinomial distribution or a logratio-normal distribution. The first property suggests that for count data sets where the random vector $\mathbf{p}$ has low variability, modelling based on either the LNM or multinomial distributions will provide similar results. The second property implies that a LNM distribution with large values of the number of trials n converges to the $\mathcal{N}_{\mathcal{S}^D}$ distribution. That is, for large n, the distribution of the random count vectors $\mathbf{X}$ on the simplex lattice $\mathcal{S}^{(n,D)}$ will be very similar to the distribution of the random vector $\mathbf{p}$ on the simplex $\mathcal{S}^D$.

3 Monte Carlo EM algorithm for logratio-normal-multinomial parameter estimation

Let $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ be an independent and identically distributed sample of multivariate count data, with N denoting the sample size. To estimate the LNM parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ it is necessary to maximise the likelihood of the observed data given by

$$L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}) = \prod_{i=1}^{N} \Pr(X = \mathbf{x}_i; n, \boldsymbol{\mu}, \boldsymbol{\Sigma}). \tag{6}$$

Denoting by $\mathbf{H}$ the non-observed ilr-coordinates, i.e. $\mathbf{H} = \{\mathbf{h}_1, \ldots, \mathbf{h}_N\}$, the EM algorithm (Dempster, Laird and Rubin, 1977) allows to maximise (6) by iteratively using an expected *augmented* likelihood $L(\boldsymbol{\mu}, \boldsymbol{\Sigma}; \mathbf{X}, \mathbf{H}) = \prod_{i=1}^{N} f(\mathbf{x}_i, \mathbf{h}_i; \boldsymbol{\mu}, \boldsymbol{\Sigma})$, where the joint probability density function of random vectors X and H is

$$f(\mathbf{x}, \mathbf{h}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) = \mathcal{N}(\mathbf{h}; \boldsymbol{\mu}, \boldsymbol{\Sigma}) \frac{(\sum_{k=1}^{D} x_k)!}{\prod_{k=1}^{D} x_k!} \prod_{k=1}^{D} \mathrm{ilr}_k^{-1}(\mathbf{h})^{x_k}.$$

In the E step, the expected value at the $(s+1)$-th iteration of the algorithm is calculated using the expression

$$Q(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}) = \mathbb{E}_{H\mid X;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}\left[\ln L(\boldsymbol{\mu},\boldsymbol{\Sigma};\mathbf{X},\mathbf{H})\right] = \mathbb{E}_{H\mid X;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}\left[\sum_{i=1}^{N}\ln\left(f(\mathbf{x}_i,\mathbf{h}_i;\boldsymbol{\mu},\boldsymbol{\Sigma})\right)\right],$$

for a random vector H conditioned to $\mathbf{X}$ with parameters $\boldsymbol{\mu}^{(s)}$ and $\boldsymbol{\Sigma}^{(s)}$ obtained in the previous iteration. In the M step, the function $Q(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)})$ is maximised with respect to the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$. By expanding $Q(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)})$, it holds that

$$\begin{aligned} Q(\boldsymbol{\mu},\boldsymbol{\Sigma} \mid \boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}) &= \mathbb{E}_{H\mid X;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}\left[\sum_{i=1}^{N}\ln\left(\mathcal{N}(\mathbf{h}_i;\boldsymbol{\mu},\boldsymbol{\Sigma})\,\frac{(\sum_{k=1}^{D}x_{ik})!}{\prod_{k=1}^{D}x_{ik}!}\prod_{k=1}^{D}\mathrm{ilr}_k^{-1}(\mathbf{h}_i)^{x_{ik}}\right)\right] \\ &= \sum_{i=1}^{N}\left\{\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}\left[\ln\left(\mathcal{N}(\mathbf{h}_i;\boldsymbol{\mu},\boldsymbol{\Sigma})\right)\right]\right\} + \\ &+ \sum_{i=1}^{N}\left\{\ln\left(\frac{(\sum_{k=1}^{D}x_{ik})!}{\prod_{k=1}^{D}x_{ik}!}\right) + \sum_{k=1}^{D}\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}\left[\ln\left(\mathrm{ilr}_k^{-1}(\mathbf{h}_i)^{x_{ik}}\right)\right]\right\}. \end{aligned}$$

To optimise the function Q with respect to the parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, it is only necessary to optimise the terms where $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ are involved. That is, the term $Q^*(\boldsymbol{\mu},\boldsymbol{\Sigma}|\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}) = \sum_{i=1}^{N}\left\{\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}\left[\ln\left(\mathcal{N}(\mathbf{h}_i;\boldsymbol{\mu},\boldsymbol{\Sigma})\right)\right]\right\}$. Using the linearity of the expectation $\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}$, this term is maximised at the basic statistics $\boldsymbol{\mu}^{(s+1)} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}[\mathbf{h}_i]$ and $\boldsymbol{\Sigma}^{(s+1)} = \frac{1}{N}\sum_{i=1}^{N}\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}[\mathbf{h}_i^{\intercal}\mathbf{h}_i] - \hat{\boldsymbol{\mu}}^{\intercal}\hat{\boldsymbol{\mu}}$. In consequence, the critical point when applying the EM algorithm here is the calculation of the expected values $\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}[\mathbf{h}_i]$ and $\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}[\mathbf{h}_i^{\intercal}\mathbf{h}_i]$.

3.1 Quasi-Monte Carlo approximation to the E step

The expected values $\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}[\mathbf{h}_i]$ and $\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}[\mathbf{h}_i^{\intercal}\mathbf{h}_i]$ are calculated using Monte Carlo approximation, then turning the EM algorithm into a Monte Carlo EM algorithm (Jank, 2005; Neath, 2013). To simplify the exposition, in this subsection we denote the expected value $\mathbb{E}_{\mathbf{h}_i\mid\mathbf{x}_i;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}$ by simply $\mathbb{E}$. The vector $\mathbb{E}[\mathbf{h}] = (\mathbb{E}[h_k])_{k=1,\ldots,D-1}$ and the matrix $\mathbb{E}[\mathbf{h}^{\intercal}\mathbf{h}] = (\mathbb{E}[h_k h_r])_{k,r=1,\ldots,D-1}$ are particular cases of the general expression

$$\mathbb{E}[\varphi(\mathbf{h})] = \int_{\mathbb{R}^{D-1}}\varphi(\mathbf{h})\,f(\mathbf{h}\mid\mathbf{x};\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)})\,\mathrm{d}\mathbf{h}, \tag{7}$$

where $\varphi:\mathbb{R}^{D-1}\to\mathbb{R}$ and $f(\mathbf{h}\mid\mathbf{x};\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}) = \frac{f(\mathbf{x},\mathbf{h};\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)})}{\Pr\left(\{X=\mathbf{x}\};n,\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}\right)}$. Moreover, note that $\Pr\left(X=\mathbf{x};n,\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}\right) = \int_{\mathbb{R}^{D-1}}f(\mathbf{x},\mathbf{h};\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)})\,\mathrm{d}\mathbf{h}$. Hence, to evaluate (7), we need to approximate the integral

$$I\left(\varphi, \mathbf{x}, \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)}\right) = \int_{\mathbb{R}^{D-1}} \varphi(\mathbf{h})\, f(\mathbf{x}, \mathbf{h}; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})\, \mathrm{d}\mathbf{h} \tag{8}$$

for different functions φ.

In Xia et al. (2013), a MCMC method based on the Metropolis algorithm is used to estimate $\mathbb{E}[\mathbf{h}]$. Although the authors approximate the second moment, $\mathbb{E}[\mathbf{h}^\intercal\mathbf{h}]$, with the square of the first moment, $\mathbb{E}[\mathbf{h}]^\intercal \cdot \mathbb{E}[\mathbf{h}]$, we here estimate the first and the second moments, i.e. $\varphi(\mathbf{h}) = \mathbf{h}$ and $\varphi(\mathbf{h}) = \mathbf{h}^\intercal\mathbf{h}$, separately in the E step. To approximate $I\left(\varphi, \mathbf{x}, \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)}\right)$, we used Monte Carlo integration with importance sampling (Caflisch, 1998). In each E step, importance sampling is performed using a normal distribution centred at $\mathbf{m} = \mathbb{E}[\mathbf{h}]$ with covariance $\mathbf{S} = \mathbb{E}[(\mathbf{h}-\mathbf{m})^\intercal(\mathbf{h}-\mathbf{m})]$ calculated in the previous E step. The integral $I\left(\varphi, \mathbf{x}, \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)}\right)$ is approximated by

$$\begin{aligned} I\left(\varphi, \mathbf{x}, \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)}\right) &= \int_{\mathbf{h}\in\mathbb{R}^{D-1}} \frac{\varphi(\mathbf{h})\, f(\mathbf{x}, \mathbf{h}; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})}{\mathcal{N}(\mathbf{h}; \mathbf{m}, \mathbf{S})} \mathcal{N}(\mathbf{h}; \mathbf{m}, \mathbf{S}) \mathrm{d}\mathbf{h} \\ &= \mathbb{E}_{\mathbf{W}\sim\mathcal{N}(\mathbf{m},\mathbf{S})}\left[\frac{\varphi(\mathbf{w})\, f(\mathbf{x}, \mathbf{w}; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})}{\mathcal{N}(\mathbf{w}; \mathbf{m}, \mathbf{S})}\right] \\ &\approx \frac{1}{M}\sum_{r=1}^{M} \frac{\varphi(\mathbf{w}_r)\, f(\mathbf{x}, \mathbf{w}_r; \boldsymbol{\mu}^{(s)}, \boldsymbol{\Sigma}^{(s)})}{\mathcal{N}(\mathbf{w}_r; \mathbf{m}, \mathbf{S})}\,, \end{aligned} \tag{9}$$

where the values $\mathbf{w}_r$, $r = 1, \ldots, M$, are sampled from a normal distribution $\mathcal{N}(\mathbf{m}, \mathbf{S})$.

We here adopt a QMC approach which, instead of using pseudo-random normal generators, employs low-discrepancy sequences to generate the random values $\mathbf{w}_r$ (Caflisch, 1998; Wang and Fang, 2003; Leobacher and Pillichshammer, 2014). A low-discrepancy sequence is an equidistributed sample defined on a particular domain that is generated at a low computational cost (Chapter 2, Leobacher and Pillichshammer (2014)). Although different low-discrepancy sequences exist, we only considered Halton and Sobol sequences (Chapter 1, Drmota and Tichy (1997)). To choose between them we followed Morokoff and Caflisch (1995), which suggests best performance of Sobol sequences when the dimension of $\mathbf{h}$ is higher than six. Halton sequences are instead recommended for lower dimensions. QMC methods have shown to improve efficiency when combined with an EM algorithm (Jank, 2005).

Appendices B and C include a comparative of the performance of different methods in a univariate but extreme case and on a number of multidimensional cases respectively. In these scenarios, the best approximations for the first and second moments were obtained using the QMC approach. By contrast, methods based on MCMC algorithms showed the worst performance and highest computing time.

4 Examples

In this section, we consider three different contexts where N multinomial observations $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ are generated from N probability vectors $\mathbf{p}_1, \ldots, \mathbf{p}_N$. In the first scenario we consider a pure multinomial process (i.e. $\mathbf{p}_1 = \cdots = \mathbf{p}_N$). In the second scenario we consider each $\mathbf{p}_i$ in Hardy-Weinberg equilibrium (i.e. $p_{i2} = 4p_{i1}p_{i3}$, which implies a correlation of *minus one* between the logratios $\ln(\frac{p_{\cdot 1}}{p_{\cdot 2}})$ and $\ln(\frac{p_{\cdot 3}}{p_{\cdot 2}})$). The last case is a real scenario where no implicit relation between $\mathbf{p}_i$'s is assumed. In these three scenarios we compare the ability to model the sample $\mathbf{X}$ using both DM and LNM compound distributions. The main aim is to investigate how probability vectors $\mathbf{p}_i$ are modelled using the expected posterior probabilities calculated using distributions DM and LNM, i.e. $\hat{\mathbf{p}}_{i,\mathrm{DM}} = \mathbb{E}_{P|X=\mathbf{x}_i;\hat{\boldsymbol{\alpha}}}[P]$ and $\hat{\mathbf{p}}_{i,\mathrm{LNM}} = \mathrm{ilr}^{-1}\left(\mathbb{E}_{H|X=\mathbf{x}_i;\hat{\boldsymbol{\mu}},\hat{\boldsymbol{\Sigma}}}[H]\right)$ respectively.

For the LNM distribution we considered two different possibilities as starting point for the EM algorithm (SP1 and SP2 below):

- SP1: Given model parameters $\boldsymbol{\mu}_t^*$ and $\boldsymbol{\Sigma}_t^*$ evaluated at iteration t and observation $\mathbf{x}$, the maximum $\mathbf{h}^*$ of $f(\mathbf{h} \mid \mathbf{x}; \boldsymbol{\mu}_t^*, \boldsymbol{\Sigma}_t^*)$ can be easily calculated. Thus, the following iterative algorithm was defined:

  1. Set $t = 0$ and initiate $\boldsymbol{\mu}_0^*$ and $\boldsymbol{\Sigma}_0^*$ using sample mean and identity matrix in logratio coordinates.
  2. For each $\mathbf{x}_i \in \mathbf{X}$ calculate $\mathbf{h}_i^*$ maximising $f(\mathbf{h} \mid \mathbf{x}_i; \boldsymbol{\mu}_t^*, \boldsymbol{\Sigma}_t^*)$.
  3. Set $\boldsymbol{\mu}_{t+1}^* = \frac{1}{N}\sum \mathbf{h}_i^*$ and $\boldsymbol{\Sigma}_{t+1}^* = \frac{1}{N}\sum (\mathbf{h}_i^* - \boldsymbol{\mu}_{t+1}^*)^\intercal (\mathbf{h}_i^* - \boldsymbol{\mu}_{t+1}^*)$.
  4. If $\|\boldsymbol{\mu}_{t+1}^* - \boldsymbol{\mu}_t^*\|_\infty > 0.001$ go to step 2. Otherwise, stop and set $\boldsymbol{\mu}_0 = \boldsymbol{\mu}_{t+1}^*$ and $\boldsymbol{\Sigma}_0 = \boldsymbol{\Sigma}_{t+1}^*$.

- SP2: Set $\boldsymbol{\mu}_0 = \overline{\mathrm{ilr}(\mathbf{p}_{\mathrm{DM}})}$ and $\boldsymbol{\Sigma}_0 = \mathrm{Cov}[\mathrm{ilr}(\mathbf{p}_{\mathrm{DM}})]$. That is, LNM estimation started using the final estimates of the mean vector and covariance matrix in logratio coordinates obtained from the DM compound distribution.

To estimate the parameters of the LNM distribution, we iterated the EM algorithm until the distance between two consecutive estimates was lower than a certain tolerance value $\tau = 0.001$. The expected values in (7) were approximated using 10000 iterations of a QMC simulation scheme based on Sobol sequences for the first and third example and Halton sequences for the second example.

4.1 A pure multinomial process

A sample $\mathbf{X} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$ was generated, with $\mathbf{x}_i \sim \mathcal{M}(\mathbf{x}; n, \boldsymbol{\pi} = \mathbf{p})$ using a unique probability vector $\mathbf{p}$. Following Martín-Fernández et al. (2015), we designed nine different settings for $\mathbf{p}$ (see details in Appendix D). For each one we considered an increasing number of multinomial trials $n \in \{50, 100, 200, 500\}$, resulting in 9×4 different scenar-

ios. For each scenario we generated 10 different replicates of $\mathbf{X}$ with $N = 1000$ count vectors from the corresponding multinomial model. In total, we used $9 \times 4 \times 10 = 360$ samples of size 1000. For each of the replicates $\mathbf{X}$, we calculated the expected value $\hat{\mathbf{p}}_{i,\text{DM}}$ and $\hat{\mathbf{p}}_{i,\text{LNM}}$. It is reasonable that for any vector of counts $\mathbf{x}_i$, expected values $\hat{\mathbf{p}}_{i\cdot}$ will be close to $\mathbf{p}$. To evaluate how close estimations $\hat{\mathbf{p}}_{i\cdot}$ were to $\mathbf{p}$ we computed the mean of the Aitchison distances, $\text{dist}_{\mathcal{A}}$, between them; or, equivalently, the mean of the Euclidean distances, $\text{dist}_{\mathcal{E}}$, between the corresponding ilr coordinates (Egozcue et al., 2003):

$$\frac{1}{1000}\sum_{i=1}^{1000}\text{dist}_{\mathcal{A}}\left(\mathbf{p},\hat{\mathbf{p}}_{i\cdot}\right) = \frac{1}{1000}\sum_{i=1}^{1000}\text{dist}_{\mathcal{E}}\left(\text{ilr}(\mathbf{p}),\text{ilr}(\hat{\mathbf{p}}_{i\cdot})\right). \tag{10}$$

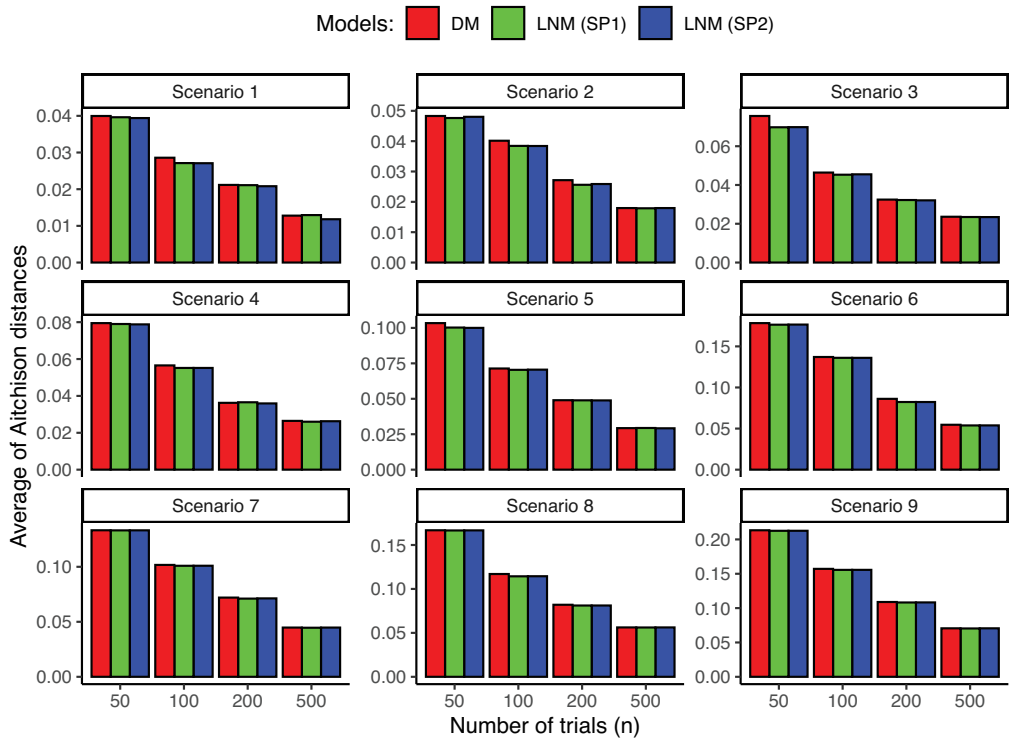

Figure 1: *Pure multinomial process: average of Aitchison distances between true* $\mathbf{p}$ *and estimated* $\hat{\mathbf{p}}_{i\cdot}$ *values using models DM (red bars), LNM with starting point SP1 (green bars) and LNM with starting point SP2 (blue bars). Nine different scenarios were set up considering four different number of trials in each one (n from 50 to 500). See Appendix D for details.*

This value was afterwards averaged across the 10 replicates. Figure 1 shows the results for the nine scenarios. As expected, DM and LNM produced similar results when modelling probabilities $\mathbf{p}$. Note that, in this example considering a multinomial setting, the estimate for $\boldsymbol{\Sigma}$ is close to the zero matrix (Property 1). Because of this, in some

scenarios after initialisations SP1 and SP2, the covariance matrix $\mathbf{\Sigma}_0$ was close to degenerate and the EM algorithm stopped in the first iteration (see Table *2*). In general, when the number of trials increases the error decreases as expected.

4.2 The Hardy-Weinberg equilibrium

In this case the variability of the unobserved vectors of probabilities $\mathbf{p}_1,\ldots,\mathbf{p}_N$ is governed by the Hardy-Weinberg equilibrium (Graffelman, and Weir, 2016). In brief, a biallelic genetic marker with alleles A and B with respective frequencies q and $(1-q)$ is in Hardy-Weinberg equilibrium if the genotype frequencies $\mathbf{x} = (f_{\mathrm{AA}}, f_{\mathrm{AB}}, f_{\mathrm{BB}})$ are given by $\mathbf{p} = (q^2, 2q(1-q), (1-q)^2)$. To obtain our sample $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_N\}$, we generated N uniform random variables q_i, taking values between 0 and 1, in six different scenarios:

1. $q_i \sim Unif(0,1)$,
2. $q_i \sim Unif(0,0.5)$,
3. $q_i \sim Unif(0,0.25)$,
4. $q_i \sim Unif(0.25,0.5)$,
5. $\mathrm{logit}(q_i) \sim Norm(0,1)$, and
6. $\mathrm{logit}(q_i) \sim Norm(1,1)$.

The probabilities $\mathbf{p}_i$ were calculated using q_i according to the Hardy-Weinberg equilibrium. Observations $\mathbf{x}_i$ were drawn from a multinomial distribution with parameters $n \in \{50,100,200,500\}$, and $\pi = \mathbf{p}_i$ (Graffelman, 2015). After fitting DM and LNM models to sample $\mathbf{X}$, we computed the expected probability vector of probabilities $\hat{\mathbf{p}}_{i,\mathrm{DM}}$ and $\hat{\mathbf{p}}_{i,\mathrm{LNM}}$ using DM and LNM models respectively.

We compared estimations $\hat{\mathbf{p}}_{i\cdot}$ to vector $\mathbf{p}_i = (q_i^2, 2q_i(1-q_i), (1-q_i)^2)$ by using the average of Aitchison distances (10) as in the previous example. The results are displayed in Figure 2. Unlike with the pure multinomial process, there is a linear relation between the three parts of the composition. Consequently, the probabilities $\mathbf{p}_i$ could be better approximated in all cases using LNM instead of DM, with negligible differences for different starting points SP1 or SP2. Again, the error decreases with increasing number of trials as expected.

To illustrate the performance of the models in presence of variability in the probability vectors $\mathbf{p}_i$, we used ternary diagrams to graphically represent the first 50 simulated allele genotype probability vectors (Figure 3) and their genotype frequency (Figure 4).

Figure 3 (left) shows probability vectors $\{\mathbf{p}_1\ldots,\mathbf{p}_{50}\}$ satisfying the Hardy-Weinberg equilibrium which were generated from the first scenario above ($q \sim Unif(0,1)$). Note that they exactly fit a (compositional) line described by the parametric equations $\{(t^2, 2t(1-t), (1-t)^2) : 0 < t < 1\}$. Figure 3 (centre) shows that estimates $\{\hat{\mathbf{p}}_{1,\mathrm{DM}},\ldots,\hat{\mathbf{p}}_{50,\mathrm{DM}}\}$ from the DM model are far more scattered with respect to the equilibrium state than those from the LNM model $\{\hat{\mathbf{p}}_{1,\mathrm{LNM}},\ldots,\hat{\mathbf{p}}_{50,\mathrm{LNM}}\}$ (Figure 3 (right)). Taht is, while the LNM model is able to capture the variability along the compositional line the DM is

not.

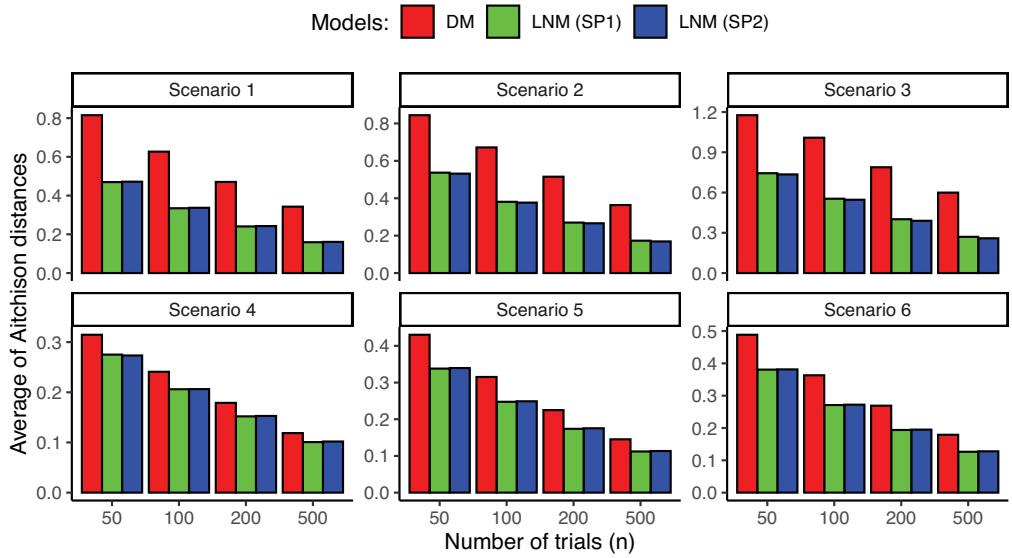

Figure 2: *Hardy-Weinberg equilibrium: average of Aitchison distances between true* $\boldsymbol{\pi}$ *and estimated* $\boldsymbol{p}_i$ *values using models DM (red bars), LNM with starting point SP1 (green bars) and LNM with starting point SP2 (blue bars). Six different scenarios were set up considering four different numbers of trials in each one (n from 50 to 500).*

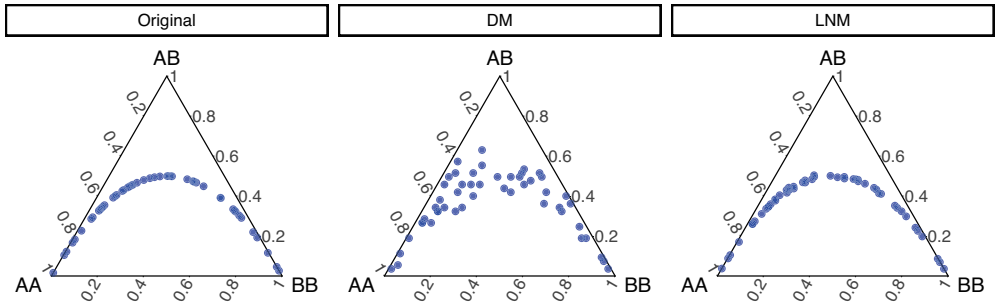

Figure 3: *Probabilities* $\boldsymbol{p}_i$ *in the Hardy-Weinberg equilibrium. Original probabilities* $\boldsymbol{p}_i$ *distributed according to Scenario 1 (left), estimates* $\hat{\boldsymbol{p}}_{i,DM}$ *using the Dirichlet-multinomial model (centre), estimates* $\hat{\boldsymbol{p}}_{i,LNM}$ *using the logratio-normal-multinomial model (right).*

This behaviour is made more evident when samples of count data are generated with both models. Figure 4 (left) shows genotype frequency data vectors $\{\mathbf{x}_1 \ldots, \mathbf{x}_{50}\}$ generated from the Hardy-Weinberg equilibrium using a multinomial distribution (Graffelman, 2015). In the centre we can see how data randomly generated from a DM model with parameter $\hat{\boldsymbol{\alpha}} = (0.642, 0.858, 0.622)$, which was estimated from sample $\mathbf{X}$, spread all over the ternary diagram. On the right-hand side, data were randomly generated using a LNM model with parameters

$$\hat{\boldsymbol{\mu}} = (-0.460, 0.343) \quad \text{and} \quad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.528 & 2.706 \\ 2.706 & 4.816 \end{pmatrix}$$

estimated from $\mathbf{X}$. In this case, the obtained sample is a fairly accurate realisation of the original data.

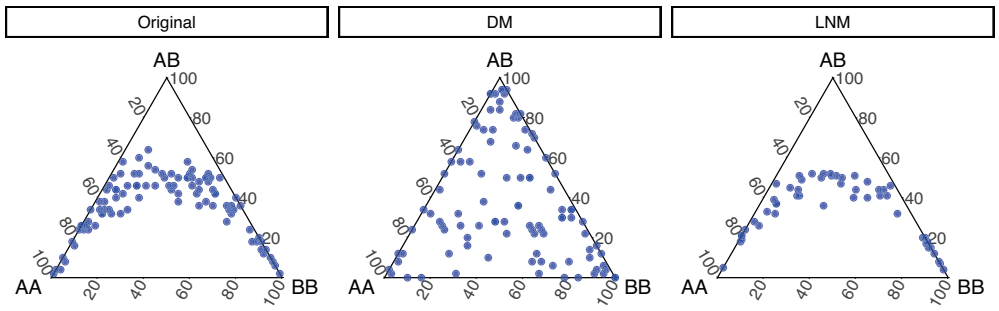

Figure 4: *Genotype frequencies $\boldsymbol{x}_i$ in the Hardy-Weinberg equilibrium. Original frequencies $\boldsymbol{x}_i$ generated in Scenario 1 (left), genotype frequencies simulated by Dirichlet-multinomial model (centre), genotype frequencies simulated by logratio-normal-multinomial model (right).*

4.3 Catalan parliamentary elections

Final vote shares by municipality from the 2015 Catalan parliamentary elections were provided by the Statistical Institute of Catalonia (http://www.idescat.cat/en/). A subsample was drawn for statistical inference based on the DM and LNM models. We aimed to evaluate by simulation the ability of an opinion poll to reflect voting intention. Table *1* shows the vote counts registered in some municipalities out of 947. We only considered votes for the six main parties obtaining seats in the parliament: Junts pel sí (jxsi), Partit socialista de Catalunya (psc), Partit popular (pp), Catalunya si que es pot (catsp), Ciutadans (cs) and Candidatura d'unitat popular (cup). Moreover, in order to have reasonable estimates for our simulations, we only considered those municipalities with at least 1000 votes registered (369 municipalities). The number of votes registered in these municipalities ranged from 1006 votes to 836687 votes.

Table 1: *Distribution of votes by municipality across political parties in 2015 Catalan parliamentary election (only some municipalities shown).*

Municipality	**jxsi**	**psc**	**pp**	**catsp**	**cs**	**cup**	**Population**
Abella de la Conca	57	11	5	3	9	18	170
Abrera	1685	1275	619	938	1771	453	12125
Àger	271	14	25	10	18	45	594
Agramunt	1940	161	185	104	252	152	5515
Aguilar de Segarra	121	0	4	5	4	18	251
Agullana	298	26	15	16	31	80	826
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Using all available data, for each municipality i we calculated the vector of probabilities $\mathbf{p}_i = \frac{1}{n'_i}\left(n'_{i,\text{jxsi}},\ldots,n'_{i,\text{cup}}\right)$, where $n'_{i,j}$ is the total number of votes to party j in municipality i and $n'_i = n'_{i,\text{jxsi}} + \cdots + n'_{i,\text{cup}}$. In this example, the vector of probabilities $\mathbf{p}_i$, $i = 1\ldots369$, was considered the gold standard. That is, the best available estimation of the vote probabilities across parties in each municipality. We created an artificial survey by selecting for each municipality a subsample consisting of n_i registered votes. More formally, we created a sample $\mathbf{X} = \{\mathbf{x}_1,\ldots,\mathbf{x}_{369}\}$ with $\mathbf{x}_i$ following a multivariate hypergeometric distribution with population size $\left(n'_{i,\text{jxsi}},\ldots,n'_{i,\text{cup}}\right)$ and sample size n_i. We considered two scenarios for n_i:

1. Proportional size: n_i as percentage of n'_i, ranging from 0.5% to 5%.
2. Constant size: n_i constant for all municipalities, with n_i ranging from 10 to 200.

In both cases, we repeated the experiment five times for each value of n_i . Given a data set $\mathbf{X}$, we compared estimates $\hat{\mathbf{p}}_{i,\text{DM}}$ and $\hat{\mathbf{p}}_{i,\text{LNM}}$ with the gold standard $\mathbf{p}_i$. Because data did not follow any particular distribution, we used three different criteria in this comparative analysis as proposed in Palarea-Albaladejo and MartÃn-Fernández (2008):

- Average of Aitchison distances: $\frac{1}{369}\sum_{i=1}^{369}\text{dist}_{\mathcal{A}}(\mathbf{p}_i,\hat{\mathbf{p}}_{i\cdot})$,
- Frobenius distance between the covariance matrix, $\mathbf{\Sigma_p} = \left(\sigma_{ij}^{\mathbf{p}}\right) \in \mathbb{R}^{5\times5}$, obtained from $\{\text{ilr}(\mathbf{p}_1),\ldots,\text{ilr}(\mathbf{p}_{369})\}$ and the covariance matrix, $\mathbf{\Sigma}_{\hat{\mathbf{p}}\cdot} = \left(\sigma_{ij}^{\hat{\mathbf{p}}\cdot}\right) \in \mathbb{R}^{5\times5}$, obtained from $\{\text{ilr}(\hat{\mathbf{p}}_{1\cdot}),\ldots,\text{ilr}(\hat{\mathbf{p}}_{369\cdot})\}$, i.e.

$$\left\|\mathbf{\Sigma_p} - \mathbf{\Sigma}_{\hat{\mathbf{p}}\cdot}\right\|_{\mathbf{F}} = \sqrt{\sum_{i=1}^{5}\sum_{j=1}^{5}(\sigma_{ij}^{\mathbf{p}} - \sigma_{ij}^{\hat{\mathbf{p}}\cdot})^2},$$

  and
- STRESS (standardised residual sum of squares) index given by

$$STRESS = \sqrt{\frac{\sum_{i=1}^{369}\sum_{j=1}^{369}\left(\text{d}_{\mathcal{A}}(\mathbf{p}_i,\mathbf{p}_j) - \text{d}_{\mathcal{A}}(\hat{\mathbf{p}}_{i\cdot},\hat{\mathbf{p}}_{j\cdot})\right)^2}{\sum_{i=1}^{369}\sum_{j=1}^{369}\text{d}_{\mathcal{A}}(\mathbf{p}_i,\mathbf{p}_j)^2}}.$$

Figure 5 shows the results for different sample sizes. The values of the three measures decreased when the size n_i increased in all cases (Fig. 5). A parallelism between the results for DM and LNM compound distributions is observed. Noticeably, the performance of DM was worse than LNM in all the cases. The alternative starting points, either SP1 or SP2, showed similar behaviour, specially using the average Aitchison dis-

tance criterium. Importantly, for a similar sample size, the results for the second scenario (constant size) were better than those for the first scenario (proportional size).

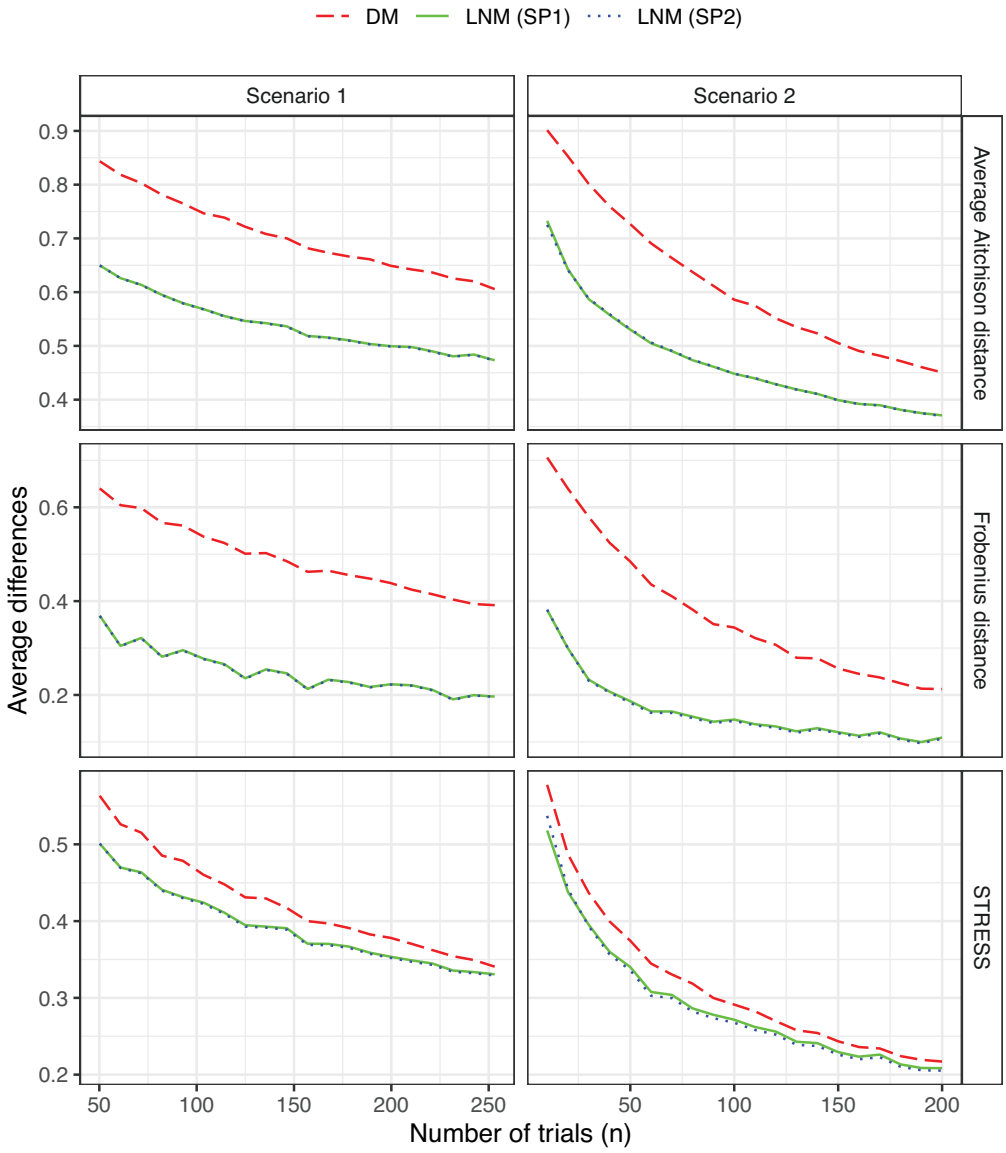

Figure 5: *Performance measures for the 2015 Catalan election example: average of Aitchison distance (top), Frobenius distance (centre), STRESS index (bottom). Estimates obtained using models DM (dashed red line), LNM (starting point SP1, dotted blue line) and LNM (starting point SP2, solid green line). Two scenarios with different sample sizes (see text for details).*

4.4 Comparison of computing times

Table 2: *Mean computing time in seconds for the three examples shown in Section 4 (95% confidence interval in parenthesis). Calculations performed on an Intel(R) Xeon(R) CPU E5-2630 v3 @ 2.40GHz. Times marked with* [1] *did not iterated trough the EM algorithm because after initialisation* Σ_0 *was singular.*

	DM	LNM (SP1)	LNM (SP2)
Pure multinomial process example			
Scenario 1 (D=9)	11.6 [4.2, 23.7]	152.6 [133.7, 172.9]	22.2 [9.6, 31.9][1]
Scenario 2 (D=12)	16.1 [3.1, 28.6]	194.7 [169.2, 223.8]	23.1 [13.6, 39.2][1]
Scenario 3 (D=15)	20.0 [10.7, 35.0]	231.3 [189.8, 310.7]	25.4 [19.1, 43.3][1]
Scenario 4 (D=16)	22.6 [6.5, 39.3]	255.9 [156.8, 334.6]	28.7 [13.6, 51.1][1]
Scenario 5 (D=20)	24.8 [11.5, 35.4]	145.3 [2.2, 310.6]	35.3 [18.8, 52.3][1]
Scenario 6 (D=25)	22.2 [12.6, 32.2]	218.2 [1.8, 535.5]	33.0 [23.7, 476.5][1]
Scenario 7 (D=30)	34.3 [9.2, 42.9]	2.4 [2.4, 2.7][1]	41.6 [10.9, 51.9][1]
Scenario 8 (D=36)	27.8 [11.2, 53.4]	3.5 [3.4, 3.6][1]	44.2 [14.9, 65.7][1]
Scenario 9 (D=50)	58.0 [11.1, 80.3]	7.0 [6.9, 9.0][1]	77.1 [45.2, 92.6][1]
Hardy-Weinberg equilibrium example			
Scenario 1 (D=3)	2.3 [1.8, 2.5]	219.2 [152.6, 286.9]	216.7 [153.8, 263.4]
Scenario 2 (D=3)	1.8 [1.6, 1.9]	878.7 [735.1, 1015.3]	565.6 [484.4, 612.3]
Scenario 3 (D=3)	2.3 [2.1, 2.6]	1319.0 [1208.2, 1548.5]	695.6 [593.3, 757.5]
Scenario 4 (D=3)	1.8 [1.6, 1.9]	66.3 [55.9, 92.7]	129.5 [89.5, 179.7]
Scenario 5 (D=3)	2.0 [1.9, 2.3]	160.7 [101.9, 208.3]	146.4 [116.1, 198.1]
Scenario 6 (D=3)	1.6 [1.5, 1.8]	327.6 [232.4, 431.0]	195.3 [140.7, 253.3]
Catalan parliamentary election example			
Scenario 1 (D=6)	1.5 [1.4, 1.7]	508.6 [430.4, 617.2]	430.2 [379.7, 510.0]
Scenario 2 (D=6)	0.8 [0.7, 1.3]	124.7 [84.3, 208.5]	186.6 [116.7, 272.2]

For the three examples above, the computing time spent on parameter estimation using LNM was higher than using DM (see Table 2). For LNM, both choices of starting points (SP1 and SP2) provided similar results, although SP1 tended to be slower. For the second example (subsection 4.2), the computing time was remarkably higher in scenarios 2 and 3 using LNM. Note that these scenarios were characterised for being the ones with the smaller probability in the first component. As expected, data dimensionality was the major factor affecting computing time.

5 Final remarks and conclusions

Count data are commonly generated in modern scientific areas such as text mining or genomic and microbiome studies based on next generation sequencing technologies. The DM distribution is a popular choice to model multivariate counts. However it may not be appropriate for complicated correlation structures because, amongst others, it imposes a negative correlation between every pair of multinomial categories. This might

not be realistic when analysing for example microbiome data (Mandal et al., 2015). Consequently, there is a need for models allowing more flexible dependence structures in multivariate count data.

The LNM compound probability distribution has been introduced in this work as a flexible model for multivariate count data. Rooted on the theoretical framework for compositional data modelling, the LNM model is fully compatible with the geometry of the simplex, the sample space where multinomial probabilities lay. Accordingly, multinomial probabilities can be conveniently mapped onto real space through logratio coordinates with respect to an orthonormal basis of the simplex for the purpose of parameter estimation. Importantly, results are invariant under changes of orthonormal basis. Parameter estimates for the LNM model cannot be computed analytically though. Different estimation approaches have been discussed and compared in this work. One based on a quasi-Monte Carlo EM algorithm is concluded to be preferable. This approach improves estimates obtained by Markov Chain Monte Carlo based on the Metropolis algorithm as used in previously works. Because inference is based on the EM algorithm, likelihood estimation can get stuck in some local maxima. Even though in the examples shown in this manuscript the global maximum is obtained, it is possible that different initialisations can be necessary to find it in particular cases.

In terms of modelling, we have shown that the LNM model produces better results than the DM. In particular, we have shown that in realistic cases LNM outperforms DM in its ability to model the underlying probabilities from the observed counts. It is important to remark that the number of parameters for a DM and a LNM grow linearly and quadratically respectively. So when the number of dimension is high, it is recommended to consider some parametrisation for the covariance matrix (Pinheiro and Bates, 1996, for an example where different restrictions on the spectral decomposition are applied to Gaussian finite mixtures see (Banfield and Raftery, 1993)).

Modelling multivariate count data using the LNM provides extra flexibility for the multinomial parameter distribution. In addition, it opens up the possibility of defining new statistical inference tools for compositional data analysis. Areas for future development include improved procedures for obtaining fast and reliable maximum likelihood estimates, e.g. along the lines of recent work by (Silverman et al., 2019). These and other questions in relation to the proposed LNM model will be addressed in future work.

References

Aitchison, J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, London (UK). Reprinted in 2003 by Blackburn Press.

Aitchison, J. and Shen, S. M. (1980). Logistic-normal distributions: Some properties and uses. *Biometrika*, 67, 261–272.

Aitchison, J. and Ho, C. H. (1989). The multivariate Poisson-Log Normal Distribution. *Biometrika*, 76, 643–653.

Banfield, J. and Raftery, A. E. (1993). Model-based Gaussian and Non-Gaussian Clustering. *Biometrics*, 49, 803–821.

Billheimer, D., Guttorp, P. and Fagan, W. F. (2001). Statistical Interpretation of Species Composition. *Journal of the American Statistical Association*, 96, 1205–1214.

Blei, D. M., and Lafferty, J. D. (2007). A Correlated Topic Model of Science. *The Annuals of Applied Statistics*, 1, 1–21.

Bouguila, N. (2008). Clustering of Count Data Using Generalized Dirichlet Multinomial Distributions. *IEEE Transactions on Knowledge and Data Engineering*, 20, 462–474.

Caflisch, R. E. (1998). Monte Carlo and quasi-Monte Carlo methods. *Acta Numerica*, 7, 1–49.

Chastin, S. F., Palarea-Albaladejo, J., Dontje, M. L. and Skelton, D. A. (2015). Combined effects of time spent in physical activity, sedentary behaviours and sleep on obesity and cardiometabolic health markers: A novel compositional data analysis approach. *PLoS ONE*, 10, e0139984.

Comas-Cufí, M., Martín-Fernández, J. A. and Mateu-Figueras, G. (2016). Logratio methods in mixture models for compositional data sets. *SORT*, 40, 349–374.

Comas-Cufí, M., Martín-Fernández, J. A. and Mateu-Figueras, G. (2019). Merging the components of a finite mixture using posterior probabilities. *Statistical Modelling*, 19, 1–31.

Connor, R. J. and Mosimann, J. E. (1969). Concepts of independence for proportions with a generalization of the Dirichlet Distribution. *Journal of the American Statistical Association*, 64, 194–206.

Dempster, A. P., Laird, N. M. and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society, Series B*, 39, 1–38.

Drmota, M. and Tichy, R. F. (1997). *Sequences, Discrepancies and Applications*. Lecture Notes in Mathematics, vol. 1651. Springer, Berlin (1997).

Edjabou, M. E., Martín-Fernández, J. A., Scheutz, C. and Astrup, T. F. (2017). Statistical analysis of solid waste composition data: Arithmetic mean, standard deviation and correlation coefficients. *Waste Management*, 69, 13–23.

Egozcue, J. J., Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35, 279–300.

Graffelman, J. (2015). Exploring Diallelic Genetic Markers: The HardyWeinberg Package *Journal of Statistical Software*, 64.

Graffelman, J. and Weir, B. S. (2016). Testing for Hardy-Weinberg equilibrium at biallelic genetic markers on the X chromosome *Heredity*, 116, 558–568.

Grantham, N. S., Guan, Y., Reich, B. J., Borer, E. T., and Gross, K. (2019). MIMIX: A Bayesian Mixed-Effects Model for Microbiome Data From Designed Experiments *Journal of the American Statistical Association*, 0, 1–16.

Holmes, I., Harris, K. and Quince, C. (2012) Dirichlet Multinomial Mixtures: Generative Models for Microbial Metagenomics *PLOS ONE*, 7, e30126.

Hughes, G., Munkvold, G. P. and Samita, S. (1998). Application of the logistic-normal-binomial distribution to the analysis of Eutypa dieback disease incidence. *International Journal of Pest Management*, 44, 35–42.

Jank, W. (2005) Quasi-Monte Carlo sampling to improve the effciency of Monte Carlo EM *Computational Statistics & Data Anaylsis*, 48, 685–701.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1997). *Discrete Multivariate Distributions* Series in probability and statistics. John Wiley & Sons, Inc, New York (1997).

Kuo, F. Y., Dunsmuir,W. T. M., Sloan, I. H., Wand, M. P. and Womersley, R. S. (2008). Quasi-Monte Carlo for Highly Structured Generalised Response Models. *Methodology and Computing in Applied Probability*, 10, 239–275.

Layton, D. F. and Siikamäki, J. (2009). Payments for ecosystem services programs: predicting landowner enrollment and opportunity cost using a beta-binomial model *Environmental and Resource Economics*, 44, 415–439.

L'Ecuyer, P. and Lemieux, C. (2002). Recent advances in randomized quasi-Monte Carlo methods. In *Modeling Uncertainty: An Examination of Stochastic Theory, Methods, and Applications*, 419–474. Kluwer Academic Publishers.

Leobacher, G. and Pillichshammer, F. (2014) *Introduction to Quasi-Monte Carlo Integration and Applications*. Compact Textbooks in Mathematics, Springer International Publishing.

Lindsay, B. G. (1995). *Mixture Models: Theory, Geometry and Applications*. Haywood, CA: Institute of Mathematical Sciences; Alexandria VA: American Statistical Association

Mandal, S., Van Treuren, W., White, R., Eggesbo, M., Knight, R. and Peddada, S. (2015) Analysis of composition of microbiomes: a novel method for studying microbial composition *Microbial Ecology in Health and Disease*, 26, 27663.

Martín-Fernández, J. A., Hron, K., Templ, M., Filzmoser, P. and Palarea-Albaladejo, J. (2015) Bayesian-multiplicative treatment of count zeros in compositional data sets *Statistical Modelling*, 15, 134–158.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). The principle of working on coordinates. In *Compositional Data Analysis*, 29–42. John Wiley & Sons, Ltd.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2013). The normal distribution in some constrained sample spaces. *SORT*, 37, 29–56.

Minka, T. P. (2004). The Dirichlet-tree distribution https://www.microsoft.com/en-us/research/publication/dirichlet-tree-distribution (last access December/2019)

Morokoff, W. J. and Caflisch, R. E. (1995). Quasi-Monte Carlo Integration. *Journal of Computational Physics*, 122, 218–230.

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

Neal, R. M. (2010). *MCMC Using Hamiltonian Dynamics*. Handbook of Markov Chain Monte Carlo, 54, 113–162.

Neath, R. C. (2013). On convergence Properties of the Monte Carlo EM Algorithm. In *Advances in Modern Statistical Theory and Applications: A Festschrift in honor of Morris L. Eaton*, 43–62. Institute of Mathematical Statistics.

Nelson, J. F. (1985). Multivariate Gamma-Poisson Models. *Journal of the American Statistical Association*, 80, 828–834.

Ongaro, A. and Migliorati, S. (2013). A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis*, 114, 412–426.

Owen, A. B. (1995) Randomly permuted (t,m,s)-nets and (t,s)-sequences In *Monte Carlo and Quasi-Monte Carlo Methods in Scientific Computing*, 299–317. Springer-Verlag.

Palarea-Albaladejo, J. and Martín-Fernández, J. A. (2008). A modified EM alr-algorithm for replacing rounded zeros in compositional data sets. *Computers & Geosciences*, 34, 902–917.

Palarea-Albaladejo J., Rooke J. A., Nevison, I. M. and Dewhurst, R. J. (2017). Compositional mixed modeling of methane emissions and ruminal volatile fatty acids from individual cattle and multiple experiments. *Journal of Animal Science*, 95, 2467–2480.

Pan, J. and Thompson, R. (2007). Quasi-Monte Carlo estimation in generalized linear mixed models. *Computational Statistics and Data Analysis*, 51, 5765–5775.

Pawlowsky-Glahn V and Egozcue JJ (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15, 384–398.

Pinheiro, JC and Bates, DM (1996). Unconstrained parametrizations for variance-covariance matrices. *Statistics and Computing*, 6, 289–296.

R Development Core Team (2015). R: A language and environment for statistical computing. *Vienna: R Foundation for Statistical Computing*, URL http://www.r-project.org (last accessed on 23 November 2017).

Robbins, H. (1964). The empirical Bayes approach to statistical decision problems. *The Annals of Mathematical Statistics*, 35, 1–20.

Robbins, H. (1980). Estimation and prediction for mixtures of the exponential distribution. *Proceedings of the National Academy of Sciences*, 77, 2382–2383.

Scheffé, H. (1958) Experiments with mixtures. *Journal of the Royal Statistical Society, series B (Methodological), 20, 344–360.*

Silverman, J. D., Durand, H. K., Bloom, R. J., Mukherjee, S. and David, L. A. (2018) Dynamic linear models guide design and analysis of microbiota studies within artificial human guts. *Microbiome, 6, 6–202.*

Silverman, J. D., Roche, K., Zachary, C. H., David, L. A. and Mukherjee, S. (2019) Bayesian Multinomial Logistic Normal Models through Marginally Latent Matrix-T Processes. *arXiv: 1903.11695.*

Wang, X. and Fang K. T. (2003). The effective dimension and quasi-Monte Carlo integration *Journal of Complexity*, 19, 101–124.

Xia, M., Chen, J., Fung K. F. and Li H. (2013). A Logistic Normal Multinomial Regression Model for Microbiome Compositional Data Analysis. *Biometrics*, 69, 1053–1063.

A Proof of properties 1 and 2

Property 1 *For a fixed* $\boldsymbol{x}$ *we have*

$$\lim_{\|\boldsymbol{\Sigma}\|\to 0} \mathcal{LNM}(\boldsymbol{x};n,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{M}\left(\boldsymbol{x};n,\mathrm{ilr}^{-1}(\boldsymbol{\mu})\right).^{2}$$

Proof. Let $\mathbf{h}$ be a real vector defined on $\mathbb{R}^{D-1}$ and let $\mathbf{x}$ be a count vector defined on $\mathcal{S}^{n,D}$. Because $\mathrm{ilr}_1^{-1}(\mathbf{h})^{x_1}\cdots\mathrm{ilr}_D^{-1}(\mathbf{h})^{x_D} \leq {x_1}^{x_1}\cdots{x_D}^{x_D}$, for any fixed $\mathbf{x}=(x_1,\ldots,x_D)$ we have:

$$\begin{aligned}\lim_{\|\boldsymbol{\Sigma}\|\to 0} \mathcal{LNM}(\mathbf{x};n,\boldsymbol{\mu},\boldsymbol{\Sigma}) &= \int_{\mathbf{h}\in\mathbb{R}^{D-1}} \frac{n!}{x_1!\cdots x_D!}\mathrm{ilr}_1^{-1}(\mathbf{h})^{x_1}\cdots\mathrm{ilr}_D^{-1}(\mathbf{h})^{x_D} \lim_{\|\boldsymbol{\Sigma}\|\to 0}\mathcal{N}(\mathbf{h};\boldsymbol{\mu},\boldsymbol{\Sigma})d\mathbf{h}\\ &= \int_{\mathbf{h}\in\mathbb{R}^{D-1}} \frac{n!}{x_1!\ldots x_D!}\mathrm{ilr}_1^{-1}(\mathbf{h})^{x_1}\cdots\mathrm{ilr}_D^{-1}(\mathbf{h})^{x_D}\delta(\mathbf{h}-\boldsymbol{\mu})\,d\mathbf{h}\\ &= \frac{n!}{x_1!\ldots x_D!}\mathrm{ilr}_1^{-1}(\boldsymbol{\mu})^{x_1}\cdots\mathrm{ilr}_D^{-1}(\boldsymbol{\mu})^{x_D} = \mathcal{M}\left(\mathbf{x};n,ilr^{-1}(\boldsymbol{\mu})\right).\end{aligned}$$

■

Property 2 *Let* $\boldsymbol{x}=(x_1,\ldots,x_D)$ *and* $x_1+\cdots+x_D=n$. *If* $\lim_{n\to\infty}\frac{x_i}{n}=\pi_i$ *and* $\pi_i>0$ *for* $1\leq i\leq D$, *then*

$$\lim_{n\to\infty} n^{D-1}\cdot\mathcal{LNM}(\boldsymbol{x};n,\boldsymbol{\mu},\boldsymbol{\Sigma}) = \mathcal{N}_{\mathcal{S}^D}(\boldsymbol{\pi};\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{\sqrt{D}}\frac{1}{\pi_1\ldots\pi_D}$$

Proof. Let $\boldsymbol{\pi}_n=(\pi_{n,1},\ldots,\pi_{n,D})=\frac{1}{n}(x_1,\ldots,x_D)$. We have that

$$\lim_{n\to\infty} n^{D-1}\mathcal{LNM}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}) = \lim_{n\to\infty}\frac{(n+D-1)!}{n!}\mathcal{LNM}(\mathbf{x};\boldsymbol{\mu},\boldsymbol{\Sigma}). \tag{11}$$

Substituting $\mathbf{x}$ by $n\boldsymbol{\pi}_n$ and using the LNM probability mass function (5), we can rewrite (11) as

$$\int_{\mathbf{h}\in\mathbb{R}^{D-1}}\mathcal{N}(\mathbf{h};\boldsymbol{\mu},\boldsymbol{\Sigma})\lim_{n\to\infty}\frac{(n+D-1)!}{n!}\frac{n!}{(n\pi_{n,1})!\ldots(n\pi_{n,D})!}\mathrm{ilr}_1^{-1}(\mathbf{h})^{n\pi_{n,1}}\ldots\mathrm{ilr}_D^{-1}(\mathbf{h})^{n\pi_{n,D}}d\mathbf{h} \tag{12}$$

Considering the change of variable given by (3), which has the Jacobian

$$d\mathbf{h} = \frac{1}{\sqrt{D}\,p_1\ldots p_D}d\mathbf{p},$$

2. $\lim_{\|\boldsymbol{\Sigma}\|\to 0}$ stands for any sequence of covariance matrices such that their highest eigenvalue goes to 0.

we can express (12) with respect to $\mathbf{p}$

$$\int_{\mathbf{p}\in\mathcal{S}^D\subset\mathbb{R}^D} \mathcal{N}(\text{ilr}(\mathbf{p});\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{\sqrt{D}\,p_1\dots p_D}\lim_{n\to\infty}\frac{(n+D-1)!}{(n\pi_{n,1})!\dots(n\pi_{n,D})!}{p_1}^{n\pi_{n,1}}\dots{p_D}^{n\pi_{n,D}}d\mathbf{p}, \tag{13}$$

where $d\mathbf{p}$ is measured using Lebesgue measure.

Note that for the Dirichlet distribution with parameters $\alpha_i = n\pi_{n,i}+1$, $1\le i\le D$, we have

$$\int_{\mathbf{p}\in\mathcal{S}^D\subset\mathbb{R}^D}\frac{(n+D-1)!}{(n\pi_{n,1})\dots!(n\pi_{n,D})!}{p_1}^{n\pi_{n,1}}\dots{p_D}^{n\pi_{n,D}}d\mathbf{p}=1, \tag{14}$$

and using the Stirling's approximation we have

$$\begin{aligned}\varphi(\mathbf{p}) &= \lim_{n\to\infty}\frac{(n+D-1)!}{(n\pi_{n,1})!\dots(n\pi_{n,D})!}{p_1}^{n\pi_{n,1}}\dots{p_D}^{n\pi_{n,D}}\\ &= \lim_{n\to\infty}\frac{(n+D-1)!}{(2\pi n)^{\frac{D-1}{2}}\sqrt{\pi_{n,1}\dots\pi_{n,D}}}\frac{{p_1}^{n\pi_{n,1}}\dots{p_{n,D}}^{n\pi_D}}{\pi_1^{n\pi_{n,1}}\dots\pi_D^{n\pi_{n,D}}}.\end{aligned}$$

Moreover, it can be seen that $\mathbf{p}=\boldsymbol{\pi}_n$ is a global maximum for ${p_1}^{\pi_{n,1}}\dots{p_D}^{\pi_{n,D}}$ when $p_1+\dots+p_D=1$. Moreover, because $\lim_{n\to\infty}\boldsymbol{\pi}_n=\boldsymbol{\pi}$, we have that

$$\begin{aligned}\varphi(\mathbf{p}) &= \lim_{n\to\infty}\frac{(n+D-1)!}{(2\pi n)^{\frac{D-1}{2}}\sqrt{\pi_{n,1}\dots\pi_{n,D}}}\left(\frac{{p_1}^{\pi_{n,1}}\dots{p_D}^{\pi_{n,D}}}{{\pi_{n,1}}^{\pi_{n,1}}\dots{\pi_{n,D}}^{\pi_{n,D}}}\right)^n =\\ &= \begin{cases}\infty & \text{when } \mathbf{p}=\boldsymbol{\pi}\\ 0 & \text{otherwise,}\end{cases}\end{aligned}$$

which implies, together with Equation 14, that φ is the Dirac delta function centred at $\boldsymbol{\pi}$.

Putting all together, (13) can be rewritten as

$$\begin{aligned}&\int_{\mathbf{p}\in\mathcal{S}^D\subset\mathbb{R}^D}\mathcal{N}(\text{ilr}(\mathbf{p});\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{\sqrt{D}\,p_1\dots p_D}\delta(\mathbf{p}-\boldsymbol{\pi})\,d\mathbf{p}=\\ &\mathcal{N}(\text{ilr}(\boldsymbol{\pi});\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{\sqrt{D}\,\pi_1\dots\pi_D}=\mathcal{N}_{\mathcal{S}^D}(\boldsymbol{\pi};\boldsymbol{\mu},\boldsymbol{\Sigma})\frac{1}{\sqrt{D}\,\pi_1\dots\pi_D}.\end{aligned}$$

■

Observe that we obtain the logratio-normal distribution on the simplex expressed with respect to the Lebesgue measure. The term $1/(\sqrt{D}\,\pi_1\dots\pi_D)$ relates the Aitchison measure with the Lebesgue measure (Mateu-Figueras et al., 2013).

B E-step convergence in a simple univariate but extreme case

Let $\mathbf{x} = (1,0)$ be an observed vector of counts, with mean $\mu = 0$ and standard deviation $\sigma = 1$. Consider the $\mathcal{S}^2$ basis given by $\mathcal{B} = \left\{\mathbf{b}_1 = \frac{1}{e^{\sqrt{1/2}}+e^{-\sqrt{1/2}}}\left(e^{\sqrt{1/2}}, e^{-\sqrt{1/2}}\right)\right\}$. The coordinates of an element $\mathbf{p} = (p_1, p_2)$ of $\mathcal{S}^2$ with respect to the basis $\mathcal{B}$, are

$$h = \text{ilr}(\mathbf{p}) = \sqrt{\frac{1}{2}} \ln \frac{p_1}{p_2},$$

and an element of $\mathcal{S}^2$ with respect to its coordinates is

$$\mathbf{p} = \text{ilr}^{-1}(h) = h \odot \mathbf{b}_1 = \frac{1}{e^{h\sqrt{1/2}} + e^{-h\sqrt{1/2}}}\left(e^{h\sqrt{1/2}}, e^{-h\sqrt{1/2}}\right).$$

Using Equation 5 we calculate the marginal probability

$$\Pr(\{\mathbf{X} = (1,0)\}; \mu = 0, \sigma = 1) = \int_{-\infty}^{\infty} \left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{h^2}{2}} \left(\frac{e^{h\sqrt{1/2}}}{e^{h\sqrt{1/2}} + e^{-h\sqrt{1/2}}}\right) dh \approx 0.50.$$

Using numerical integration we obtain an approximation of these expected values:

$$\mathbb{E}_{H|X;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}[h] = \int_{-\infty}^{\infty} h \frac{\left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{h^2}{2}} \left(\frac{e^{h\sqrt{1/2}}}{e^{h\sqrt{1/2}}+e^{-h\sqrt{1/2}}}\right)}{\Pr(\{\mathbf{X} = (1,0)\}; \mu = 0, \sigma = 1)} dh \approx 0.5136.$$

$$\mathbb{E}_{H|X;\boldsymbol{\mu}^{(s)},\boldsymbol{\Sigma}^{(s)}}\left[h^2\right] = \int_{-\infty}^{\infty} h^2 \frac{\left(\frac{1}{\sqrt{2\pi}}\right) e^{-\frac{h^2}{2}} \left(\frac{e^{h\sqrt{1/2}}}{e^{h\sqrt{1/2}}+e^{-h\sqrt{1/2}}}\right)}{\Pr(\{\mathbf{X} = (1,0)\}; \mu = 0, \sigma = 1)} dh \approx 1.0.$$

After performing a fixed number of simulations, we compare the estimate and the variance of the error when approximating the expected values $\mathbb{E}_{h|\mathbf{x},\mu,\sigma}(h)$ and $\mathbb{E}_{h|\mathbf{x},\mu,\sigma}\left(h^2\right)$ using five different Monte Carlo approaches: MC method via importance sampling as described in Section 3.1, MC method via importance sampling and antithetic variates (Caflisch, 1998), QMC method using Halton low-discrepancy sequences, MCMC method based on the Metropolis algorithm with a standardised gaussian proposal (Xia et al., 2013), and MCMC method based on the Hamiltonian algorithm (Chapter 5, Neal (2010)). For QMC estimation, the variability was estimated using *scrambling* techniques (see Owen (1995); L'Ecuyer and Lemieux (2002) for further details). Importance sampling for MC and QMC was conducted using $m = \mathbb{E}_{h|\mathbf{x},\mu,\sigma}(h)$ and $s = 1$. MCMC methods were initiated at $h_0 = \mathbb{E}_{h|\mathbf{x},\mu,\sigma}(h)$.

Table 3: *Mean and standard deviation (in parenthesis) of 500 different approximations to* $\mathbb{E}_{h|\boldsymbol{x},\mu,\sigma}(h)$ *and* $\mathbb{E}_{h|\boldsymbol{x},\mu,\sigma}\left(h^2\right)$ *when* $\boldsymbol{x}=(1,0)$, $\mu=0$ *and* $\sigma=1$. *Computing time is shown with respect to the MC method.*

Method	First moment	Second moment	Time
Numerical approximation	0.5135884	1.0000000	
MC	0.5136272 (0.02412)	1.0008430 (0.03998)	×1.00
MC (Antithetic variates)	0.5136542 (0.00148)	1.0017754 (0.04047)	×0.96
QMC	0.5135818 (0.00071)	0.9999877 (0.00171)	×2.03
MCMC (Metropolis)	0.5133551 (0.02928)	0.9996497 (0.04716)	×6.70
MCMC (Hamiltonian)	0.5163096 (0.04394)	1.0037255 (0.07058)	×90.12

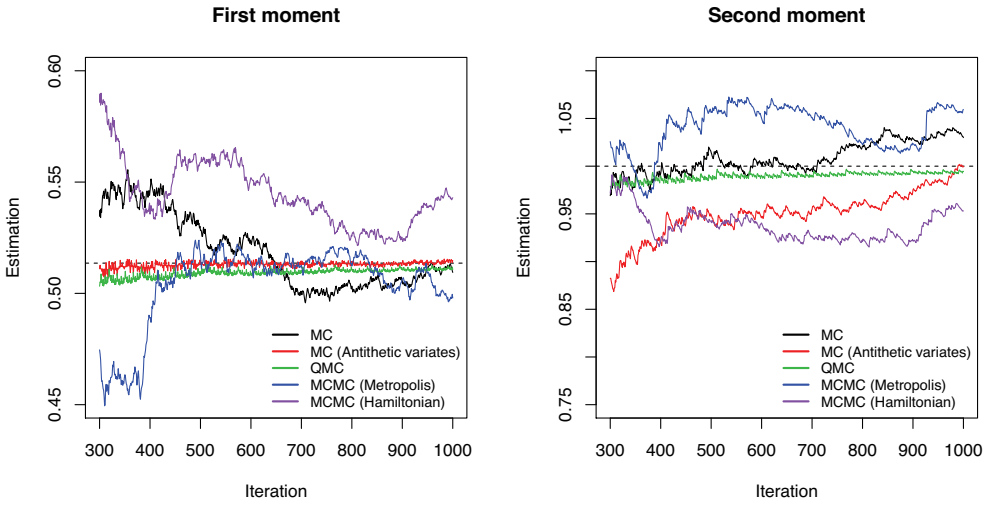

Figure 6: *Approximation of moments:* $\mathbb{E}_{h|\boldsymbol{x},\mu,\sigma}(h)$ *(left);* $\mathbb{E}_{h|\boldsymbol{x},\mu,\sigma}\left(h^2\right)$ *(right). For* $\boldsymbol{x}=(1,0)$, $\mu=0$ *and* $\sigma=1$ *using different MC techniques to approximate the E step: Monte Carlo (MC) via importance sampling (black), MC via importance sampling and antithetic variates (red), QMC using Halton sequences (green), MCMC based on the Metropolis algorithm (blue) and MCMC based on the Hamiltonian algorithm (purple). Horizontal dashed line (black) is approximation calculated by numerical integration.*

Figure 6 shows the behaviour of the methods for the first 1000 iterations in one simple approximation. The horizontal dashed line (in black) represents the expected values calculated by numerical integration. This exercise was repeated 500 times. Table 3 shows the mean and the standard deviation of the corresponding 500 approximations obtained by each procedure in the first 1000 iterations. In addition, a comparison of the computing time was conducted. The computing time was very similar for all methods, except for MCMC methods. The best results produced by the classical MC method. Regarding to the approximation of the first and second moment, QMC estimation clearly outperformes the other approaches. Remarkably, the MCMC algorithms has the worst performance. The standard deviations obtained by the Metropolis algorithm (0.044 and 0.071) were the largest and far from the standard deviations obtained by QMC. Figure 6

illustrates this behaviour. Note that the lines for the methods whose standard deviation was close to zero are close to the horizontal line representing the exact values.

C E-step convergence in multivariate cases

To evaluate the performance of the estimation procedures, we set up a simulation study parametrised by the following five parameters:

- Dimension of the random vector $\mathbf{H}$. We considered dimension $d \in \{1,5,25,125\}$.
- Multinomial sample size, $n \in \{10,100\}$.
- Location of parameter $\boldsymbol{\mu}$. Concretely, we parameterised the Aitchison norm for the mean of the multivariate normal (MVN) distribution, $\lambda \in \{0,1,2\}$.
- Variability of parameter $\boldsymbol{\Sigma}$. To this end, we parameterised the quotient between the trace and the dimension of the covariance matrix of the MVN distribution, $\nu \in \{0.5,1,2\}$.
- Agreement between count $\mathbf{x}$ and parameter $\boldsymbol{\mu}$. We considered two scenarios, a first scenario were count $\mathbf{x}$ was generated by a multinomial distribution with parameter $\boldsymbol{\pi} = \text{ilr}(\boldsymbol{\mu})$, and a second scenario were count $\mathbf{x}$ was generated by a multinomial distribution with parameter $\boldsymbol{\pi} = \text{ilr}(-\boldsymbol{\mu})$. We parameterised the two scenarios with a parameter $\xi \in \{0,1\}$ to modelate the two multinomial distributions with $\boldsymbol{\pi} = \text{ilr}^{-1}((2\xi-1)\boldsymbol{\mu})$. Parameter ξ measures the change from a situation with disagreement between $\mathbf{x}$ and $\boldsymbol{\mu}$ to a situation with agreement between them.

In each of the previous 144 scenarios we repeated the following simulation 100 times:

1. A vector $\boldsymbol{\mu} \in \mathbb{R}^d$ was uniformly generated from the d-sphere with radius λ, i.e. $\{\boldsymbol{\mu} \in \mathbb{R}^d;\ \|\boldsymbol{\mu}\|_2 = \lambda\}$.
2. A covariance matrix $\boldsymbol{\Sigma} \in \mathbb{R}_{d\times d}$ was generated as $\boldsymbol{\Sigma} = \frac{\nu}{\text{tr}(\mathbf{A})/d}\mathbf{A}$, where $\mathbf{A} \sim \text{Wishart}(d,\mathbf{I}_d)$ (ensuring that $\frac{\text{tr}(\boldsymbol{\Sigma})}{d} = \nu$).
3. A vector $\mathbf{X}$ was generated following a multinomial distribution with sample size n and probability $\boldsymbol{\pi} = \text{ilr}^{-1}((2\xi-1)\boldsymbol{\mu})$.
4. We approximated the first and second moment of the random variable $\mathbf{H}$ conditional to $\mathbf{X}$, $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ using Monte Carlo, Monte Carlo with antithetic variate, Quasi Monte Carlo and MCMC (using the Metropolis-Hastings algorithm). Approximations were conducted generating 100 random variables.

In each scenario, for the first and second moment, gold standards were obtained with 100000 replicates using standard Monte Carlo integration and MCMC. We evaluated the accuracy of each method by calculating the infinity norm of the difference between the approximation and the gold standard.

Table 4: *Results obtained after adjusting a linear model to the logarithm of the error.*

	Dependent variable: ln(error)	
	First moment	Second moment
	Effect (95% CI)	Effect (95% CI)
MC (Reference)	1	1
MC-AV	0.404 (0.378, 0.429)	0.888 (0.861, 0.915)
QMC	0.443 (0.417, 0.469)	0.612 (0.585, 0.639)
MCMC	2.182 (2.156, 2.208)	2.014 (1.987, 2.041)
d	1.020 (1.020, 1.020)	1.023 (1.023, 1.024)
n	0.994 (0.994, 0.994)	0.989 (0.989, 0.989)
λ	1.156 (1.145, 1.167)	1.532 (1.521, 1.544)
ν	1.583 (1.569, 1.598)	1.851 (1.835, 1.866)
ξ	1.056 (1.037, 1.074)	1.176 (1.157, 1.195)
R^2	0.574	0.600

To assess the results we fitted a linear model to investigate differences in logarithmic error with respect to the methods used. Results were further adjusted by the five parameters: d, n, λ, ν and ξ (Table 4). The table shows the relative effect of each parameter when estimating the error. On average, QMC produced 44.3% and 61.2% of the error to estimate the first and second moment respectively in comparison to standard MC. In contrast, MCMC methods doubled the error in both moments with respect to MC (2.182 and 2.014). The use of antithetic variables (MC-AV) provided the best results when approximating the first moment. In relation to the parameters, as it is expected, the higher the dimension the higher the error. We also observe the high impact of the MVN variability, ν, parameter in both moments (1.58 and 153 respectively). In minor measure, the same occurred for the norm, λ, and the disagreement, ξ. On the contrary, the higher the sample size, n, the lower the error.

To have a visual summary of the results, Figure 7 shows different boxplots of the parameters d, λ and ν. As seen in Table 4, this graphic illustrates how antithetic variates perform specially well in low dimensions when estimating the first moment. QMC method performs well in almost all scenarios, when estimating both moments.

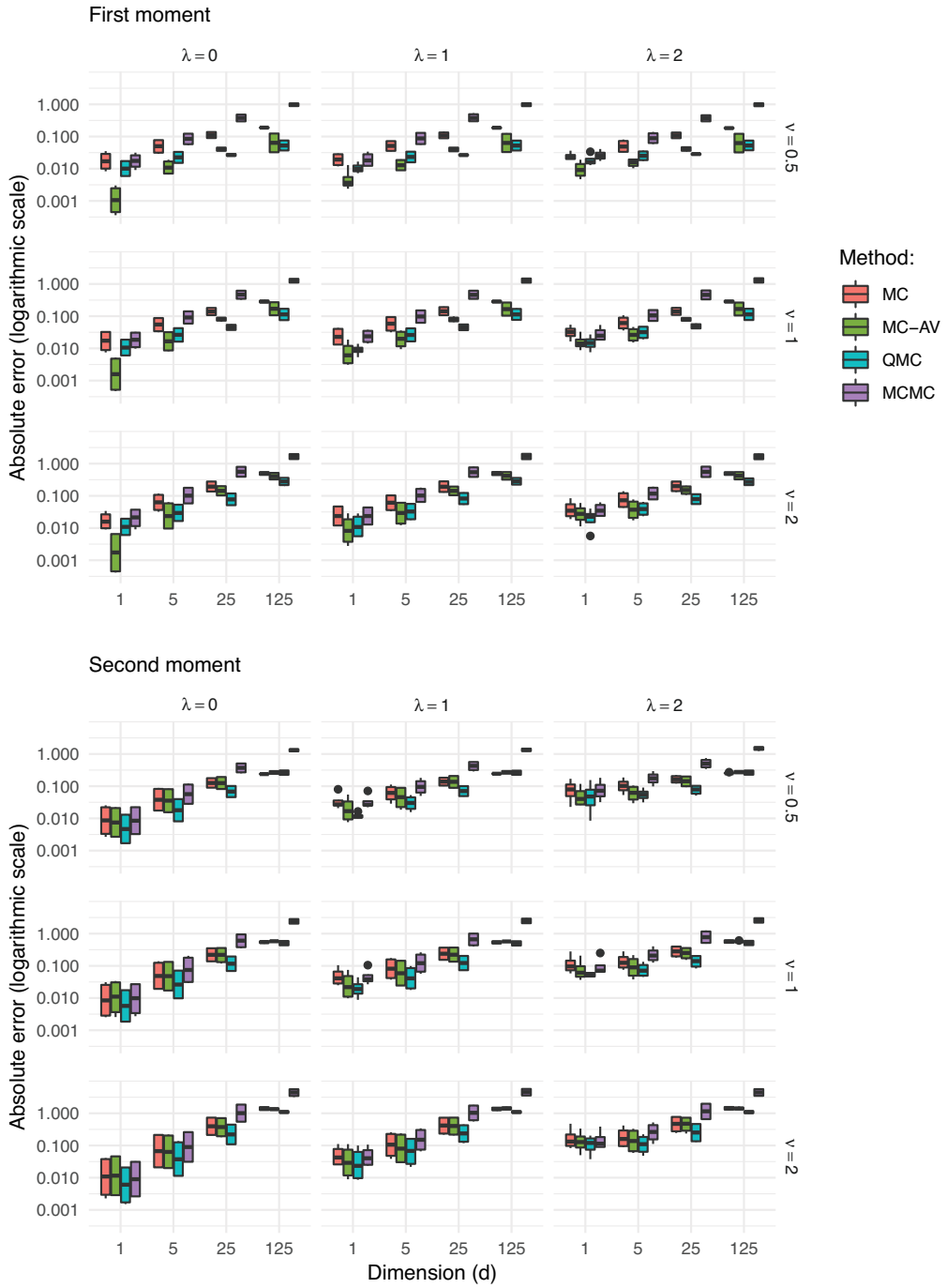

Figure 7: *Results obtained in the simulation study for the first and second moment. Results are shown for each method with respect to parameters d, λ, ν.*

D Choices for the probability vector p

Table 5: *Probability vectors* ***p*** *used in the first example. Column k shows the scenario number, column D refers to the number of components. The column on the right-hand side accounts for the initial numbers of trials.*

k	*D*	**p**
1	9	0.057 0.077 0.078 0.105 0.105 0.105 0.141 0.141 0.191
2	12	0.066 0.071 0.072 0.076 0.078 0.078 0.084 0.086 0.087 0.096 0.097 0.109
3	15	0.024 0.033 0.033 0.044 0.044 0.044 0.059 0.059 0.059 0.080 0.080 0.080 0.108 0.108 0.145
4	16	0.024 0.031 0.031 0.041 0.041 0.041 0.056 0.056 0.056 0.056 0.075 0.075 0.075 0.102 0.102 0.13
5	20	0.016 0.020 0.020 0.028 0.028 0.028 0.037 0.037 0.037 0.037 0.050 0.050 0.050 0.050 0.068 0.068 0.068 0.092 0.092 0.124
6	25	0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.01 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.02 0.05 0.05 0.05 0.05 0.50
7	30	0.021 0.023 0.023 0.025 0.025 0.026 0.028 0.028 0.028 0.028 0.030 0.030 0.031 0.031 0.031 0.033 0.034 0.034 0.034 0.034 0.037 0.038 0.038 0.038 0.042 0.042 0.042 0.047 0.047 0.052
8	36	0.019 0.019 0.020 0.020 0.021 0.021 0.021 0.021 0.022 0.022 0.023 0.023 0.024 0.024 0.024 0.024 0.024 0.026 0.026 0.027 0.027 0.028 0.028 0.029 0.029 0.030 0.032 0.032 0.033 0.033 0.037 0.037 0.038 0.043 0.043 0.050
9	50	0.02 (50 times)