Statistics and Computing (2024) 34:93
https://doi.org/10.1007/s11222-024-10412-w

ORIGINAL PAPER

Robust score matching for compositional data

Janice L. Scealy[1] · Kassel L. Hingee[1] · John T. Kent[2] · Andrew T. A. Wood[1]

Received: 27 September 2023 / Accepted: 17 February 2024 / Published online: 13 March 2024

Abstract
The restricted polynomially-tilted pairwise interaction (RPPI) distribution gives a flexible model for compositional data. It is particularly well-suited to situations where some of the marginal distributions of the components of a composition are concentrated near zero, possibly with right skewness. This article develops a method of tractable robust estimation for the model by combining two ideas. The first idea is to use score matching estimation after an additive log-ratio transformation. The resulting estimator is automatically insensitive to zeros in the data compositions. The second idea is to incorporate suitable weights in the estimating equations. The resulting estimator is additionally resistant to outliers. These properties are confirmed in simulation studies where we further also demonstrate that our new outlier-robust estimator is efficient in high concentration settings, even in the case when there is no model contamination. An example is given using microbiome data. A user-friendly R package accompanies the article.

Keywords Zeros · Log-ratios · PPI model · Outliers

1 Introduction

The polynomially-tilted pairwise interaction (PPI) model for compositional data was introduced by Scealy and Wood (2023). It is a flexible model for compositional data because it can model high levels of right skewness in the marginal distributions of the components of a composition, and it can capture a wide range of correlation patterns. Empirical investigations in Scealy and Wood (2023) showed that this distribution can successfully describe the behaviour of real data in many settings. They illustrated its effectiveness on a set of microbiome data. Some other recent related articles are Yu et al. (2021) and Weistuch et al. (2022).

Many articles in the literature analysing microbiome data, including for example Cao et al. (2019), He et al. (2021), Mishra and Muller (2022) and Liang et al. (2022), assume that zeros are effectively outliers since they replace all zero counts by 0.5 (or they use some other arbitrary constant to impute the zeros), then take a log-ratio transformation of the proportions and apply Euclidean data analysis methods. This is a form of Winsorisation and depending on the target of inference, this zero replacement method can often lead to bias in parameter estimators. There are typically huge numbers of zeros in microbiome data and we argue that they should not be automatically treated as outliers, but rather they should be treated as legitimate datapoints that occur with relatively high probability.

The purpose of this article is to develop a novel method of estimation for the PPI model with several attractive features: (a) it is tractable, (b) it is insensitive to zero values in any of the components of a data composition, and (c) it is resistant to outliers. Outliers can occur when the majority of the dataset is highly concentrated in a relatively small region of the simplex. An observation not close to the majority would be deemed an outlier. The method is based on score matching estimation (SME) after an additive log-ratio transformation, plus the inclusion of additional weights in the estimating equations for resistance to outliers. Our approach uses the additive log-ratio transformation as a device to obtain parameter estimators for the PPI model and we are not transforming the data itself prior to the analysis as in the Aitchison (1986) approach. Although the method is mathematically well-defined for the full PPI model, it is helpful in practice to focus on a restricted version of the PPI model (the RPPI model) both for identifiability reasons and because the

✉ Janice L. Scealy
janice.scealy@anu.edu.au

[1] Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT 2601, Australia

[2] Department of Statistics, University of Leeds, Leeds LS2 9JT,, UK

restricted model was shown by Scealy and Wood (2023) to provide a good fit to microbiome data.

The article is organized as follows. The PPI model is presented in Sect. 2. The important distinction between zero components and outliers is explored in Sect. 3. Section 4 gives the score matching algorithm and Sect. 5 gives the modifications needed for resistance to outliers. An application to microbiome data analysis is given in Sect. 6, with simulation studies in Sect. 7. Further technical details and simulation results are given in the Supplementary Material which also includes a document that reproduces the numerical results in this article.

A package designed for the Comprehensive R Archive Network (CRAN 2022) is under development and available at github.com/kasselhingee/scorecompdir. The package contains our new additive log-ratio score matching estimator and its robustified version, other score matching estimators, and a general capacity for implementing score matching estimators.

2 The PPI distribution

The $(p-1)$-dimensional *simplex* in $\mathbb{R}^p$ is defined by

$$\Delta^{p-1} = \left\{ \boldsymbol{u} = (u_1, u_2, \ldots, u_p)^\top \in \mathbb{R}^p : u_j \geq 0, \quad \sum_{j=1}^{p} u_j = 1 \right\}, \tag{1}$$

where a *composition* $\boldsymbol{u}$ contains p nonnegative *components* adding to 1. The boundary of the simplex consists of compositions for which one or more components equal zero. The *open simplex* Δ_0^{p-1} excludes the boundary so that $u_j > 0, \ j = 1, \ldots, p$.

The *polynomially-tilted pairwise interaction (PPI)* model of Scealy and Wood (2023) on Δ_0^{p-1} is defined by the density

$$f(\boldsymbol{u}; \boldsymbol{D}, \boldsymbol{\beta}) = \frac{1}{c_1(\boldsymbol{D}, \boldsymbol{\beta})} \left(\prod_{j=1}^{p} u_j^{\beta_j} \right) \exp\left(\boldsymbol{u}^\top \boldsymbol{D} \boldsymbol{u} \right), \quad \boldsymbol{u} \in \Delta_0^{p-1}, \tag{2}$$

with respect to $d\boldsymbol{u}$, where $d\boldsymbol{u}$ denotes Lebesgue measure in $\mathbb{R}^p$ on the hyperplane $\sum u_j = 1$. The density is the product of a Dirichlet factor and an exp-quadratic (i.e. the exponential of a quadratic) factor. To ensure integrability, the Dirichlet parameters must satisfy $\beta_j > -1, \ j = 1, \ldots, p$. Note that if $-1 < \beta_j < 0$, the Dirichlet factor blows up as $u_j \to 0$. The matrix $\boldsymbol{D}$ in the quadratic form is a symmetric matrix. Due to the constraint $\sum u_j = 1$ it may be assumed without loss of generality that $\mathbf{1}^\top \boldsymbol{D} \mathbf{1} = 0$.

If the last component is written in terms of the earlier components, $u_p = 1 - \sum_1^{p-1} u_j$, then (2) can be written in the alternative form

$$f(\boldsymbol{u}; \boldsymbol{A}_L, \boldsymbol{\beta}, \boldsymbol{b}_L) = \frac{1}{c_2(\boldsymbol{A}_L, \boldsymbol{b}_L, \boldsymbol{\beta})} \left(\prod_{j=1}^{p} u_j^{\beta_j} \right) \exp\left(\boldsymbol{u}_L^\top \boldsymbol{A}_L \boldsymbol{u}_L + \boldsymbol{u}_L^\top \boldsymbol{b}_L \right), \quad \boldsymbol{u} \in \Delta_0^{p-1}, \tag{3}$$

with respect to the same Lebesgue measure $d\boldsymbol{u}$, where $\boldsymbol{u}_L = (u_1, u_2, \ldots, u_{p-1})^\top$, $\boldsymbol{A}_L$ is a $(p-1) \times (p-1)$-dimensional symmetric matrix and $\boldsymbol{b}_L$ is a $(p-1)$-dimensional vector.

The full PPI model contains $(p^2 + 3p - 2)/2$ parameters. Although the parameters are mathematically identifiable, in practice it can be difficult to estimate all of them accurately. Hence it is useful to consider a *restricted PPI (RPPI)* with a smaller number of free parameters. The RPPI model contains $q = (p+2)(p-1)/2$ free parameters (the same as for the $(p-1)$-dimensional multivariate normal distribution) and is defined as follows. First, order the components so that the most abundant component u_p is listed last. Then set

$$\boldsymbol{b}_L = 0, \quad \beta_p = 0. \tag{4}$$

The above restriction on the parameters leads to a model which is similar to a generalised gamma distribution in $p-1$ dimensions. This model was shown by Scealy and Wood (2023) to provide a reasonably good fit to microbiome data.

3 Zeros and outliers

Two types of extreme behaviour in compositional data are zeros and outliers, and it is helpful to distinguish between these two concepts.

An outlier is defined to be an observation which has low probability density under the PPI model fitted to the bulk of the data on the simplex. Outliers can occur when the majority of the dataset is highly concentrated in a relatively small region of the simplex. In particular, if most of the data are highly concentrated in the middle of the simplex with small variance, then an observation close to or on the boundary would be deemed to be an outlier.

On the other hand if the marginal distribution for the jth component has a nonvanishing probability density as u_j tends to 0 (e.g. in the PPI model with $\beta_j \leq 0$), then a composition $\boldsymbol{u}$ with $u_j = 0$ would not be considered to be an outlier.

Although the PPI model has no support on the boundary of the simplex, we may still want to fit the model to data sets for which some of the compositions have components which are exact zeros. There are two main ways to think about the presence of zeros in the data. First, they may be due to

measurement error; a measurement of zero corresponds to a "true" composition lying in the interior of the simplex. Second the data may arise as counts from a multinomial distribution where the probability vector is viewed as a latent composition coming from the PPI model. See Sects. 4.3, 6 and Scealy and Wood (2023) for further details on the multinomial latent variable model. Then zero counts can occur even though the probability vector lies in Δ_0^{p-1}.

The presence of zero components in data poses a major problem for maximum likelihood estimation for the PPI model. In particular, the derivative of the log-likelihood function with respect to β_j for a single composition $\boldsymbol{u}$,

$$\partial \log f(\boldsymbol{u}; \boldsymbol{D}, \boldsymbol{\beta})/\partial\beta_j = \log u_j - \partial \log c_1/\partial\beta_j$$

is unbounded as $u_j \to 0$, which leads to singularities in the maximum likelihood estimates.

Hence we look for alternatives to maximum likelihood estimation. One promising general approach is score matching estimation (SME), due to Hyvarinen (2005). A version of SME was used by Scealy and Wood (2023) that involved downweighting observations near the boundary of the simplex. However, their method was somewhat cumbersome due to the requirement to specify a weight function and their estimator is inefficient when many of the parameters β_j, $j = 1, 2, \ldots, p-1$ are close to -1. This parameter setting is relevant to microbiome data applications.

This article uses another version of SME that we call ALR-SME because it involves an additive log-ratio transformation when constructing the estimators. ALR-SME is tractable and is insensitive to zeros, in the sense that the influence function is bounded as $u_j \to 0$ for any j.

Further, following the method of Windham (1995) it is possible to robustify ALR-SME to outliers by incorporating suitable weights in the estimating equations. Details are given in Sect. 5. In this article we make a distinction between robustness to zeros and robustness to outliers, and for clarity we often describe these forms of robustness as *insensitive to zeros* and *resistant to outliers*, respectively.

4 Additive log-ratio score matching estimation

In this section we recall the general construction of the score matching estimator due to Hyvarinen (2005), and then apply it to data from the RPPI distribution after first making an additive log-ratio transformation.

4.1 The score matching estimator

The construction of the score matching estimator starts with the *Hyvarinen divergence*, defined by

$$\Phi(g, g_0) = \frac{1}{2}\int_{\boldsymbol{y}\in\mathbb{R}^{p-1}} \{\nabla \log g(\boldsymbol{y}) - \nabla \log g_0(\boldsymbol{y})\}^2 g_0(\boldsymbol{y}) d\boldsymbol{y} \quad (5)$$

where g and g_0 are probability densities on $\mathbb{R}^{p-1}$ subject to mild regularity conditions (Hyvarinen 2005). Note that $\Phi(g, g_0) = 0$ if and only if $g = g_0$.

Let

$$g(\boldsymbol{y}) = g(\boldsymbol{y}; \boldsymbol{\pi}) \propto \exp\{\boldsymbol{\pi}^\top \boldsymbol{t}(\boldsymbol{y})\} \quad (6)$$

define an exponential family model, where $\boldsymbol{\pi}$ is a q-dimensional parameter vector and $\boldsymbol{t}(\boldsymbol{y})$ is a q-dimensional vector of sufficient statistics. Then for a given density g_0, the "best-fitting" model $g(\boldsymbol{y}; \boldsymbol{\pi})$ to g_0 can be defined by minimizing (5) over $\boldsymbol{\pi}$. Since $\nabla \log g(\boldsymbol{y})$ is linear in $\boldsymbol{\pi}$, Φ is quadratic in $\boldsymbol{\pi}$. Differentiating Φ with respect to $\boldsymbol{\pi}$, and setting the derivative to $\boldsymbol{0}$ yields the estimating equations

$$\boldsymbol{W}\boldsymbol{\pi} - \boldsymbol{d} = \boldsymbol{0},$$

where $\boldsymbol{W}$ and $\boldsymbol{d}$ have elements

$$w_{k_1,k_2} = \int \sum_{j=1}^{p-1} (\partial t_{k_1}(\boldsymbol{y})/\partial y_j)(\partial t_{k_2}(\boldsymbol{y})/\partial y_j) g_0(\boldsymbol{y}) d\boldsymbol{y},$$
$$k_1, k_2 = 1, \ldots, q, \quad (7)$$

$$d_k = \Delta t_k(\boldsymbol{y}) = \int \sum_{j=1}^{p-1} (\partial^2 t_k(\boldsymbol{y})/\partial y_j^2) g_0(\boldsymbol{y}) d\boldsymbol{y},$$
$$k = 1, \ldots, q. \quad (8)$$

The Laplacian in (8) arises after integration by parts in (5). Hence the the best-fitting value of $\boldsymbol{\pi}$ is

$$\boldsymbol{\pi} = \boldsymbol{W}^{-1}\boldsymbol{d}.$$

Given data $\boldsymbol{y}_i$, $i = 1, \ldots, n$, with elements y_{ij}, $j = 1, \ldots, p-1$, the integrals can be replaced by empirical averages to yield the estimating equations

$$\hat{\boldsymbol{W}}\boldsymbol{\pi} - \hat{\boldsymbol{d}} = \boldsymbol{0}, \quad (9)$$

where $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{d}}$ have elements

$$\hat{w}_{k_1,k_2} = \sum_{i=1}^{n}\sum_{j=1}^{p-1}(\partial t_{k_1}(\boldsymbol{y}_i)/\partial y_j)(\partial t_{k_2}(\boldsymbol{y}_i)/\partial y_j),$$
$$k_1, k_2 = 1, \ldots, q, \tag{10}$$
$$\hat{d}_k = \sum_{i=1}^{n}\sum_{j=1}^{p-1}(\partial^2 t_k(\boldsymbol{y}_i)/\partial y_j^2), \quad k = 1, \ldots, q. \tag{11}$$

Solving the estimating equations (9) yields the score matching estimator (SME)

$$\hat{\boldsymbol{\pi}} = \hat{\boldsymbol{W}}^{-1}\hat{\boldsymbol{d}}. \tag{12}$$

4.2 Additive log-ratio transformed compositions

To make use of this result for distributions on the simplex, it is helpful to make an additive log-ratio (ALR) transformation from $\boldsymbol{u} \in \Delta_0^{p-1}$ to $\boldsymbol{y} = (y_1, y_2, \ldots, y_{p-1})^\top \in \mathbb{R}^{p-1}$ where

$$y_j = \log(u_j/u_p), \quad j = 1, \ldots, p-1.$$

This transformation was popularized by Aitchison (1986). The logistic-normal distribution for $\boldsymbol{u}$, or equivalently the normal distribution for $\boldsymbol{y}$ has often been suggested as a model for compositional data (e.g., Aitchison 1986). However, it should be noted that the RPPI distribution has very different properties. In particular, we do not advocate the use of logistic-normal models in situations where zero or very-near-zero compositional components occur frequently. See Scealy and Welsh (2014) for relevant discussion and see Appendix A.3 (Supplementary Material) for further details on the choice of metric and transformation in score matching.

The transformed RPPI distribution has density proportional to

$$\exp\left(\frac{\exp(\boldsymbol{y})^\top \boldsymbol{A}_L \exp(\boldsymbol{y})}{(1+\sum_{k=1}^{p-1}\exp(y_k))^2}\right) \times \left(\frac{1}{1+\sum_{k=1}^{p-1}\exp(y_k)}\right)$$
$$\prod_{j=1}^{p-1}\left(\frac{\exp(y_j)}{1+\sum_{k=1}^{p-1}\exp(y_k)}\right)^{\beta_j+1} \tag{13}$$

with respect to Lebesgue measure $d\boldsymbol{y} = dy_1 \cdots dy_{p-1}$ on R^{p-1}, where

$$\exp(\boldsymbol{y}) = (\exp(y_1), \exp(y_2), \ldots, \exp(y_{p-1}))^\top$$

and we have used the constraints (4). The density (13) forms a full exponential family with canonical parameter vector

$$\boldsymbol{\pi} = (a_{11}, a_{22}, \ldots, a_{(p-1)(p-1)}, a_{12}, a_{13}, \ldots, a_{(p-2)(p-1)},$$
$$1+\beta_1, 1+\beta_2 \ldots, 1+\beta_{p-1})^\top$$

with $q = p(p-1)/2 + (p-1)$ parameters, where a_{ij} refers to the i, jth element of $\boldsymbol{A}_L$. The corresponding sufficient statistic, $\boldsymbol{t}(\boldsymbol{y}) = (\boldsymbol{t}_1(\boldsymbol{y})^\top, \boldsymbol{t}_2(\boldsymbol{y})^\top, \boldsymbol{t}_3(\boldsymbol{y})^\top)^\top$ will now be specified: $\boldsymbol{t}_1(\boldsymbol{y})$ is a $(p-1)$-vector with jth element

$$t_{1j}(\boldsymbol{y}) = \exp(2y_j)\Big/\left\{1+\sum_{k=1}^{p-1}\exp(y_k)\right\}^2,$$
$$j = 1, \ldots, p-1;$$

$\boldsymbol{t}_2(\boldsymbol{y})$ is a $(p-1)(p-2)/2$-vector with typical element

$$t_{2jk}(\boldsymbol{y}) = 2\exp(y_j)\exp(y_k)\Big/\left\{1+\sum_{\ell=1}^{p-1}\exp(y_\ell)\right\}^2,$$
$$1 \le j < k \le p-1;$$

and $\boldsymbol{t}_3(\boldsymbol{y})$ is a $(p-1)$-vector with typical element

$$t_{3j} = y_j - \log\left\{1+\sum_{k=1}^{p-1}\exp(y_k)\right\}, \qquad j = 1, \ldots, p-1.$$

The elements of $\hat{\boldsymbol{W}}$ and $\hat{\boldsymbol{d}}$ in (9) can be expressed in terms of linear combinations of powers and products of the u_{ij} which are the elements of the data vectors $\boldsymbol{u}_i, i = 1, 2, \ldots, n$; see the equations (20) and (21) in Appendix A.1 (Supplementary Material). We refer to the resulting score matching estimator as the ALR-SME. Note that there are no $\log(u_{ij})$ or ratios involving u_{ij} terms in (20) and (21). The ALR-SME estimator is very different to the standard maximum likelihood estimator for Aitchsion's logistic normal distribution. We use log-ratios as merely a device in the derivations to obtain our new score matching estimators (we are not actually transforming the data in the analysis since the PPI distribution is defined directly on the simplex).

4.3 Consistency

Next we state a consistency result for the ALR-SME when applied to the multinomial latent variable model. Let $\boldsymbol{x}_i$, $i = 1, \ldots, n$, be independent multinomial count vectors from different multinomial distributions, where the probability vectors $\boldsymbol{u}_i$ are taken independently from the RPPI model. Let $m_i = x_{i1} + \cdots + x_{ip}$ denote the total count from the ith multinomial vector. That is, we assume the conditional probability mass function of $\boldsymbol{x}_i = (x_{i1}, x_{i2}, \ldots, x_{ip})^\top$ given $\boldsymbol{u}_i$ is $f(\boldsymbol{x}_i|\boldsymbol{u}_i) = m_i!\prod_{j=1}^{p}\{u_{ij}^{x_{ij}}/x_{ij}!\}$, where the $\boldsymbol{u}_i = (u_{i1}, u_{i2}, \ldots, u_{ip})^\top$ are unobserved latent variables. This model is relevant for analysing microbiome data; see Sect. 6. Consider estimating the parameters $\boldsymbol{A}_L$ and $\beta_1, \beta_2, \ldots, \beta_{p-1}$ using the ALR-SME where the known proportions $\hat{\boldsymbol{u}}_i = \boldsymbol{x}_i/m_i$ are used as substitutes for the unknown

true compositions $\boldsymbol{u}_i$ for $i = 1, 2, \ldots, n$. Note that we do not need the extra restrictive conditions in part (III) Theorem 3 of Scealy and Wood (2023) for estimating $\boldsymbol{\beta}$. The proof of Theorem 1 below is given in Appendix A.2 (Supplementary Material).

Theorem 1 *Let $\hat{\boldsymbol{\pi}}$ denote the ALR-SME of $\boldsymbol{\pi}$ (12) based on the (unobserved) compositional vectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n$ and let $\hat{\boldsymbol{\pi}}^\dagger$ denote the ALR-SME of $\boldsymbol{\pi}$ based on the observed vectors of proportions $\hat{\boldsymbol{u}}_1, \ldots, \hat{\boldsymbol{u}}_n$. Assume that for some constants $C_1 > 0$ and $\alpha > 1$, $\inf_{i=1,\ldots,n} m_i \geq C_1 n^\alpha$.*

Then as $n \to \infty$

$$||n^{1/2}\left(\hat{\boldsymbol{\pi}}^\dagger - \hat{\boldsymbol{\pi}}\right)|| = o_p(1). \tag{14}$$

Theorem 1 above shows that $\hat{\boldsymbol{\pi}}^\dagger$ is asymptotically equivalent to $\hat{\boldsymbol{\pi}}$ to leading order. Note that Theorem 1 does not assume that the population latent variable distribution is a RPPI distribution, but if the RPPI model is correct then $\hat{\boldsymbol{\pi}}^\dagger$ and $\hat{\boldsymbol{\pi}}$ are both consistent estimators of $\boldsymbol{\pi}$ under the conditions of Theorem 1. Asymptotic normality of $\hat{\boldsymbol{\pi}}$ also follows directly from similar arguments to Scealy and Wood (2023). Theorem 1 applies even when the observed data has a large proportion of zeros. This has important implications for analysing microbiome count data with many zeros. Scealy and Wood (2023) were unable to use score matching based on the square root transformation to estimate $\boldsymbol{\beta}$ when analysing real microbiome data because the extra conditions needed for consistency did not look credible for the real data (there was an extra assumption needed on the marginal distributions of the components of the $\boldsymbol{u}_i$). Here, using the ALR-SME we are now able to estimate $\boldsymbol{\beta}$ directly using score matching. See Sects. 6 and 7 for further details.

It is also insightful to compare the ALR-SME to standard maximum likelihood estimation. The maximum likelihood estimator for $\boldsymbol{A}_L$ and $\boldsymbol{\beta}$ based on an iid sample from model (3) solves the estimating equation

$$\sum_{i=1}^{n}\left(\boldsymbol{t}^*(\boldsymbol{u}_i) - \frac{\partial}{\partial \boldsymbol{\pi}}\log\{c_2(\boldsymbol{A}_L, \boldsymbol{0}, \boldsymbol{\beta})\}\right) = \boldsymbol{0}, \tag{15}$$

where $\boldsymbol{t}^*(\boldsymbol{u})$ is defined at (19) in Appendix A.1 (Supplementary Material). Denote the maximum likelihood estimator of $\boldsymbol{\pi}$ by $\hat{\boldsymbol{\pi}}_{ML}$. The estimator $\hat{\boldsymbol{\pi}}_{ML}$ is difficult to calculate due to the intractable normalising constant c_2. Theorem 1 also does not hold for $\hat{\boldsymbol{\pi}}_{ML}$ due to the presence of the $\log(u_j)$ terms in $\boldsymbol{t}^*(\boldsymbol{u})$ which are unbounded at zero. That is, we cannot simply replace $\boldsymbol{u}_i$ by $\hat{\boldsymbol{u}}_i$ within (15) to obtain a consistent estimator for the multinomial latent variable model. This is a major advantage of the ALR-SME because it leads to computationally simple and consistent estimators, whereas $\hat{\boldsymbol{\pi}}_{ML}$ with the latent variables $\boldsymbol{u}_i$, $i = 1, 2, \ldots, n$ each replaced with $\hat{\boldsymbol{u}}_i$ is inconsistent and computationally not tractable.

4.4 Comments on SME

Score matching estimation has been defined here for probability densities whose support is all of $\mathbb{R}^d$. This construction can be extended in various ways, and we mention two possibilities here that are relevant for compositional data.

First, the unbounded region $\mathbb{R}^{p-1}$ in (5) can be replaced by a bounded region such as Δ_0^{p-1}. However, there is a price to pay. The integration by parts which underlies the Laplacian term in (8) now includes boundary terms. Scealy and Wood (2023) introduced a weighting function which vanishes on the boundary of the simplex. The effect of this weighting function is to eliminate the boundary terms. However, the weighting function also lessens the contribution of data near the boundary to the estimating equations.

Second, the Hyvarinen divergence in (5) implicitly uses a Riemannian metric in $\mathbb{R}^{p-1}$, namely Euclidean distance. Other choices of Riemannian metric lead to different estimators. Some comments on these choices in the context of compositional data are discussed in Appendix A.3 (Supplementary Material).

5 An ALR-SME that is resistant to outliers

Although the simplex is a bounded space, outliers/influential points can still occur when the majority of the data is highly concentrated, or equivalently has low dispersion, in certain regions of the simplex. In the case of microbiome data (see Sect. 6), there are a small number of abundant components which have low concentration (e.g. *Actinobacteria* and *Proteobacteria*) and these components should be fairly resistant to outliers. However, the components *Spirochaetes*, *Verrucomicrobia*, *Cyanobacteria/Chloroplast* and *TM7* are highly concentrated at or near zero and any large values away from zero can be influential. For the highly concentrated microbiome components distributed close to zero, these marginally look to be approximately gamma or generalised gamma distributed; see Figs. 1 and 2 in Sect. 6. Hence there is a need for the Dirichlet component of the density in the RPPI model (3).

We now develop score matching estimators for the RPPI model (3) that are resistant to outliers. Assume that the first k^* components of $\boldsymbol{u}$ are highly concentrated near zero where it is expected that possibly $\beta_1 < 0, \beta_2 < 0, \ldots, \beta_{k^*} < 0$. The remaining components $u_{k^*+1}, u_{k^*+2}, \ldots, u_p$ are assumed to have relatively low concentration. By low concentration we mean moderate to high variance and by high concentration we mean small variance. The robustification which follows is only relevant for highly concentrated components near zero which is why we are distinguishing between the different cases. See Scealy and Wood (2021) for further discussion on standardised bias robustness under high concentration which

is relevant to all compact sample spaces including the simplex. When $k^* < p-1$ partition

$$\boldsymbol{A}_L = \begin{pmatrix} \boldsymbol{A}_{KK} & \boldsymbol{A}_{KR} \\ \boldsymbol{A}_{RK} & \boldsymbol{A}_{RR} \end{pmatrix},$$

where $\boldsymbol{A}_{KK}$ is a $k^* \times k^*$ matrix, $\boldsymbol{A}_{KR}$ is a $k^* \times (p-1-k^*)$ matrix, $\boldsymbol{A}_{RK}$ is a $(p-1-k^*) \times k^*$ matrix and $\boldsymbol{A}_{RR}$ is a $(p-1-k^*) \times (p-1-k^*)$ matrix. When $k^* = p-1$ then $\boldsymbol{A}_L = \boldsymbol{A}_{KK}$. The (unweighted) estimating equations for the ALR-SME are given by (9) and can be written slightly more concisely as

$$\boldsymbol{0} = \frac{1}{n}\sum_{i=1}^{n} \left(\boldsymbol{W}_1(\boldsymbol{u}_i)\boldsymbol{\pi} - \boldsymbol{d}_1(\boldsymbol{u}_i)\right), \tag{16}$$

where the elements of $\boldsymbol{W}_1(\boldsymbol{u}_i)$ and $\boldsymbol{d}_1(\boldsymbol{u}_i)$ are functions of $\boldsymbol{u}_i$ and are defined at equations (20) and (21) in Appendix A.1 (Supplementary Material) for the RPPI model and are given in a more general form though Eqs. (9)–(11).

Windham (1995) approach to creating robustified estimators is to use weights which are proportional to a positive power of the probability density function. The intuition behind this approach is that outliers under a given distribution will typically have small likelihood and hence a small weight, whereas observations in the central region of the distribution will tend to have larger weights. The Windham (1995) method is an example of a density-based minimum divergence estimator, but with the advantage that the normalising constant in the density does not need to be evaluated in order to apply it. See Windham (1995), Basu et al. (1998), Jones, et al. (2001), Choi et al. (2000), Ribeiro and Ferrari (2020), Kato and Eguchi (2016) and Saraceno et al. (2020) for further discussion and insights. In the setting of the RPPI model for compositional data, there is a choice to be made between the probability densities to use in the weights, that is to use (3) or (13), or in other words should we choose the measure $d\boldsymbol{u}$ or $d\boldsymbol{y}$. We prefer $d\boldsymbol{u}$ because $d\boldsymbol{y}$ places zero probability density at the simplex boundary and thus always treats zeros as outliers which is not a good property with data concentrated near the simplex boundary.

For the RPPI distribution, taking a power of the density (3) is a bad idea because for those β_j which are negative the weights will diverge to infinity as u_j tends to 0. To circumvent this issue we only use the exp-quadratic factor in (3) to define the weights. This choice of weighting function is a compromise between wanting the weight of an observation to be smaller if the probability density is smaller and needing to avoid infinite weights on the boundary of the simplex. In fact, typically $\boldsymbol{u}_K^\top \boldsymbol{A}_{KK}\boldsymbol{u}_K$, where $\boldsymbol{u}_K = (u_1, u_2, \ldots, u_{k^*})^\top$, is highly negative whenever $\boldsymbol{u}$ has a large value in any of the components that are highly concentrated near zero in distribution. It is thus sufficient to use just the $\exp(\boldsymbol{u}_K^\top \boldsymbol{A}_{KK}\boldsymbol{u}_K)$ factor of (3) in the weights (the influence function in Theorem 2 below confirms this behaviour). Including all elements of $\boldsymbol{A}_L$ in the weights leads to a large loss in efficiency, so the weights in our robustified ALR-SME estimator are $\exp(c\boldsymbol{u}_{i,K}^\top \boldsymbol{A}_{KK}\boldsymbol{u}_{i,K})$, $i = 1, \ldots, n$, where $\boldsymbol{u}_{i,K} = (u_{i1}, u_{i2}, \ldots, u_{ik^*})^\top$. The weighted form of estimating Eq. (16) is then

$$\sum_{i=1}^{n} \exp\left(c\boldsymbol{u}_{i,K}^\top \boldsymbol{A}_{KK}\boldsymbol{u}_{i,K}\right)\left(\boldsymbol{W}_1(\boldsymbol{u}_i)\boldsymbol{H}\boldsymbol{\pi} - \boldsymbol{d}_1(\boldsymbol{u}_i)\right) = \boldsymbol{0}, \tag{17}$$

where $\boldsymbol{H}$ is a $q \times q$ diagonal matrix with diagonal elements either equal to $c+1$ or 1 (the elements corresponding to the parameters $\boldsymbol{A}_{KK}$ are $c+1$ and the rest are 1). The estimating Eq. (17) has a very simple form here (i.e. given the weights, the estimating equations are linear in $\boldsymbol{\pi}$), whereas the version based on the maximum likelihood estimator does not have such a nice linear form leading to a much more complicated influence function calculation and its interpretation (e.g. Jones, et al. 2001).

An algorithm similar to that in Windham (1995) can be used to solve (17) and involves iteratively solving weighted versions of the score matching estimators. In summary this algorithm is

1. Set $r = 1$ and initialise the parameters: $\hat{\boldsymbol{\beta}}^{(0)}$ and $\hat{\boldsymbol{A}}_L^{(0)}$ (i.e. choose starting values such as the unweighted ALR-SME). Then repeat steps 2–5 until convergence.
2. Calculate the weights $\tilde{w}_i = \exp\left(c\boldsymbol{u}_{i,K}^\top \hat{\boldsymbol{A}}_{KK}^{(r-1)}\boldsymbol{u}_{i,K}\right)$ for $i = 1, 2, \ldots, n$ and normalise the weights so that the weights sum to 1 across the sample. Also calculate the additional tuning constants $\boldsymbol{d}_\beta = -c\hat{\boldsymbol{\beta}}^{(r-1)}$, $\boldsymbol{d}_{A_1} = -c\hat{\boldsymbol{A}}_{RR}^{(r-1)}$, $\boldsymbol{d}_{A_2} = -c\hat{\boldsymbol{A}}_{RK}^{(r-1)}$ and $\boldsymbol{d}_{A_3} = -c\hat{\boldsymbol{A}}_{KR}^{(r-1)}$.
3. Calculate weighted score matching estimates. That is, replace all sample averages with weighted averages using the normalised weights $\tilde{w}_i$ calculated in step 2. Denote the resulting estimates as $\tilde{\boldsymbol{\beta}}^{(r)}$ and $\tilde{\boldsymbol{A}}_L^{(r)}$.
4. The estimates in step 3 are biased and we need to do the following bias correction:

$$\hat{\boldsymbol{\beta}}^{(r)} = \frac{\tilde{\boldsymbol{\beta}}^{(r)} - \boldsymbol{d}_\beta}{c+1}, \quad \text{and} \quad \hat{\boldsymbol{A}}_{KK}^{(r)} = \frac{\tilde{\boldsymbol{A}}_{KK}^{(r)}}{c+1}$$

and

$$\hat{\boldsymbol{A}}_{RR}^{(r)} = \frac{\tilde{\boldsymbol{A}}_{RR}^{(r)} - \boldsymbol{d}_{A_1}}{c+1}, \quad \hat{\boldsymbol{A}}_{RK}^{(r)} = \frac{\tilde{\boldsymbol{A}}_{RK}^{(r)} - \boldsymbol{d}_{A_2}}{c+1},$$

$$\text{and} \quad \hat{\boldsymbol{A}}_{KR}^{(r)} = \frac{\tilde{\boldsymbol{A}}_{KR}^{(r)} - \boldsymbol{d}_{A_3}}{c+1}.$$

This correction is simple because the model is an exponential family; see Windham (1995) for further details.

5. $r \to r+1$

Step 4 in this new robust score matching algorithm above is similar to applying the inverse of τ_c in Windham (1995). The tuning constants $\boldsymbol{d}_\beta$, $\boldsymbol{d}_{A_1}$, $\boldsymbol{d}_{A_2}$ and $\boldsymbol{d}_{A_3}$ are required due to our use of a factor of the density in the weights.

This modified version of Windham (1995) method is particularly useful when any β_j's are negative in order to avoid infinite weights at zero. When the data is concentrated in the simplex interior (i.e. we expect $\beta_j > 0$, $j = 1, 2, \ldots, p$) then the model density is bounded and we can apply the Windham (1995) method without modification, although efficiency gains may be possible from using well-chosen factors of the model density.

In order to complete the description of the robustified ALR-SME, we need to choose the robustness tuning constant c. In related settings (Kato and Eguchi 2016) use cross validation and Saraceno et al. (2020) calculate theoretical optimal values for a Gaussian linear mixed model. Basak et al. (2021) report that choosing the optimal tuning constant is challenging in general when choosing density power divergence tuning parameters. We agree with the view of (Muller and Welsh 2005, page 1298) that choice of model selection criteria or estimator selector criterion should be independent of the estimation method, otherwise we may excessively favour particular estimators. This is an issue with the Kato and Eguchi (2016) method which is based on an arbitrary choice of divergence which could favor the optimal estimator under that divergence. Instead we use a simulation based method to choose c; see Sect. 6. We also need to decide on the value of k^*; see Sect. 6 for a guide.

We next examine the theoretical properties of our new robustified estimator. The proof of Theorem 2 below is given in Appendix A.4 (Supplementary Material). Let $\mathcal{F}$ denote the set of probability distributions on the unit simplex $\Delta^{p-1} \subset \mathbb{R}^p$, where $p \geq 3$. Let F_0 denote the population probability measure for a single observation from Δ^{p-1} and write δ_z for the degenerate distribution on Δ^{p-1} which places unit probability on $z \in \Delta^{p-1}$. Consider the ALR-SME functional $\boldsymbol{\theta} : \mathcal{F} \to \Theta \subseteq \mathbb{R}^q$. It is assumed that $\boldsymbol{\theta}$ is well defined for all z at $(1-\lambda)F_0 + \lambda\delta_z$ provided $\lambda \in (0, 1)$ is sufficiently small. Then the influence function for $\boldsymbol{\theta}$ and $F_0 \in \mathcal{F}$ at z is defined by

$$\mathbf{IF}_{\boldsymbol{\theta}; F_0}(z) = \lim_{\lambda \to 0} \frac{1}{\lambda} \left(\boldsymbol{\theta} \left\{ (1-\lambda)F_0 + \lambda\delta_z \right\} - \boldsymbol{\theta}\left(F_0\right) \right).$$

Theorem 2 *Suppose that the population distribution F_0 on Δ^{p-1} is absolutely continuous with respect to Lebesgue measure on Δ^{p-1}. Also assume that $k^* = p-1$ for exposition simplicity which implies that all of the first $p-1$ components are concentrated at/near zero. (The proof for the case of $k^* < p-1$ is similar and is not presented here.) Then*

$$\mathbf{IF}_{\boldsymbol{\pi}; F_0}(z) = -\left(\boldsymbol{G}(\boldsymbol{\pi}_0)\right)^{-1} \exp\left(c\boldsymbol{t}^{(a)}(z)^\top \boldsymbol{\pi}_0 \right) \left\{ \boldsymbol{W}_1(z)\boldsymbol{H}\boldsymbol{\pi}_0 - \boldsymbol{d}_1(z) \right\},$$

where $\boldsymbol{\pi}_0$ is the solution to the population estimating equation corresponding to (17) (see equation (27) in Appendix A.4 (Supplementary Material)) and the functions $\boldsymbol{G}(\boldsymbol{\pi}_0)$ and $\boldsymbol{t}^{(a)}(z)$ are defined in Appendix A.4 (Supplementary Material).

The functions $\boldsymbol{t}^{(a)}(z)$, $\boldsymbol{W}_1(z)$ and $\boldsymbol{d}_1(z)$ contain linear combinations of low order polynomial products, for example terms like $z_1^{r_1} z_2^{r_2} z_3^{r_3}$, where $r_1 \geq 0, r_2 \geq 0, r_3 \geq 0$ and $r_1 + r_2 + r_3$ is small. Therefore the above influence function is always bounded for all $z \in \Delta^{p-1}$ including for any points on the simplex boundary, even when $c = 0$. The $\boldsymbol{t}^{(a)}(z)^\top \boldsymbol{\pi}_0$ in Theorem 2 is equal to $z_K^\top \boldsymbol{A}_{KK} z_K$, where $z_K = (z_1, z_2, ..., z_{k^*})^\top$. For many PPI models, $z_K^\top \boldsymbol{A}_{KK} z_K = \boldsymbol{t}^{(a)}(z)^\top \boldsymbol{\pi}_0 < 0$, which means that for the components of $\boldsymbol{u}$ that are highly concentrated near zero in distribution, any large value away from zero in these components will be down-weighted and have less influence on the estimator. This leads to large efficiency gains in both the contaminated and uncontaminated cases. See Sect. 7 for further details.

It is useful to compare Theorem 2 with the influence function for $\hat{\boldsymbol{\pi}}_{ML}$. The maximum likelihood estimator is a standard M-estimator with influence function of the form

$$-\boldsymbol{B}^{-1}\left(\boldsymbol{t}^*(z) - \frac{\partial}{\partial \boldsymbol{\pi}_0} \log c_2 \right), \tag{18}$$

where $\boldsymbol{B}$ is a matrix function of the model parameters (e.g. Maronna et al. 2006, page 71) and $\boldsymbol{t}^*(z)$ is defined at (19) in Appendix A.1 (Supplementary Material). The vector $\boldsymbol{t}^*(z)$ contains the functions $\log(z_1)$, $\log(z_2)$, ..., $\log(z_{p-1})$ and the influence function (18) is unbounded if any z_j approaches 0, $j = 1, 2, \ldots, p-1$. Therefore maximum likelihood estimation for the PPI model is highly sensitive to zeros. Maximum likelihood estimation is also highly sensitive to zeros for the gamma, Beta, Dirichlet and logistic normal distributions for similar reasons.

6 Microbiome data analysis

Microbiome data is challenging to analyse due to the presence of high right skewness, outliers and zeros in the marginal distributions of the bacterial species (e.g. Li 2015; He et al. 2021). Typically microbiome count data is either modelled using a multinomial model with latent variables (e.g. Li 2015; Martin et al. 2018; Zhang and Lin 2019) or the sample counts

are normalised and treated as approximately continuous data since the total counts are large (e.g. Cao et al. 2019; He et al. 2021). Here we analyse real microbiome count data by fitting a RPPI multinomial latent variable model using the normalised microbiome counts as estimates of the latent variables; see Sect. 4.3.

In this section we analyse a subset of the longitudinal microbiome dataset obtained from a study carried out in a helminth-endemic area in Indonesia (Martin et al. 2018). In summary, stool samples were collected from 150 subjects in the years 2008 (pre-treatment) and in 2010 (post-treatment). The 16 s rRNA gene from the stool samples was processed and resulted in counts of 18 bacterial phyla. Whether or not an individual was infected by helminth was also determined at both time points. We restricted the analysis to the year 2008 for individuals infected by helminths which resulted in a sample size of $n = 94$, and we treated these individuals as being independent.

Martin et al. (2018) analysed the five most prevalent phyla and pooled the remaining categories. Scealy and Wood (2023) analysed a different set of four phyla including two with a high number of zeros and pooled the remaining categories. Here for demonstrative purposes we will first analyse the same data components as in Scealy and Wood (2023) with the $p = 5$ components representing *TM7*, *Cyanobacteria/Chloroplast*, *Actinobacteria*, *Proteobacteria* and *pooled*. The percentage of zeros in each category are 38%, 41%, 0%, 0% and 0% respectively. Call this *Dataset1*. Then for demonstrative purposes we will also analyse a second dataset with $p = 5$ denoted as *Dataset2* which contains the components *Spirochaetes*, *Verrucomicrobia*, *Cyanobacteria/Chloroplast*, *TM7* and *pooled*. The percentage of zeros in each category for *Dataset2* are 77%, 75%, 41%, 38% and 0% respectively. Let x_{ij}, $i = 1, 2, \ldots, 94$ and $j = 1, 2, 3, 4, 5$ represent the sample counts for a given dataset with total count $m_i = 2000$. The estimated sample proportions were calculated as follows: $\hat{u}_{ij} = x_{ij}/m_i$, where $i = 1, 2, \ldots, 94$ and $j = 1, 2, 3, 4, 5$.

Figure 1 is similar to (Scealy and Wood 2023, Figure 3), the only difference being that we have now included the two large proportions in $\hat{u}_{i1}$ and $\hat{u}_{i2}$ which were deleted by Scealy and Wood (2023) prior to their analysis because they identified them as outliers. The estimates of β_1 and β_2 in Scealy and Wood (2023) were negative and close to -1 and the components $\hat{u}_{i1}$ and $\hat{u}_{i2}$ are highly concentrated mostly near zero. The components $\hat{u}_{i3}$ and $\hat{u}_{i4}$ for this dataset have low concentration. Therefore it makes sense here to choose $k^* = 2$.

Figure 2 contains histograms of the sample proportions in *Dataset2*. The first four components are highly concentrated near zero and we would expect that β_1, β_2, β_3 and β_4 are negative. Therefore it makes sense to choose $k^* = 4$ for this dataset.

6.1 Choice of tuning constant *c*

For each dataset we let c range over a grid from 0 up to 1.5 and we fitted the model for each value of c. We simulated a single large sample of size $R = 10{,}000$ under the fitted model (3) for each value of c and rounded the simulated data as follows: $\hat{u}^r_{ij} = \text{round}(\hat{u}_{ij} m_i)/m_i$, where $\hat{u}_{ij}$ denotes the simulated proportion under the fitted RPPI model for $i = 1, 2, \ldots, R$. This mimics the discreteness in the data; see Scealy and Wood (2023) Section 7. Then we compared the simulated proportions with the true sample proportions. Similar to the view of Muller and Welsh (2005) (page 1298) we are interested in fitting the core of the data and we are not specifically interested in fitting in the upper tails which is where outliers can occur in this setting. This means we need to choose a criterion that is not sensitive to the upper tail. When comparing the simulated proportions with the true sample proportions we deleted all observations above the 95% quantile cutoff in the marginal distribution proportions.

For each dataset, c was chosen to give a compromise between fitting all components to give a small value of the Kolmogorov–Smirnov test statistic and keeping variation in the weights as small as possible to preserve efficiency. See Table 1 for the chosen values of c for each dataset. Note that the p value for *Proteobacteria* is quite small. This is not surprising as this was also the worst fitting component in Table 5 in Scealy and Wood (2023) in their analysis.

Table 2 contains the parameter estimates for *Dataset1*. The standard errors (SE) were calculated using a parametric bootstrap by simulating under the fitted multinomial latent variable model. Use of robust non-parametric bootstrap methods such as those in Muller and Welsh (2005) and Salibian-Barrera et al. (2008) is challenging here due to the large numbers of zeros in the data and for that reason we prefer the parametric bootstrap. The parametric bootstrap SE estimates are expected to be a little larger than the ones in Scealy and Wood (2023) since they used asymptotic standard errors which tended to be a slight underestimation as shown in their simulation study. The new robustified ALR-SME of $\boldsymbol{\beta}$ are reasonably close to the simulation/grid search estimates in Scealy and Wood (2023). Interestingly, β_3 is not significantly different from zero; Scealy and Wood (2023) set this parameter to zero based on visual inspection of plots. As expected the estimates of β_1 and β_2 are negative and are highly significant. We no longer need to treat $\boldsymbol{\beta}$ as a tuning constant and we can now estimate its standard errors which is an advantage of the new robustified ALR-SME method.

Table 3 contains the parameter estimates for *Dataset2*. Note that we cannot apply the score matching estimators of Scealy and Wood (2023) to this dataset as every datapoint has a component equal to zero, which means the manifold boundary weight functions in Scealy and Wood (2023) evaluate to zero. However, our new robustified ALR-SME method can

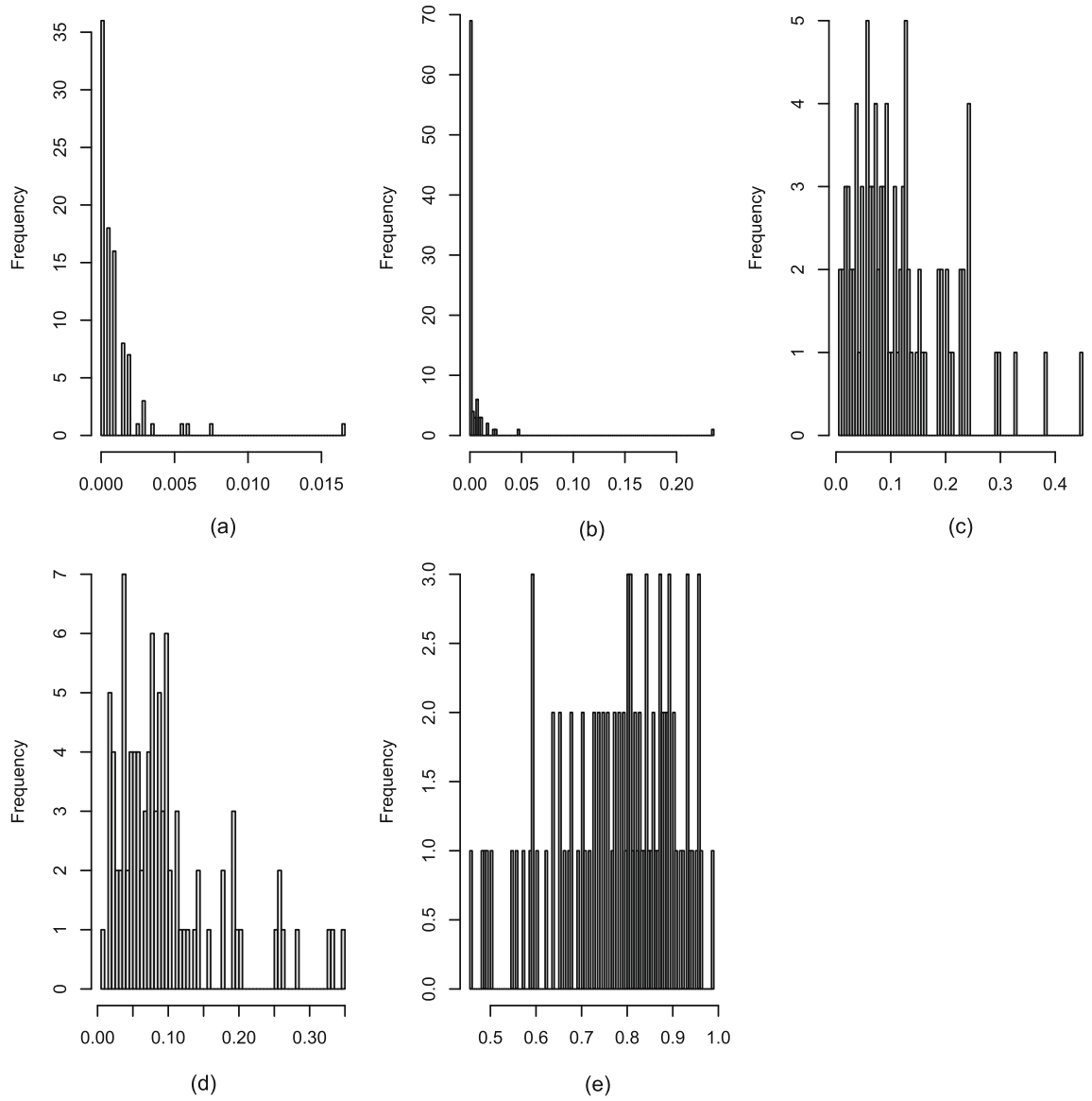

Fig. 1 *Dataset1* histograms of sample proportions $\hat{u}_{ij}$, $j = 1, 2, 3, 4, 5$

handle this dataset with massive numbers of zeros. Again, we used the parametric bootstrap to calculate the standard error estimates. As expected all of β_1, β_2, β_3 and β_4 are negative and highly significantly different from zero. The $\boldsymbol{A}_L$ parameter estimates are all insignificant. So for this dataset perhaps a Dirichlet model might be appropriate. Note that this is not surprising due to the massive numbers of zeros and relatively small sample size; there is not much information available to estimate $\boldsymbol{A}_L$.

7 Simulation

In this section we explore the properties of the new robustified ALR-SME and we compare them with the score matching estimator of Scealy and Wood (2023) based on the manifold boundary weight function $\min(u_1, u_2, \ldots u_p, a_c^2)$, with various choices of their tuning constant $a_c \in [0, 1]$. We consider eight different simulation settings and in each case we simulated $R = 1000$ samples and for each sample we calculated multiple different score matching estimates and calculated estimated root mean squared errors (RMSE). The eight different simulation settings are now described. We focus on dimension $p = 5$ only.

Simulation 1: The model is the continuous RPPI model (3) with $\boldsymbol{b}_L = \boldsymbol{0}$ and $\beta_5 = 0$ fixed (not estimated). We set $\beta_1 = -0.80$, $\beta_2 = -0.85$, $\beta_3 = 0$, $\beta_4 = -0.2$ and $\boldsymbol{A}_L$ is equal to the parameter estimates given in Table 3 of Scealy and Wood (2023). This model was the best fitting model for *Dataset1* in Scealy and Wood (2023) after they deleted two

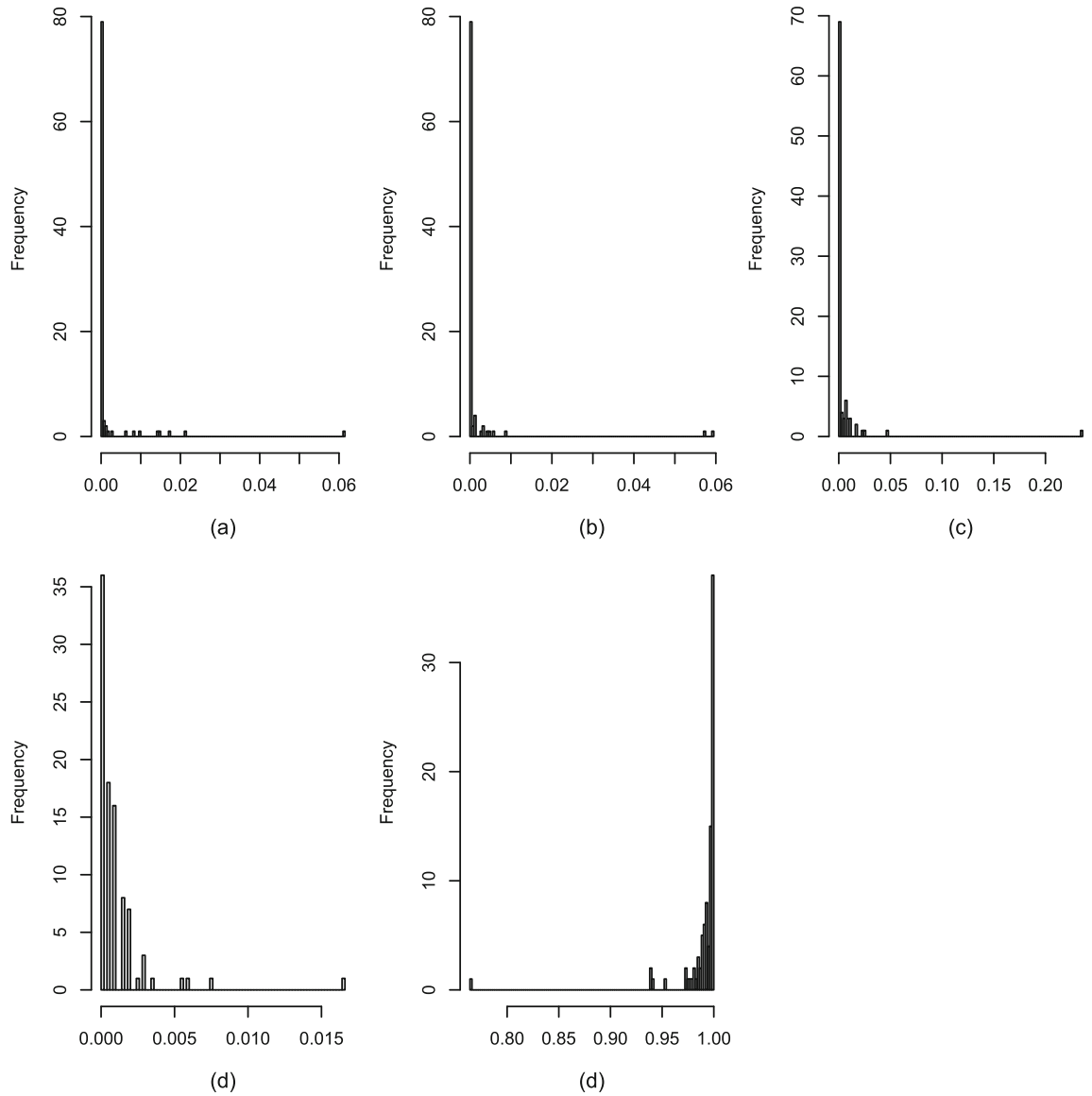

Fig. 2 *Dataset2* histograms of sample proportions $\hat{u}_{ij}$, $j = 1, 2, 3, 4, 5$

outliers. We set the sample size to $n = 92$ which is consistent with Scealy and Wood (2023). For this model we calculated the new ALR-SME, which is denoted as $c = 0$ in Table 4. We also calculated the new robustified ALR-SME with tuning constants set to $c = 0.01$ and $c = 0.7$. Then we calculated the score matching estimators of Scealy and Wood (2023) with tuning constants set to $a_c = 0.01$, $a_c = 0.000796$ and $a_c = 1$; see columns 5, 6 and 8 in Table 4. Note that $\boldsymbol{\beta}$ is not estimated in columns 5, 6 and 8 and is treated as known and set equal to the true $\boldsymbol{\beta}$ (Scealy and Wood 2023 treated $\boldsymbol{\beta}$ as a tuning constant in their real data application). The 7th column in Table 4 denoted by a_c given $\hat{\boldsymbol{\beta}}$ is a hybrid two step estimator. That is, first we calculated the estimate of $\boldsymbol{A}_L$ and $\boldsymbol{\beta}$ using the robustified ALR-SME with $c = 0.7$, then in the second step we updated the $\boldsymbol{A}_L$ estimate conditional on the robust $\boldsymbol{\beta}$ estimate using the Scealy and Wood (2023) estimator with $a_c = 0.000796$.

Simulation 2: The model is the multinomial latent variable model with $m_i = 2000$ for $i = 1, 2, \ldots, n$ with $n = 92$. The latent variable distribution is set equal to the same RPPI model used in *Simulation 1*. We calculated the same score matching estimators as in *Simulation 1* but instead of using $\boldsymbol{u}_i$ in estimation we plugged in the discrete simulated proportions $\hat{\boldsymbol{u}}_i = \boldsymbol{x}_i/m_i$.

Simulation 3: The same setting as *Simulation 1* except we replace 5.4% of the observations with the outlier $\boldsymbol{u}_i = (0.4, 0.4, 0, 0, 0.2)^\top$.

Simulation 4: The same setting as *Simulation 2* except we replace 5.4% of the observations with the outlier $\hat{\boldsymbol{u}}_i = \boldsymbol{x}_i/m_i = (0.4, 0.4, 0, 0, 0.2)^\top$.

Table 1 Kolmogorov–Smirnov test results (upper 5% quantile removed)

Dataset1	$c = 0.7$	*Dataset2*	$c = 1.25$
TM7	0.14 (p value = 0.07)	*Spirochaetes*	0.088 (p value = 0.50)
Cyanobacteria/chloroplast	0.13 (p value = 0.11)	*Verrucomicrobia*	0.053 (p value = 0.96)
Actinobacteria	0.11 (p value = 0.24)	*Cyanobacteria/chloroplast*	0.058 (p value = 0.93)
Proteobacteria	0.16 (p value = 0.021)	*TM7*	0.058 (p value = 0.93)

Table 2 Parameter estimates and standard errors for *Dataset1*

Parameter	Estimate	Estimate/SE	Parameter	Estimate	Estimate/SE
a_{11}	− 60529.900	− 0.382	a_{23}	73.34150	0.254
a_{12}	12432.5000	0.255	a_{24}	84.06690	0.356
a_{13}	430.21700	0.546	a_{33}	− 22.03470	− 2.62
a_{14}	− 411.20500	− 0.466	a_{34}	− 9.08582	− 0.861
a_{22}	− 4934.7000	− 0.303	a_{44}	− 22.16840	− 2.33
β_1	− 0.770598	− 10.8	β_3	− 0.079014	− 0.408
β_2	− 0.870535	− 15.0	β_4	− 0.149064	− 0.878

Table 3 Parameter estimates and standard errors for *Dataset2*

Parameter	Estimate	Estimate/SE	Parameter	Estimate	Estimate/SE
a_{11}	− 141.924	− 0.033	a_{23}	− 38106.8	− 0.668
a_{12}	− 16586	− 0.170	a_{24}	11709.2	0.062
a_{13}	− 5877.63	− 0.265	a_{33}	− 5184.47	− 0.071
a_{14}	− 11524.5	− 0.081	a_{34}	8260.35	0.025
a_{22}	− 9856.69	− 0.107	a_{44}	− 216660.00	− 0.137
β_1	− 0.904976	− 27.7	β_3	− 0.740065	− 10.4
β_2	− 0.909160	− 28.4	β_4	− 0.464586	− 4.60

Simulation 5: The model is the continuous RPPI model (3) with $\boldsymbol{b}_L = \mathbf{0}$ and $\beta_5 = 0$ fixed (not estimated), and remaining parameters set equal to the values given in Table 3. For this model we calculated the new ALR-SME which is denoted as $c = 0$ in Table 6. We also calculated the new robustified ALR-SME with tuning constants set to $c = 0.01$, $c = 0.25$, $c = 0.5$, $c = 0.75$, $c = 1$ and $c = 1.25$. The sample size is the same as *Dataset2* which is $n = 94$.

Simulation 6: The model is the multinomial latent variable model with $m_i = 2000$ for $i = 1, 2, \ldots, n$ with $n = 94$. The latent variable distribution is set equal to the same RPPI model used in *Simulation 5*. We calculated the same score matching estimators as in *Simulation 5* but instead of using $\boldsymbol{u}_i$ in estimation we plugged in the discrete simulated proportions $\hat{\boldsymbol{u}}_i = \boldsymbol{x}_i / m_i$.

Simulation 7: The same setting as *Simulation 5* except we replace 5.3% of the observations with the outlier $\boldsymbol{u}_i = (0.4, 0.3, 0.2, 0.1, 0)^\top$.

Simulation 8: The same setting as *Simulation 6* except we replace 5.3% of the observations with the outlier $\hat{\boldsymbol{u}}_i = \boldsymbol{x}_i / m_i = (0.4, 0.3, 0.2, 0.1, 0)^\top$.

We now discuss the simulation results in Tables 4 and 5. These models are motivated from *Dataset1*. *Dataset1* has two components that are highly right skewed concentrated near zero and three components with low concentration; see Fig. 1. The RMSE's are of a similar order when comparing the first half of Table 4 with the corresponding cells in the second half of Table 4 and similarly this also occurs within Table 5. This is not surprising because m_i is large compared with n and the approximation $\hat{\boldsymbol{u}}_i$ for $\boldsymbol{u}_i$ is reasonable. Hence the estimates are insensitive to the large numbers of zeros in $\hat{u}_{i1}$ and $\hat{u}_{i2}$. When comparing Table 4 with 5 most of the corresponding cells are fairly similar apart from $c = 0$ which has huge RMSE's in Table 5. The robustified ALR-SME with $c > 0$ are clearly resistant to the outliers, whereas the unweighted estimator with $c = 0$ does not exhibit good resistance to outliers. Interestingly, the most efficient estimate of $\boldsymbol{\beta}$ is given by $c = 0.7$ even when there are no outliers. The efficiency gains for β_1 and β_2 are substantial when comparing $c = 0$ (no weights) to $c = 0.7$. So the message here is that the weighted version of the ALR-SME is valuable for improving efficiency for estimating the components of $\boldsymbol{\beta}$ that are negative and close to -1. In the continuous case arguably the Scealy and Wood

Table 4 Simulation results *Dataset1* RMSE's

Parameter	$c = 0$	$c = 0.01$	$c = 0.7$	$a_c = 0.01$	$a_c = 0.000796$	a_c given $\hat{\boldsymbol{\beta}}$	$a_c = 1$
Simulation 1: RPPI model (continuous)							
a_{11}	95,000	88,600	99,000	95,200	81,300	81,900	293,000
a_{22}	6920	6440	7720	6700	5430	5460	18,100
a_{33}	21.9	20.5	18	27.8	20.8	21.5	104
a_{44}	18.7	17.7	16.6	22.6	16.2	16.6	75.6
a_{12}	14,800	14,200	22,500	14,800	12,400	12,500	42,900
a_{13}	1240	1150	974	1390	1170	1190	4600
a_{14}	1010	963	879	953	807	833	2690
a_{23}	276	262	272	285	242	250	706
a_{24}	294	271	238	306	248	251	855
a_{34}	15.3	14.7	14.7	16.4	12.4	13	53
β_1	0.185	0.173	0.0685	–	–	0.0685	–
β_2	0.142	0.133	0.0581	–	- -	0.0581	–
β_3	0.264	0.255	0.245	–	–	0.245	–
β_4	0.22	0.214	0.202	–	–	0.202	–
Simulation 2: multinomial latent variable model (discrete)							
a_{11}	57,700	55,600	85,700	97,800	97,800	98,200	253,000
a_{22}	5540	5200	6240	8890	8890	8910	18,900
a_{33}	16.5	15.7	15.1	36.7	36.7	37.1	100
a_{44}	16.7	15.9	15.3	36.6	36.6	36.8	84.7
a_{12}	12,200	11,800	14,300	17,600	17,600	17,700	43,400
a_{13}	846	807	844	1480	1480	1490	4060
a_{14}	876	842	803	1200	1200	1210	2530
a_{23}	273	260	241	389	389	392	804
a_{24}	248	231	214	417	417	418	882
a_{34}	14.2	13.7	13.9	22.4	22.4	22.6	51.3
β_1	0.162	0.153	0.0631	–	–	0.0631	–
β_2	0.137	0.129	0.0543	–	–	0.0543	–
β_3	0.25	0.243	0.235	–	–	0.235	–
β_4	0.214	0.209	0.197	–	–	0.197	–

(2023) estimator with $a_c = 0.000796$ is the most efficient for estimating $\boldsymbol{A}_L$, whereas in the discrete multinomial case the $c = 0.7$ estimator is arguably the most efficient for $\boldsymbol{A}_L$.

We now consider the simulation results in Tables 6 and 7. These models are motivated from *Dataset2*. *Dataset2* has four components that are highly right skewed concentrated near zero and one component highly concentrated near one; see Fig. 2. The Scealy and Wood (2023) estimators are omitted because their manifold boundary weight functions evaluate to zero, or very close to zero, for most datapoints in most simulated samples. The RMSE's are roughly of a similar order when comparing the first half of Table 6 with the corresponding cells in the second half of Table 6 and similarly this also occurs within Table 7. This is not surprising because m_i is large compared with n and the approximation $\hat{\boldsymbol{u}}_i$ for $\boldsymbol{u}_i$ is reasonable. Hence the estimates were insensitive to the large numbers of zeros in $\hat{u}_{i1}$, $\hat{u}_{i2}$, $\hat{u}_{i3}$ and $\hat{u}_{i4}$. When comparing Table 6 with 7 most of the corresponding cells are fairly similar apart from $c = 0$ which has huge RMSE's in Table 7. The unweighted ALR-SME is not resistant to outliers, whereas the estimators with $c > 0$ are clearly resistant to the outliers. Interestingly, the most efficient estimate of $\boldsymbol{\beta}$ is arguably given by $c = 1$ or $c = 1.25$ even when there are no outliers. Again the message here is that the weighted version of the ALR-SME is valuable for improving efficiency for estimating the components of $\boldsymbol{\beta}$ that are negative and close to -1. The most efficient estimator for $\boldsymbol{A}_L$ is arguably $c = 0.5$ or $c = 0.75$.

Appendix A.5 (Supplementary Material) contains additional simulation results for dimension $p = 10$ and with a broader range of outlier contaminations (4%, 12% and 45%). This simulation also confirms that the robustified ALR-SME with $c > 0$ are resistant to the outliers, whereas the unweighted estimator with $c = 0$ does not exhibit good

Table 5 Simulation results *Dataset1* RMSE's

Parameter	$c = 0$	$c = 0.01$	$c = 0.7$	$a_c = 0.01$	$a_c = 0.000796$	a_c given $\hat{\boldsymbol{\beta}}$	$a_c = 1$
Simulation 3: RPPI model (continuous) with outliers							
a_{11}	3,530,000	93,200	109,000	100,000	85,600	86,500	315,000
a_{22}	3,630,000	6810	7680	7000	5710	5750	19,600
a_{33}	1090	21.4	18.8	29.2	21.7	22.4	114
a_{44}	2770	18.9	17.3	23.9	17.2	17.6	83.7
a_{12}	3,620,000	15,200	23,300	15,800	13,200	13,400	46,600
a_{13}	47,700	1200	1020	1460	1240	1250	4960
a_{14}	120,000	1000	920	1010	850	880	2920
a_{23}	96,400	277	278	297	255	263	771
a_{24}	135,000	289	241	321	263	267	927
a_{34}	1370	15.5	15.4	17.1	13	13.5	57.3
β_1	19.3	0.181	0.0701	–	–	0.0701	–
β_2	28.9	0.137	0.06	–	–	0.06	–
β_3	17.6	0.264	0.256	–	–	0.256	–
β_4	19.2	0.224	0.211	–	–	0.211	–
Simulation 4: multinomial latent variable model (discrete) with outliers							
a_{11}	3,050,000	57,200	90,800	105,000	105,000	105,000	281,000
a_{22}	3,140,000	5500	6680	10,100	10,100	10,100	22,400
a_{33}	1020	16.3	15.7	41.8	41.8	42.2	114
a_{44}	2360	17	16.1	41.8	41.8	42.1	95.7
a_{12}	3,130,000	12,300	15,600	18,900	18,900	18,900	49,400
a_{13}	42,900	825	868	1630	1630	1650	4590
a_{14}	93,500	880	857	1270	1270	1280	2840
a_{23}	85,200	270	251	410	410	413	878
a_{24}	116,000	247	225	494	494	495	1060
a_{34}	1210	14.5	14.4	25	25	25.2	58.2
β_1	16.5	0.157	0.0649	–	–	0.0649	–
β_2	27.2	0.132	0.0555	–	–	0.0555	–
β_3	16.8	0.251	0.246	–	–	0.246	–
β_4	16	0.22	0.206	–	–	0.206	–

resistance to outliers. When there are no outliers, the ALR-SME with $c > 0$ is also often more efficient than the $c = 0$ estimator.

8 Conclusion

We proposed a log-ratio score matching estimator that produces consistent estimates for $\boldsymbol{A}_L$ and the first $p-1$ elements of $\boldsymbol{\beta}$ for the RPPI model and the multinomial model with RPPI latent probability vectors. This estimator was insensitive to the huge number of zeroes often encountered in microbiome data, and even performed well when every datapoint had a component that was zero. Our new estimator and modelling approach does not require treating zeros as outliers, which is an improvement on the treatment of zeros in the standard Aitchison log-ratio approach based on the logistic normal distribution. The robustified version of our estimator remained insensitive to zeros, improved resistance to outliers and also improved efficiency over unweighted ALR-SME for well-specified data. We recommend using our estimators when there are many components, many of which have concentrations at/near zero (i.e. many β_j, $j = 1, 2, \ldots, p-1$ are close to -1).

Table 6 Simulation results *Dataset2* RMSE's

Parameter	$c = 0$	$c = 0.01$	$c = 0.25$	$c = 0.5$	$c = 0.75$	$c = 1$	$c = 1.25$
Simulation 5: RPPI model (continuous)							
a_{11}	1950	1940	1670	1490	1400	1330	1840
a_{22}	39,900	39,600	34,000	30,300	29,900	34,700	48,100
a_{33}	4790	4750	3970	3740	4160	5110	6080
a_{44}	122,000	120,000	97,800	92,700	101,000	128,000	252,000
a_{12}	128,000	126,000	107,000	94,600	88,500	104,000	220,000
a_{13}	15,300	15,200	13,000	12,000	12,500	14,900	16,000
a_{14}	29,200	29,000	24,200	21,500	22,900	26,500	29,600
a_{23}	84,900	84,200	70,600	63,200	60,100	61,800	68,400
a_{24}	86,200	85,500	74,300	69,400	70,800	77,200	120,000
a_{34}	23,300	23,100	20,400	20,300	22,700	28,100	34,900
β_1	0.0807	0.0796	0.0607	0.0508	0.047	0.0462	0.0467
β_2	0.0869	0.0857	0.0659	0.055	0.0496	0.0476	0.048
β_3	0.125	0.123	0.0931	0.0805	0.0771	0.0786	0.0816
β_4	0.197	0.193	0.145	0.126	0.123	0.123	0.129
Simulation 6: multinomial latent variable model (discrete)							
a_{11}	1840	1830	1580	1410	1300	1310	1330
a_{22}	38,300	38,000	32,900	30,800	43,500	49,700	52,300
a_{33}	3950	3910	3290	3360	3800	4500	5430
a_{44}	93,400	93,500	96,000	97,300	102,000	115,000	143,000
a_{12}	176,000	174,000	140,000	116,000	99,900	88,500	87,700
a_{13}	12,100	12,000	10,100	9090	8660	9600	11,000
a_{14}	24,700	24,500	20,200	18,500	17,100	18,100	19,500
a_{23}	72,600	71,800	58,700	50,600	46,100	45,800	46,800
a_{24}	80,700	80,000	67,200	59,800	63,700	64,700	69,200
a_{34}	18,300	18,100	15,800	16,100	17,400	20,400	24,900
β_1	0.0634	0.0625	0.0474	0.0395	0.036	0.0345	0.0345
β_2	0.0642	0.0633	0.0476	0.0395	0.0357	0.0349	0.0348
β_3	0.104	0.103	0.0789	0.0701	0.0681	0.0675	0.0687
β_4	0.145	0.145	0.153	0.152	0.148	0.144	0.142

Table 7 Simulation results *Dataset2* RMSE's

Parameter	$c = 0$	$c = 0.01$	$c = 0.25$	$c = 0.5$	$c = 0.75$	$c = 1$	$c = 1.25$
Simulation 7: RPPI model (continuous) with outliers							
a_{11}	4,500,000	2260	1950	1740	1630	1650	2490
a_{22}	8.61e+08	75,200	65,900	60,100	58,400	62,600	82,200
a_{33}	3,510,000	4850	4050	3890	4600	5250	6420
a_{44}	1.24e+08	124,000	101,000	96,300	109,000	132,000	176,000
a_{12}	6.49e+08	142,000	121,000	107,000	99,700	109,000	256,000
a_{13}	58,600,000	16,800	14,200	13,100	13,300	15,300	17,500
a_{14}	74,300,000	30,700	25,600	22,900	29,100	32,700	34,800
a_{23}	6.48e+08	89,400	74,700	66,500	62,100	63,900	70,100
a_{24}	6.53e+08	94,100	81,400	75,400	76,600	80,200	101,000
a_{34}	35,100,000	24,000	21,100	20,900	24,600	28,500	35,800
β_1	222	0.0826	0.0632	0.0529	0.0487	0.048	0.0484
β_2	222	0.0888	0.0684	0.0573	0.0516	0.0496	0.0498
β_3	254	0.128	0.097	0.0841	0.0805	0.0818	0.0833
β_4	131	0.202	0.151	0.131	0.126	0.126	0.13
Simulation 8: multinomial latent variable model (discrete) with outliers							
a_{11}	2,930,000	2340	2030	1820	1500	1600	1800
a_{22}	3.46e+08	62,600	52,800	46,900	56,100	59,600	61,200
a_{33}	2,520,000	4050	3370	3340	3680	4500	5300
a_{44}	69,800,000	95,000	97,200	98,900	104,000	116,000	143,000
a_{12}	2.72e+08	182,000	146,000	121,000	104,000	92,300	91,100
a_{13}	54,100,000	13,900	11,600	10,200	10,400	10,700	12,300
a_{14}	66,200,000	25,400	20,900	18,500	17,700	19,500	20,200
a_{23}	2.66e+08	79,700	65,200	56,200	51,600	51,200	51,700
a_{24}	2.76e+08	84,400	70,300	62,300	68,000	70,300	72,700
a_{34}	2.6e+07	18,600	16,100	16,000	17,200	20,700	25,100
β_1	175	0.0659	0.0499	0.0415	0.0378	0.0366	0.0363
β_2	131	0.0663	0.05	0.0413	0.0376	0.0367	0.0363
β_3	186	0.108	0.0831	0.0731	0.0701	0.0703	0.0701
β_4	107	0.148	0.155	0.153	0.148	0.145	0.142

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s11222-024-10412-w.

Acknowledgements This work was supported by Australian Research Council Grant DP220102232. We thank two referees for their constructive comments that helped to improve the final manuscript.

Author Contributions J.L.S., K.L.H., J.T.K. and A.T.A.W. wrote the main manuscript text. J.L.S. prepared Figs. 1 and 2. All authors reviewed the manuscript.

Funding Open Access funding enabled and organized by CAUL and its Member Institutions

Declarations

Conflict of interest The authors declare no competing interests.

References

Aitchison, J.: The Statistical Analysis of Compositional Data, Monographs on Statistics and Applied Probability, vol. 25. Chapman & Hall, London (1986)

Basak, S., Basu, A., Jones, M.C.: On the 'optimal' density power divergence tuning parameter. J. Appl. Stat. **48**, 536–556 (2021)

Basu, A., Harris, I.R., Hjort, N.L., Jones, M.C.: Robust and efficient estimation by minimising a density power divergence. Biometrika **85**, 549–559 (1998)

Cao, Y., Lin, W., Li, H.: Large covariance estimation for compositional data via composition-adjusted thresholding. J. Am. Stat. Assoc. **114**, 759–772 (2019)

Choi, E., Hall, P., Presnell, B.: Rendering parametric procedures more robust by empirically tilting the model. Biometrika **87**, 453–465 (2000)

CRAN: The comprehensive R archive network. https://cran.r-project.org (2022). Accessed 7 Dec 2022

He, Y., Liu, P., Zhang, X., Zhou, W.: Robust covariance estimation for high-dimensional compositional data with application to microbial communities analysis. Stat. Med. **40**(15), 3499–3515 (2021)

Hyvarinen, A.: Estimation of non-normalised statistical models by score matching. J. Mach. Learn. Res. **6**, 695–709 (2005)

Jones, M.C., Hjort, N.L., Harris, I.R., Basu, A.: A comparison of related density-based minimum divergence estimators. Biometrika **88**, 865–873 (2001)

Kato, S., Eguchi, S.: Robust estimation of location and concentration parameters for the von Mises-Fisher distribution. Stat. Pap. **57**, 205–234 (2016)

Li, H.: Microbiome, metagenomics, and high-dimensional compositional data analysis. Annu. Rev. Stat. Its Appl. **2**, 73–94 (2015)

Liang, W., Wu, Y., Xiaoyan, M.: Robust sparse precision matrix estimation for high-dimensional compositional data. Stat. Probab. Lett. **184**, 109379 (2022)

Martin, I., Uh, H.-W., Supali, T., Mitreva, M., Houwing-Duistermaat, J.J.: The mixed model for the analysis of a repeated-measurement multivariate count data. Stat. Med. **38**, 2248–2268 (2018)

Maronna, R., Martin, D., Yohai, V.: Robust Statistics: Theory and Methods. Wiley, Chichester (2006)

Mishra, A., Muller, C.L.: Robust regression with compositional covariates. Comput. Stat. Data Anal. **165**, 107315 (2022)

Muller, S., Welsh, A.H.: Outlier robust model selection in linear regression. J. Am. Stat. Assoc. **100**, 1297–1310 (2005)

Ribeiro, T.K.A., Ferrari, S.L.P.: Robust estimation in beta regression via maximum L_q-likelihood (2020). arXiv:2010.11368

Salibian-Barrera, M., Van Aelst, S., Willems, G.: Fast and robust bootstrap. Stat. Methods Appl. **17**, 41–71 (2008)

Saraceno, G., Ghosh, A., Basu, A., Agostinelli, C.: Robust estimation under linear mixed models: the minimum density power divergence approach (2020). arXiv:https://arxiv.org/pdf/2010.05593pdf

Scealy, J.L., Welsh, A.H.: Colours and cocktails: compositional data analysis. 2013 Lancaster lecture. Aust. N. Z. J. Stat. **56**, 145–169 (2014)

Scealy, J.L., Wood, A.T.A.: Analogues on the sphere of the affine-equivariant spatial median. J. Am. Stat. Assoc. **116**, 1457–1471 (2021)

Scealy, J.L., Wood, A.T.A.: Score matching for compositional distributions. J. Am. Stat. Assoc. **118**, 1811–1823 (2023)

Weistuch, C., Zhu, J., Deasy, J.O., Tannenbaum, A.R.: The maximum entropy principle for compositional data. BMC Bioinform. **23**, 1–13 (2022)

Windham, M.P.: Robustifying model fitting. J. R. Stat. Soc. B **57**, 599–609 (1995)

Yu, S., Drton, M., Shojaie, A.: Interaction models and generalized score matching for compositional data (2021). arXiv:2109.04671

Zhang, J., Lin, W.: Scalable estimation and regularization for the logistic normal multinomial model. Biometrics **75**, 1098–1108 (2019)