Score Matching for Compositional Distributions

Janice L. Scealy and Andrew T. A. Wood

Research School of Finance, Actuarial Studies and Statistics, Australian National University, Canberra, ACT, Australia

ABSTRACT

Compositional data are challenging to analyse due to the non-negativity and sum-to-one constraints on the sample space. With real data, it is often the case that many of the compositional components are highly right-skewed, with large numbers of zeros. Major limitations of currently available models for compositional data include one or more of the following: insufficient flexibility in terms of distributional shape; difficulty in accommodating zeros in the data in estimation; and lack of computational viability in moderate to high dimensions. In this article, we propose a new model, the polynomially tilted pairwise interaction (PPI) model, for analysing compositional data. Maximum likelihood estimation is difficult for the PPI model. Instead, we propose novel score matching estimators, which entails extending the score matching approach to Riemannian manifolds with boundary. These new estimators are available in closed form and simulation studies show that they perform well in practice. As our main application, we analyse real microbiome count data with fixed totals using a multinomial latent variable model with a PPI model for the latent variable distribution. We prove that, under certain conditions, the new score matching estimators are consistent for the parameters in the new multinomial latent variable model.

1. Introduction

Modern scientific measuring systems commonly record compositional data, which are $p$-dimensional vectors $u$ defined on the unit simplex

$$\Delta^{p-1} = \left\{(u_1, u_2, \ldots, u_p)^\top \in \mathbb{R}^p : u_i \geq 0, \quad \sum_{i=1}^p u_i = 1 \right\}.$$  

Compositional data arise naturally in diverse areas including geochemistry, economics, and biology. Having available suitable methods for analysing such data are of importance in many fields.

Various different distributions have been proposed for modelling compositional data which take the constraints within $\Delta^{p-1}$ into account. In summary, these are the Dirichlet distribution and its generalisations (e.g., Krzysztofowicz and Reese 1993; Ongaro et al. 2020), the logistic normal distribution (Aitchison 1986) and the folded Kent distribution based on the square root transformation (Scealy and Welsh 2011; 2014). Butler and Glasbey (2008) projected a multivariate Gaussian distribution onto the simplex and Leininger et al. (2013) applied a different projection (after truncation at zero) of a multivariate Gaussian distribution onto the simplex. More recently Tsagris and Stewart (2020) projected and folded a power-transformed multivariate Gaussian distribution onto the simplex. Maximum likelihood (ML) estimation for the projected, truncated and folded distribution parameters is not straightforward due to the complexities of the densities and, unless $p$ is fairly small, these methods are generally not tractable. ML estimation for the Dirichlet, logistic normal distribution and the Tsagris and Stewart (2020) model cannot be applied whenever there are zeros in the data since the log-likelihood is not finite. Stewart and Field (2010) and Bear and Billheimer (2016) both modified the logistic normal model to handle zeros by separating the components into two parts and modelling the zeros separately. However, these two methods based on maximum likelihood estimation were applied only in the case where the dimension was small and Bear and Billheimer (2016) assumed that one of the components was always non zero. Another possibility is to use moment estimators for the Dirichlet distribution. These estimators are still well-behaved even when some of the observation vectors lie on the boundary of the simplex, but the Dirichlet correlation structure is too restrictive for this approach to be generally useful.

In summary, there is currently an absence of parametric models which satisfy all three of the following criteria: (i) a flexible model, in the sense of permitting a wide variety of distributional shapes; (ii) zeros in the data can be accommodated in estimation; and (iii) tractable estimation for substantial $p$. In this article, we define a new parametric model, the polynomially tilted pairwise interaction (PPI) model valid for continuous variables, which satisfies all of criteria (i)–(iii). However, except in special cases, ML estimation is difficult for the PPI model due to the complexity of the normalizing constant. Novel estimators are proposed which are based on the score matching idea (see next paragraph) but extended to Riemannian manifolds with boundary, of which the simplex (1) is a particular case. These estimators are consistent, quite efficient, available in...
closed form and fast and simple to compute. This constitutes a major advance.

Hyvarinen (2005) developed a set of score matching estimators for densities on Euclidean space. A key advantage of this type of estimator is that it avoids the calculation of the normalizing constant in the model. In its original form, score matching assumed that the probability density function is differentiable over the entire space $\mathbb{R}^p$. Hyvarinen (2007) extended score matching to densities defined on the nonnegative orthant by introducing weights into the objective function that are zero on the boundary of the sample space. Yu et al. (2019) extended this approach to allow non smooth weights in order to improve estimation efficiency. Liu et al. (2020) investigated score matching on subsets of Euclidean spaces with more complicated truncation boundaries and applied a weight function defined by the distance from a point in the domain to the boundary of the domain. Yu et al. (2020) explored score matching on general distance from a point in the domain to the boundary of the cation boundaries and applied a weight function defined by the fact that the marginal log-likelihood is not tractable.

2. Polynomially Tilted Pairwise Interaction Model

2.1. Definition of the Model

Assume that we have an iid (independent and identically distributed) sample of size $n$, $u_1 = (u_{11}, u_{12}, \ldots, u_{1p})^\top \in \Delta^{p-1}$, $i = 1, 2, \ldots, n$. Each $u_i \in \Delta^{p-1}$ is assumed to have the following density with respect to Lebesgue measure, $d\Delta^{p-1}$, on $\Delta^{p-1}$:

$$
\frac{1}{c_1(A^*, \beta)} \left( \prod_{j=1}^p u_j^{\beta_j} \right) \exp \left( u^\top A^* u \right),
$$

where $c_1(A^*, \beta)$ is the normalizing constant and $A^*$ is a general $p \times p$ symmetric matrix. It is assumed that one of the eigenvalues of $A^*$ is zero due to the constraint that the compositional components sum to one and the eigenvector associated with the zero eigenvalue is proportional to $(1, 1, \ldots, 1)^\top$. Note that the matrix $A^*$ does not necessarily need to be negative semidefinite due to the fact that the simplex is a bounded space. Model (2) is a canonical exponential family which makes the score matching estimation methods developed in Section 3 so tractable. These estimators can also be used in a powered PPI model with $u^\top A^* u$ in (2) replaced by $(u_1^a, u_2^a, \ldots, u_p^a)A^*(u_1^a, u_2^a, \ldots, u_p^a)^\top$, where $\alpha > 0$ is known.

The vector $\beta = (\beta_1, \beta_2, \ldots, \beta_p)^\top$ consists of shape parameters which satisfy $\beta_j > -1$ for $j = 1, 2, \ldots, p$. This model is permutation invariant and very flexible. An equivalent density with respect to $d\Delta^{p-1}$, which is sometimes more convenient, is

$$
\frac{1}{c_2(A, b, \beta)} \left( \prod_{j=1}^p u_j^{\beta_j} \right) \exp \left( u^\top Au + b^\top u \right),
$$

where $A$ is a $p \times p$ symmetric matrix with $i, j$th element denoted by $a_{ij}$ and $b$ is a $p \times 1$ vector with $j$th element denoted by $b_j$. To account for the constraint and for identifiability we set the last row and column of $A$ to zero, that is $a_{ij} = 0$ and $a_{pj} = 0$ for $i, j = 1, \ldots, p$, and also $b_p = 0$. Define $A_L$ to be the $(p-1) \times (p-1)$ matrix with the last row and column in $A$ removed and define $b_L$ to be the $(p-1) \times 1$ vector with the last element in $b$ removed.
In the following subsections, we discuss submodels of model (3) and compare them with other well-known models. Sampling algorithms for generating random observations from these distributions are given in Appendix A.1 (supplementary material).

2.2. Comparison With Other Models

The logistic normal distribution mentioned in the Introduction is flexible in the sense that it can model general location and covariance structures. In particular, the logistic normal probability density function has approximately elliptical contours of constant probability when the model is highly concentrated in the interior of the simplex. The polynomially tilted pairwise interaction (PPI) model (3) also has the following property: provided the components of $\beta$ are not too large, the ellipticity is largely controlled through the parameter $A_L$; for example, see Figure 1(d) and Section 2.3.

The logistic normal distribution has a major deficiency: it is not flexible at the boundaries of the simplex because it assumes that the probability density function is always zero at the boundaries. In contrast, the new PPI model (3) is rather more flexible at the boundaries and can handle three different types of behavior. Suppose that $u_j \rightarrow 0$ in such a way that each $u_k, k \neq j$, remains bounded away from 0. Then (i) if $\beta_j < 0$ then the probability density function becomes infinite as $u_j \rightarrow 0$; (ii) if $\beta_j = 0$ then the probability density function is bounded away from 0 and $\infty$ when $u_j \rightarrow 0$; and (iii) if $\beta_j > 0$ then the probability density function approaches zero when $u_j \rightarrow 0$. Note that the trichotomy (i), (ii), and (iii) may be thought of heuristically as “Dirichlet behavior” at the boundary of the simplex.

In summary, the Dirichlet model, obtained by setting $A^* = 0$ in (2), or $A = 0$ and $b = 0$ in (3), cannot account for general covariance structures while the logistic normal does not have the flexibility of the Dirichlet at the boundary of $\Delta^{p-1}$. The advantage of the PPI model over the other two models is that it simultaneously allows Dirichlet behavior at the boundary of $\Delta^{p-1}$ and general covariance structures.

2.3. Example for $p = 3$

We now demonstrate the flexibility of model (3) and the approximate interpretation of the parameters in the case where the data is relatively highly concentrated and some of the components are distributed close to the boundary. The dimension $p = 3$ is chosen for demonstrative purposes and a similar exploratory approach can be used in higher dimensions. Further discussion of the $\beta_j$’s is given in Section 4. Let

$$ A_L = \begin{pmatrix} -166.6667 & 117.8511 \\ 117.8511 & -333.3333 \end{pmatrix} \quad \text{and} \quad b_L = \begin{pmatrix} 11.71573, 51.71573 \end{pmatrix}^T, \tag{4} $$

and then define

$$ \Sigma = -\frac{1}{2} A_L^{-1} = \begin{pmatrix} 0.004 & 0.001414214 \\ 0.001414214 & 0.002 \end{pmatrix} \quad \text{and} \quad \mu = -\frac{1}{2} A_L^{-1} b_L = (0.12, 0.12)^T. $$

We next consider four choices of $\beta$: $\beta = (-0.8, -0.8, -0.5)^T$, $\beta = (-0.5, -0.5, -0.5)^T$, $\beta = (0, 0, -0.5)^T$ and $\beta = 0$ (0, 0, 0)^T$.

Figure 1. Bivariate distributions. (a)–(d): Plots of $u_2$ versus $u_1$ for the case $\beta = (-0.8, -0.8, -0.5)^T$, $\beta = (-0.5, -0.5, -0.5)^T$, $\beta = (0, 0, -0.5)^T$ and $\beta = (0, 0, 0)^T$, respectively. We do not use a Ternary plot here because the other boundary lies outside the plotting region.
(0, 0, 0)\top. For each of these choices of \( \beta \), we simulated a single sample of size \( n = 10,000 \) from model (3) with the other parameters \( A_L \) and \( b_L \) fixed, as given in (4). Figures 1 and 2 contain the simulated marginal and bivariate distributions. These plots show that \( u_1 \) and \( u_2 \) are distributed close to the zero boundary and the third component \( u_3 \) is distributed close to one. For the components \( u_1 \) and \( u_2 \), the different choices of \( \beta \) lead to rather different marginal distribution shapes near the zero boundary; see Figure 2. As seen in Figure 1 (d), when \( \beta = (0, 0, 0)\top \) the contours of constant probability are roughly elliptical and the model for \( (u_1, u_2)\top \) is approximately bivariate Gaussian with mean \( \mu \) and covariance \( \Sigma \). When \( \beta_1 = -0.8 \) and \( \beta_2 = -0.8 \), the marginal distributions of \( u_1 \) and \( u_2 \) are highly right skewed. Figure 1 shows that \( A_L \) and \( b_L \) roughly control the covariance and location of the distribution of \( (u_1, u_2)\top \). \( \beta_1 \) controls the shape of the density at and near \( u_1 = 0 \) and similarly \( \beta_2 \) controls the shape of the density at and near \( u_2 = 0 \). Note that \( \beta_3 \) has negligible influence on the shape of the distributions in this example because \( u_3 \) is not distributed close to zero (e.g., cases \( \beta = (0, 0, -0.5)\top \) and \( \beta = (0, 0, 0)\top \) have similar density contours and shapes).

2.4. Approximate Model on \( \mathbb{R}_+^{p-1} \) and Interpretation of \( A_L \)

In many applications of compositional data, there is often one dominant component which is distributed close to one and many small components are distributed close to zero. Without loss of generality, assume that the data is ordered such that the most abundant component is \( u_p \), while \( u_1, \ldots, u_{p-1} \) are the less abundant components. If we assume \( \beta_p = 0 \) in model (3), then \( u_L = (u_1, u_2, \ldots, u_{p-1})\top \) is approximately distributed on \( \mathbb{R}_+^{p-1} \) since \( u_1 + u_2 + \cdots + u_{p-1} < 1 \) with high probability, so that the sum-to-one constraint has little impact on \( u_1, \ldots, u_{p-1} \). Then, for \( j < k \leq p - 1 \), \( u_j \) and \( u_k \) are approximately conditionally independent given all other variables in \( u_L \) if and only if \( a_{jk} = a_{kj} = 0 \) in \( A_L \) (e.g., Yu et al. 2019).

2.5. Truncated Gaussian Distribution

If we preset \( \beta = 0 \) and assume that \( A_L \) is negative definite, then we have a \((p - 1)\)-dimensional truncated Gaussian distribution and, in this case, if \( A_L \) is large in magnitude and \( \mu = -\frac{1}{2}A_L^{-1}b_L \) is not too close to the boundary of the simplex, then the distribution is approximately multivariate Gaussian with a general covariance matrix equal to \(-\frac{1}{2}A_L^{-1}\) and mean vector \( \mu \). This model provides a practical alternative to the folded Kent distribution and may be useful for modelling certain types of ecological and expenditure compositional datasets where some of the components are distributed close to the simplex boundaries with only small to moderate right skewness (e.g., Sceal and Welsh, 2011, 2014; 2017).

3. Score Matching

The square root transformation, which for some purposes is more convenient than working with \( u \in \Delta^{p-1} \), is given by
\(z = (\sqrt{u_1}, \sqrt{u_2}, \ldots, \sqrt{u_p})^\top\) and is a bijection from the simplex, \(\Delta^{p-1}\), to \(S^{p-1}_+\), the positive orthant of the unit sphere, that is,
\[
S^{p-1}_+ = \{z = (z_1, z_2, \ldots, z_p)^\top \in \mathbb{R}^p : \|z\| = 1, z_j \geq 0, \ j = 1, 2, \ldots, p\},
\]
where \(\| \cdot \|\) is the Euclidean norm. Note that \(\Delta^{p-1}\) and \(S^{p-1}_+\) are \((p - 1)\)-dimensional diffeomorphic Riemannian manifolds with boundary, and \(\Delta^{p-1}\) is flat, while \(S^{p-1}_+\) is curved. Define \(z^2 = (z_1^2, z_2^2, \ldots, z_p^2)^\top\). The density of \(z\) is given by
\[
\frac{2^p}{c_2(A, b, \beta)} \left( \prod_{j=1}^p z_j^{1+2\beta_j} \right) \exp \left( z^2 A z^2 + b^\top z^2 \right),
\]
\[
z = (z_1, z_2, \ldots, z_p)^\top \in S^{p-1}_+ ,
\]
with respect to Lebesgue measure (i.e., unnormalized uniform measure) on \(S^{p-1}_+\).

### 3.1. Manifolds With Boundary

Mardia et al. (2016) defined consistent score matching estimators for the Bingham, Kent, and other directional distributions. The method is similar to the score matching estimators proposed by Hyvarinen (2005, 2007), but adapted to handle estimation on a Riemannian manifold \(M\). Note that model (6) is also a directional distribution. However, we cannot directly apply the Mardia et al. (2016) estimators to estimate the parameters in model (6) because the manifold \(S^{p-1}_+\) has a boundary and Theorem 1 in Mardia et al. (2016) does not cover this case. We now modify the estimation approach proposed in Mardia et al. (2016) to handle the boundary in order to estimate the parameters in model (6).

It will be convenient for the remainder of the article to assume that the manifold of interest, \(M\), is isometrically embedded in a Euclidean space \(\mathbb{R}^P\) and that we use Euclidean coordinates, \(z\), to describe \(M\). For all functions \(u : M \rightarrow \mathbb{R}\) that we consider below, it is assumed that \(u\) can be smoothly extended to a function \(\tilde{u}\) in a neighborhood of \(M\) in \(\mathbb{R}^P\).

Let \(f(z)\) and \(\tilde{f}_0(z)\), \(z \in M \subset \mathbb{R}^P\), be two probability densities on a Riemannian manifold \(M\), defined with respect to the unnormalized uniform measure on \(M\), denoted \(dM(z)\). It is assumed that \(f\) and \(\tilde{f}_0\) are everywhere nonzero and twice continuously differentiable, that is, all components of \(\nabla z \nabla z^\top f\) are continuous functions of \(z \in M\). Write \(P = P(z) = (p_{ij})\) for the orthogonal projection of a point in \(\mathbb{R}^P\) onto the \(\xi\)-dimensional subspace of \(\mathbb{R}^P\) parallel to the tangent space at \(z \in M\), so that \(P\) has rank \(\xi \leq p\). As \(P\) is an orthogonal projection, \(P^2 = P = P^\top\). The weighted Hyvarinen divergence between \(f\) and \(\tilde{f}_0\) is defined by
\[
\Phi(f, \tilde{f}_0) = \frac{1}{2} \int_M \tilde{f}_0(z) \tilde{h}(z)^2 \left\| P \left( \nabla z \log(f) - \nabla z \log(\tilde{f}_0) \right) \right\|^2 dM(z) \times P \left( \nabla z \log(f) \right)^\top dM(z) \times P \left( \nabla z \log(\tilde{f}_0) \right) dM(z).
\]
\[
\tilde{h}(z)^2 = \min \left( \prod_{j=1}^p z_j^2, a_j^2 \right) ,
\]
where \(\| \cdot \|\) is the Euclidean norm and \(\tilde{V}_z\) is the usual Euclidean gradient, for example, \(V_z \tilde{u} = (\partial \tilde{u}/\partial z_1, \ldots, \partial \tilde{u}/\partial z_p)^\top\) with \(z = (z_1, \ldots, z_p)^\top\). It is assumed that the weight function \(\tilde{h}^2\) in (7) is smooth and zero on the boundary, \(\partial M\), or is one of the weight functions (9), (11), or (12).

### 3.2. Choice of \(\tilde{h}\) Function

There are various choices of \(\tilde{h}^2\) that we could make. Similarly to Hyvarinen (2007), a simple choice which guarantees a zero contribution from the relevant integral over the boundary of \(S^{p-1}_+\) is
\[
\tilde{h}(z)^2 = \prod_{j=1}^p z_j^2 .
\]
However, the problem with this function is that it places too much weight on the interior of the simplex and will lead to a loss of efficiency, which is expected to be more dramatic when the dimension increases and the majority of the data is distributed close to the boundaries. Similarly to Yu et al. (2019), we could use a capped weight function
\[
\tilde{h}(z)^2 = \min \left( \prod_{j=1}^p z_j^2, a_j^2 \right) ,
\]
which reduces the weight in the middle of the simplex. The parameter \(a_j\) needs to be chosen such that 0 ≤ \(a_j\) ≤ 1. Note that the optimal choice of \(a_j\) will usually be somewhere in between 0 and 1, depending on the data.

Following Liu et al. (2020), another choice of weight function would be the distance from the datapoint to the boundary. On the simplex scale this (Euclidean) distance is defined in the following lemma, which is proved in Appendix A.2 (supplementary material).

\textbf{Lemma 1.} Suppose \(u = (u_1, \ldots, u_p)^\top \in \text{Int} \{\Delta^{p-1}\}\), the relative interior of \(\Delta^{p-1}\), and let \(\partial \Delta^{p-1} = \{u \in \Delta^{p-1} : u_j = 0\ \text{for some} \ j \subset \mathbb{R}^p\}\) denote the boundary of \(\Delta^{p-1}\), defined by \(\partial \Delta^{p-1} = \Delta^{p-1} \setminus \text{Int} \{\Delta^{p-1}\}\). Define \(C_p = \sqrt{p/(p - 1)}\). Then
\[
\tilde{h}(u)^2 = \min_{v \in \partial \Delta^{p-1}} \|u - v\| = C_p \min(u_1, \ldots, u_p).\]

The weight function defined in (10) transformed to the sphere scale is then (ignoring constants)
\[
\tilde{h}(z)^2 = \min(z_1^2, \ldots, z_p^2) .
\]
We could also put a cap in to reduce the weight in the middle of the simplex
\[
\tilde{h}(z)^2 = \min(z_1^2, \ldots, z_p^2, a_j^2) .
\]
In (7), \( \tilde{f} \) represents the density of the model that we would like to fit to data and \( \tilde{f}_0 \) is the unknown population density. To estimate the parameters in model \( \tilde{f} \), we minimize the objective function \( \Phi(\tilde{f}; \tilde{f}_0) \) with respect to the parameters in \( \tilde{f} \). The apparent difficulty caused by dependence of (7) on the unknown \( \tilde{f}_0 \) is resolved by (i) applying Theorem 1 below and (ii) replacing population moments with respect to \( \tilde{f}_0 \) in (13) by sample moments. See Appendix A.3 (supplementary material) for our proof of Theorem 1. For the expressions used in (13) and (14) for the Riemannian inner product and the Laplacian, \( \tilde{\Delta} \), expressed in Euclidean coordinates; see, for example, Mardia et al. (2016).

**Theorem 1.** Suppose \( M \) is a manifold with boundary \( \partial M \) and assume that \( \tilde{f} \) and \( \tilde{f}_0 \) are smooth functions of \( z \) which are positive in the interior of \( M \). Additionally, assume that one of the following two conditions hold:

(C1) \( \tilde{h}^2 : M \to \mathbb{R} \) is a smooth function which is zero on \( \partial M \), or

(C2) \( M = S_p^{p-1} \) or \( M = \Delta^{p-1} \) and \( \tilde{h}^2 \) is of the form (9), (11) or (12).

Then \( \text{argmin}_z \Phi(\tilde{f}; \tilde{f}_0) = \text{argmin}_z \Psi(\tilde{f}; \tilde{f}_0) \), where in both cases the argmin is taken over the unknown parameters in \( \tilde{f} \). \( \Phi(\tilde{f}; \tilde{f}_0) \) is defined in (7) and

\[
\Psi(\tilde{f}; \tilde{f}_0) = \frac{1}{2} \int_{S_p^{p-1}} \tilde{f}_0(z) \tilde{h}(z)^2 \begin{bmatrix} \nabla_z \log(\tilde{f}(z)) \end{bmatrix}^\top \left( P \begin{bmatrix} \nabla_z \log(\tilde{f}(z)) \end{bmatrix} \right) dM(z) \\
+ \int_{S_p^{p-1}} \tilde{f}_0(z) \left( \tilde{h}(z)^2 \tilde{\Delta} \log(\tilde{f}) + \begin{bmatrix} \nabla_z \tilde{h}(z) \end{bmatrix}^2 \right)^\top \\
\times \left( P \begin{bmatrix} \nabla_z \log(\tilde{f}(z)) \end{bmatrix} \right) dM(z),
\]

where for a real-valued function \( \tilde{u}(z) \) expressed as a function of Cartesian coordinates, the Laplacian, \( \tilde{\Delta} \), on \( M \) is defined by

\[
\tilde{\Delta} \tilde{u}(z) = \sum_{i=1}^p \sum_{j=1}^p \sum_{k=1}^p p_{ij}(z) \frac{\partial}{\partial z_i} \left( p_{jk}(z) \frac{\partial \tilde{u}(z)}{\partial z_k} \right).
\]

Here, we consider \( M = S_{p+1}^{p-1} \) in (5) as being embedded in \( \mathbb{R}^p \) in the standard way and work with Cartesian coordinates \( z = (z_1, \ldots, z_p) \) \( \top \) in \( \mathbb{R}^p \). In (13) with \( M = S_{p+1}^{p-1} \), the appropriate projection matrix \( P \) is given by \( P = I_p - z z^\top \) for \( z \in S_{p+1}^{p-1} \), where \( I_p \) is the \( p \times p \) identity matrix.

Another possibility is to perform the score matching directly on the simplex, \( \Delta^{p-1} \), rather than on \( S_{p+1}^{p-1} \). This entails choosing \( P = I_p - p^{-1} I_p z z^\top \), where \( I_p \) is the \( p \)-vector of ones. Although this leads to a viable alternative approach, in our numerical examples, the score matching estimators tended to be more efficient on \( S_{p+1}^{p-1} \) than on \( \Delta^{p-1} \) and, for this reason, we decided to focus on the former in this article.

### 3.4. Estimators for the PPI Model

Here, we give a more explicit expression for the score matching estimator. Let \( q^* \) denote the total number of parameters in the model to be estimated; see Appendix A.5 (supplementary material) for the full model, A.6 (supplementary material) for the restricted model with \( \beta \) fixed and A.7 (supplementary material) for the Dirichlet model. The model can be written in the following form:

\[
\log(\tilde{f}(z)) = \pi^\top \tilde{t}(z) + \tilde{t}_0(z) + c_0,
\]

where \( c_0 \) is a constant; \( \tilde{t}(z) \) is a \( q^* \times 1 \) vector of sufficient statistics; \( \pi \) is a \( q^* \times 1 \) vector of parameters; \( \tilde{t}_0(z) \) is an offset function which does not depend on \( \pi \) and consists of the sufficient statistics of any \( \beta \)’s that we decide not to estimate (see Section 4). Omitting terms which do not depend on \( \pi \), the objective function \( \Psi(\tilde{f}; \tilde{f}_0) \) reduces to

\[
\Psi(\tilde{f}; \tilde{f}_0) = \frac{1}{2} \int_{S_p^{p-1}} \tilde{f}_0(z) \tilde{h}(z)^2 \begin{bmatrix} \nabla_z (\pi^\top \tilde{t}(z)) \end{bmatrix}^\top P(z) \\
\times \begin{bmatrix} \nabla_z (\pi^\top \tilde{t}(z)) \end{bmatrix} dS_p^{p-1} \\
- \int_{S_p^{p-1}} \tilde{f}_0(z) \left( \tilde{h}(z)^2 \pi^\top \tilde{t}(z) \right) dS_p^{p-1} \\
- \left( \int_{S_p^{p-1}} \tilde{f}_0(z) \left( \nabla_z \tilde{h}(z)^2 \right)^\top P(z) \nabla_z (\pi^\top \tilde{t}(z)) \right) dS_p^{p-1} \\
\times P(z) \begin{bmatrix} \nabla_z \tilde{t}_0(z) \end{bmatrix} dS_p^{p-1},
\]

\[
\frac{1}{2} \pi^\top W \pi - \pi^\top \left( d^{(1)} + d^{(2)} + d^{(6)} \right),
\]

where \( W \) is a \( q^* \times q^* \) matrix representing the terms in the first integral, \( d^{(1)}, d^{(2)} \) and \( d^{(6)} \) are \( q^* \times 1 \) vectors representing the terms in the last three integrals, respectively, and \( P(z) = I_p - z z^\top \). Replacing population moments in (15) by sample moments and minimizing the sample analogue of (15) over \( \pi \) gives

\[
\hat{\pi} = \hat{W}^{-1} (d^{(1)} + d^{(2)} + d^{(6)}) = \hat{W}^{-1} \hat{d},
\]

which is the score matching estimator, where

\[
\hat{W} = \frac{1}{n} \sum_{i=1}^n R(z_i), \quad d^{(j)} = -\frac{1}{n} \sum_{i=1}^n r_j(z_i), \quad j = 1, 2, 6;
\]

\[
R(z) = \tilde{h}(z)^2 \begin{bmatrix} \nabla_z \tilde{t}(z) \end{bmatrix}^\top P(z) \nabla_z \tilde{t}(z)^\top, \\
\hat{r}_1(z) = \tilde{h}(z)^2 \tilde{\Delta} \tilde{t}(z),
\]

\[
R_2(z) = \begin{bmatrix} \nabla_z \tilde{t}(z) \end{bmatrix}^\top P(z) \nabla_z \tilde{h}(z)^2, \\
r_6(z) = \tilde{h}(z)^2 \begin{bmatrix} \nabla_z \tilde{t}(z) \end{bmatrix}^\top P(z) \nabla_z \tilde{t}_0(z);
\]

\( \hat{\Delta} \) is the Laplacian operator, see (14), specialized to \( S_{p+1}^{p-1} \), which is applied componentwise in \( \tilde{\Delta} \tilde{t}(z) \) in (18) and is evaluated at \( z = z_i \) and, for example, \( \nabla_z \tilde{t}(z)^\top \) is the \( p \times q^* \) matrix with element \( (j,k) \) given by \( \partial \tilde{t}_j(z)/\partial z_k \) evaluated at \( z = z_i \). The
components in (18) and (19) are given explicitly in Appendices A.5–A.7 (supplementary material).

This estimator is applicable even when \( n \) and \( p \) are large (assuming \( p < n \)) as long as we are able to invert \( \hat{\mathbf{W}} \), with regularization if necessary. We could also constrain \( A_L \) in the model to have a special structure, for example, diagonal with \( b_L = 0 \), if we wish to simplify the model. It is also possible to deal with the case \( p > n \) by inserting a penalty term into the objective function (e.g., a ridge penalty).

### 3.5. Asymptotic Normality and Standard Errors

Theorem 2 establishes asymptotic normality for the estimator \( \hat{\pi} = \hat{\mathbf{W}}^{-1} \hat{a} \) and gives an expression for the standard errors. The proof is very similar to Yu et al. (2019, theor. 6) and is omitted.

Recall the definitions in (16)–(19) and define \( \mathbf{W}_0 = E_0 \left( \mathbf{W} \right) \), \( \mathbf{d}_0 = E_0 \left( \hat{a} \right) \), \( \pi_0 = \mathbf{W}_0^{-1} \mathbf{d}_0 \) and

\[
\Sigma_0 = \lim_{n \to \infty} nE_0 \left\{ \left( \mathbf{W} \pi_0 - \hat{a} \right) \left( \mathbf{W} \pi_0 - \hat{a} \right)^\top \right\},
\]

where \( E_0 \) denotes expectation with respect to the population distribution. If condition (C1) or (C2) in Theorem 1 holds, then \( \mathbf{W}_0 \) and \( \mathbf{d}_0 \) both have all elements finite.

**Theorem 2.** Suppose that

(C1) \( \mathbf{W}_0, \mathbf{W}_0^{-1}, \mathbf{d}_0 \) and \( \pi_0 \) exist and are entry-wise finite, and

(C2) there exists an \( n_0 \) such that, for \( n \geq n_0 \), \( \hat{\mathbf{W}} \) is a.s. invertible.

Then, the minimum of \( \frac{1}{2} \pi^\top \hat{\mathbf{W}} \pi - \pi^\top \hat{a} \) is a.s. unique with closed form solution \( \hat{\pi} = \hat{\mathbf{W}}^{-1} \hat{a} \). Moreover, \( \hat{\pi} \to_{a.s.} \pi_0 \) and \( \sqrt{n} (\hat{\pi} - \pi_0) \to_{d} N(0, \mathbf{W}_0^{-1} \Sigma_0 \mathbf{W}_0^{-1}) \) as \( n \to \infty \). Also, \( \mathbf{var}(n^{1/2} \hat{\pi}) \) may be estimated consistently using \( \hat{\mathbf{W}}^{-1} \hat{\Sigma}_0 \hat{\mathbf{W}}^{-1} \), where

\[
\hat{\Sigma}_0 = \frac{1}{n} \sum_{i=1}^{n} \left( R(z_i) \hat{\pi} - r(z_i) \right) \left( R(z_i) \hat{\pi} - r(z_i) \right)^\top
\]

and \( r(z) = -r_1(z) - r_2(z) - r_6(z) \).

### 4. Zeros and Practical Identifiability

#### 4.1. Zeros in the Data

The PPI model is a model for continuous variables and the probability of observing zero values in the sample is zero. However, in practice zeros can arise, for example, due to being structural or due to measurement error. The new score matching estimators as defined in Section 3.4 can still be applied to estimate the parameters in model (3) even when we have some \( u_j = 0 \) and \( \beta_j < 0 \). These boundary datapoints get either constant (non-zero) weight or zero weight in the sample moment calculations; see Appendix A.5 (supplementary material). This is very similar to what happens when the moment estimator for the Dirichlet distribution is applied: specifically, the resulting estimators are insensitive to the presence of exact or approximate zeros.

#### 4.2. Decisions Concerning \( \beta \)

A key question when using the PPI model is deciding how to treat the components of \( \beta \): specifically, to decide which components of \( \beta \) should be estimated, as part of the score matching procedure, and which \( \beta \) components should be treated as tuning parameters and fixed in advance. The motivation for the proposals below comes principally from extensive computational experience with the PPI model.

For those components of the observation vectors \( \mathbf{u}_j \in \Delta^{p-1} \) that have many zeros and/or points very close to zero, along with right skewness, for the PPI model to perform well it is vital that the corresponding components of \( \beta \) are estimated. Conversely, for those components of the \( \mathbf{u}_j \) which are always, or nearly always, well away from zero, from a practical point of view there is a lack of identifiability between the corresponding components of \( \beta \) and the parameters in \( A_L \) and \( b_L \). Intuitively, unless component \( j \) of the \( \mathbf{u}_j \) has many points at or near zero, there is not enough information in the data to estimate component \( \beta_j \) in addition to \( A_L \) and \( b_L \). Moreover, we have found that, in the case of more dominant components, the results are insensitive to the choice of \( \beta_j \) provided they are chosen to be of at most modest size. Thus, regarding \( \beta \), two points need to be decided before model fitting: (i) Which components of \( \beta \) should be estimated and which should be fixed in advance? (ii) What value(s) should we choose for those components of \( \beta \) that we decide to fix rather than estimate?

We first focus on question (ii). For simplicity and in the absence of motivation for doing otherwise, we suggest fixing those \( \beta_j \) we decide to treat as tuning parameters to have a common value, say \( \tau_0 \), for example \( \tau_0 = 0 \) or \( \tau_0 = 1 \).

We now consider question (i). In many data applications, there will be one, or a small number, of dominant component(s) whose values are away from zero and whose sum is distributed close to one. In this case, the \( \beta_j \)’s corresponding to the dominant components should be fixed and not estimated. For example, the models with \( \beta = (0, 0, -0.5)^\top \) and \( \beta = (0, 0, 0)^\top \) give very similar contours of constant probability in Figure 1 (c) and (d) and it will be difficult to estimate \( \beta_j \) in small samples. To decide whether or not to estimate \( \beta_j \), we suggest to first look at plots of the sample marginal distribution of each \( u_j, j = 1, 2, \ldots, p \).

If \( u_j \) is distributed close to zero with high right skewness, then \( \beta_j \) needs to be estimated (see, e.g., Figure 2 (a), (b), (d), and (e)). If the right skewness in \( u_j \) is moderate/small such as in Figure 2 (g) and (h), then a truncated Gaussian model may be appropriate where \( \beta_j = 0 \) is fixed and not estimated. Estimating \( \beta_j \) is more important when it is clear that \( \beta_j < 0 \) from the plots and in this case the model is effectively a continuous analogue of a discrete zero inflated model. In zero-inflated negative binomial and Poisson models, it only makes sense to include the zero inflation factor/component in the density if there is a large spike at zero which is not well explained by the other parameters in the model. See Section 5 for further discussion of practical identifiability.

Depending on the dataset, we may occasionally wish to fix all the \( \beta_j \) in advance, in cases where the dataset is highly concentrated in the interior of the simplex; and for datasets where there is no dominant component and all components of the \( u_j \) have many zeros, we may wish to estimate all of the \( \beta_j \). For
most datasets, however, we anticipate that the best option will be to estimate some of the components and fix some of the components.

4.3. All of the $\beta_j > 0$ Are Fixed

If we assume that $\beta_j > 0$ for $j = 1, 2, \ldots, p$ are all fixed and known in the true model $\theta_0$, then the weight function is not needed. That is, we can set $\hat{h}(x) = 1$ in Theorem 1 because the boundary integral over $\partial M$ is zero; see Appendix A.3 (supplementary material). In this case, the score matching estimator is equivalent to Mardia et al. (2016). However, the assumption that the density is zero on the boundary is too strong in general.

5. Simulation

In this section, we examine the performance of estimator (16) with $\hat{h}^2$ given by (12) for $p = 3$. In each case we simulated $R = 1000$ samples from model (3). In all cases $A_L$ and $b_L$ were set equal to (4) and we chose three different values of $\beta$: $\beta = (-0.8, -0.8, -0.5)^T$, $\beta = (-0.5, -0.5, -0.5)^T$ and $\beta = (0, 0, -0.5)^T$, the same choices of $\beta$ as in Section 2.3. For each of the three models, we calculated the score matching estimator for two different sample sizes: $n = 100$ and then $n = 1000$. The score matching estimator requires choice of the cap $a_c$. We considered a range of different values of $a_c$ which corresponded to the 1th, 5th, 10th, 20th, 50th, and 80th percentiles of the distribution of $\hat{h}$ defined at (11) (these are denoted by 1%, 5%, 10%, 20%, 50%, and 80% in Tables 1 and 2). These percentiles were calculated by simulating a single large sample from the model and calculating the percentiles of the empirical distribution of $\hat{h}$ defined at (11). For each sample and $a_c$ choice, we calculated the score matching estimator assuming (i) $\beta_1$ and $\beta_2$ are estimated with $\beta_3$ fixed (known); and (ii) $\beta_1$, $\beta_2$ and $\beta_3$ are all fixed. The columns EST and FIX in Tables 1 and 2 refer to case (i) and case (ii), respectively.

Tables 1 and 2 contain the estimated root mean squared errors (RMSEs) of the score matching estimator from the $R$ simulated samples. The estimated relative biases are given in Appendix A.10 (supplementary material) and are all small. As expected for each choice of $\beta$, the RMSE’s decrease as $n$ increases. When estimating $\beta_1$ and $\beta_2$ (see EST columns), the best choice of $a_c$ depends on $\beta$ (see highlighted columns) and ranges from 20% to 1%. Clearly, a well-chosen $a_c$ is essential when estimating $\beta_1$ and $\beta_2$ because the RMSE’s can be very inflated with a poor choice of $a_c$. When $\beta_1$ and $\beta_2$ are fixed (see FIX columns) any choice of $a_c$ between 1% and 20% gives similar small RMSEs for estimating $A_L$ and $b_L$ (choice of $a_c$ is less critical in this case).

Interestingly, when $\beta_1 = -0.8$ and $\beta_2 = -0.8$, the optimal $a_c$ for EST is 20% and in this case the RMSE’s for $A_L$ and $b_L$ are almost identical when comparing EST versus FIX. So, in this

| $a_c$ | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. |
|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| 1%   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 5%   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 10%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 20%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 50%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| 80%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |

| $\beta$ | $a_c$ | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. |
|---------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| (-0.8,-0.8,-0.5) | 1%   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.8,-0.8,-0.5) | 5%   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.8,-0.8,-0.5) | 10%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.8,-0.8,-0.5) | 20%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.8,-0.8,-0.5) | 50%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.8,-0.8,-0.5) | 80%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |

| $\beta$ | $a_c$ | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. | Est. | Fix. |
|---------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|------|
| (-0.5,-0.5,-0.5) | 1%   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.5,-0.5,-0.5) | 5%   |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.5,-0.5,-0.5) | 10%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.5,-0.5,-0.5) | 20%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.5,-0.5,-0.5) | 50%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
| (-0.5,-0.5,-0.5) | 80%  |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |      |
Table 2. Simulation results: RMSEs.

|         | $\alpha_c = 1\%$ |         | $\alpha_c = 5\%$ |         | $\alpha_c = 10\%$ |         | $\alpha_c = 20\%$ |         | $\alpha_c = 50\%$ |         | $\alpha_c = 80\%$ |
|---------|------------------|---------|------------------|---------|------------------|---------|------------------|---------|------------------|---------|------------------|
|         | EST              | FIX     | EST              | FIX     | EST              | FIX     | EST              | FIX     | EST              | FIX     | EST              | FIX     |
| $a_{11}$| 46.5             | 33.8    | 50.8             | 34.3    | 53.1             | 35.1    | 55.4             | 35.1    | 60               | 38.5    | 64.1             | 40.7    |
| $a_{22}$| 101              | 60.7    | 107              | 61.3    | 113              | 62      | 119              | 63.3    | 129              | 67.9    | 140              | 72.9    |
| $a_{12}$| 31.5             | 31.2    | 32.2             | 31.8    | 33               | 32.5    | 33.8             | 33.3    | 35.7             | 35.2    | 37.6             | 37.2    |
| $b_1$   | 17.7             | 8.6     | 20.3             | 8.1     | 23.7             | 8.8     | 28.6             | 13.3    | 42.2             | 13.7    | 52.3             | 15      |
| $b_2$   | 40.2             | 12.8    | 43               | 13      | 46.5             | 13.3    | 48.2             | 13.7    | 52.3             | 15      | 56.8             | 16.1    |
| $\beta_1$| 0.802            |         | 0.846            |         | 0.908            |         | 0.979            |         | 1.1              |         | 1.19             |         |
| $\beta_2$| 1.98             |         | 2.11             |         | 2.23             |         | 2.36             |         | 2.55             |         | 2.76             |         |

case, we can effectively fix $\beta$ at its estimate when doing inference and standard error calculations for $A_1$ and $b_L$ and estimation of $\beta_1$ and $\beta_2$ will not make much difference.

As $\beta_1$ and $\beta_2$ increases from $-0.8$ to $0$ the optimal RMSE's of $\beta_1$ and $\beta_2$ in the EST column increase substantially and the RMSE's for $b_L$ also get more inflated for EST. This is due to the fact that as $\beta$ increases there is less probability near the zero boundary and $\beta$ is becoming less identifiable for small samples.

The above simulation relates to $\rho = 3$ only. See Appendix A.11 (supplementary material) for a wider range of simulation settings which covers various different models and dimensions. Appendix A.11 (supplementary material) also contains a comparison of various different estimators, from which we concluded that estimator (16) with $\hat{h}$ given by (12) performed the best in general. These additional simulations also showed that the standard error estimator based on (20) performed well in small- to moderate-sized samples.

6. Multinomial Model With Latent Variables

In this case, we observe multivariate counts $x_i = (x_{i1}, x_{i2}, \ldots, x_{ip})$ $i = 1, 2, \ldots, n$, for $p$ taxa on $n$ subjects where the total count for each individual, $m_i = \sum_{j=1}^p x_{ij}$, is assumed known and treated as non-random. We assume that, for $i = 1, 2, \ldots, n$, each $x_i$ given $u_i \in \Delta_{p-1}$ and $m_i$ are independent with multinomial probability mass function given by $f(x_i|u_i) = m_i! \prod_{j=1}^p (u_{ij}/x_{ij})$. We also assume that the $u_i$ are iid for $i = 1, 2, \ldots, n$, each with probability density function given by (3).

This model is fully flexible which can be seen from the marginal moment structure:

$$\mathbb{E}(x_{ij}/m_i) = \mathbb{E}(u_{ij}) + \frac{\mathbb{E}(u_{ij}) - \mathbb{E}(u_{ij}^2)}{m_i}. \quad (21)$$

6.1. Parameter Estimation

We now consider parameter estimation in the multinomial latent variable model with latent variable distribution given by (6). The main goal here is to estimate the parameters of the latent variable distribution, that is, $A_1$, $b_L$ and those components of $\beta$ that we decide to estimate rather than fix in advance. In many applications of this multinomial latent variable model, including the microbiome dataset considered in the next section, the $m_i$ are all large compared with $n$, in which case it is reasonable to approximate the unobserved random vectors $z_{1i} = u_{1i}, \ldots, z_{ni} = u_{ni}$, in the notation of (6), by the observed vectors of proportions $\hat{z}_{ij} = \hat{u}_{ij} = m_i^{-1} x_{ij}$ where for $i = 1, \ldots, n$, the $x_{ij}$ are independent and each $x_i$ is a multinomial vector with number of trials $m_i$ and probability vector $z_i$. Then, we apply the score matching procedure to the observed data $\hat{z}_{11}, \ldots, \hat{z}_{nn}$.

Theoretical justification is given in Theorem 3.

Let $\hat{h}(z)^{2+e}$ denote the obvious modification of the weight functions $\tilde{h}(z)^2$ defined in Section 3.2; for example in (11),

$$\hat{h}(z)^{2+e} = \min(z_1^{1+e}, \ldots, z_p^{1+e}).$$

Write $\mathcal{B}_{\text{est}} \subseteq \{1, \ldots, p\}$ for the set of indices, $j$, of those $\beta_j$ we choose to estimate. In the multinomial latent variable model with latent variable distribution (3), consider the unknown parameters to be estimated, $A_1$, $b_L$ and $\beta_j, j \in \mathcal{B}_{\text{est}}$, contained in the parameter vector $\pi$.

**Theorem 3.** Let $\hat{\pi}$ denote the score matching estimator of $\pi$ based on the (unobserved) compositional vectors $z_1^1, \ldots, z_n^n$ and let $\hat{\pi}^\dagger$ denote the score matching estimator of $\pi$ based on the observed vectors of proportions $\hat{z}_1^1, \ldots, \hat{z}_n^n$. Assume that for some constants $C_1 > 0$ and $\alpha > 1$, $\inf_{i=1,\ldots,n} m_i \geq C_1 n^\alpha$.

1. Assume weight function (8) or (9) is used. Then as $n \to \infty$,

$$||n^{1/2} (\hat{\pi} - \pi) || = o_p(1). \quad (22)$$

2. Assume $\mathcal{B}_{\text{est}}$ is empty and the weight function (11) or (12) is used. Then (22) holds as $n \to \infty$.

3. Assume weight function $\hat{h}(z)^{2+e}$ is used, where $e \in [0, 2)$ and $\hat{h}(z)^2$ is one of the functions (11) or (12). Also assume (i) the marginal distribution of $z_j$ conditional on $z_j > 0$ is absolutely continuous with density $f_j$; (ii) for each $j \in \mathcal{B}_{\text{est}}$,...
\[ f_j(z) \leq C_0 z^{-\delta-1}, \] for some constants \( C_0, \delta > 0; \) and (iii) suppose that \( a \delta > 1. \) Then (22) holds as \( n \to \infty. \)

(IV) Assume weight function \( \tilde{h}(z)^{2+\epsilon} \) is used, where \( \epsilon \geq 2 \) and \( \tilde{h}(z)^2 \) is one of the functions (11) or (12). Then (22) holds as \( n \to \infty. \)

It is shown, under the conditions of Theorem 3, that \( \hat{\pi} \) based on weight functions (8), (9), (11) or (12) is asymptotically equivalent to \( \pi \) to leading order. Note that Theorem 3 does not assume that the population latent variable distribution is a PPI distribution, but if the PPI model is correct then \( \hat{\pi} \) and \( \pi \) are both consistent estimators of \( \pi \) under the conditions of Theorem 3. Moreover, for each \( j \), the population marginal distribution of \( z_j \) is permitted to have a non-zero point mass at \( z_j = 0; \) Theorem 3 applies even when the observed data has a large proportion of zeros; and there seems to be quite close correspondence between numerical results in Appendix A.11 (supplementary material) and Theorem 3. The estimator \( \hat{\pi} \) based on the weight functions (11) or (12) (with \( \epsilon = 0 \)) is not recommended for estimating \( \beta \) and performed poorly in simulations; see Appendix A.11 (supplementary material) for further discussion. If the population distribution is of PPI model form then we can put \( \delta = 2(1 + \beta_0) \) where \( \beta_0 = \inf_{j \in B_{\text{all}}} \beta_j. \)

Finally, we mention that estimation based on the \( \hat{\beta} \) was fixed in advance as long as \( n \) was relatively small compared to the \( m_j. \)

### 6.2. Marginal Moment Matching Estimator

If \( m_j > n \) typically fails to hold then \( \hat{\pi} \) defined in Section 6.1 is not appropriate. An alternative score matching method based on weight function (8) and matching marginal moments is described in Appendix A.8 (supplementary material). This estimator is consistent when the \( m_j \) have a common bound as \( n \to \infty \). However, it was found to be inefficient in the simulation study reported in Appendix A.11 (supplementary material) and is not recommended when \( n \) is small.

One way to improve the efficiency of this marginal moment matching estimator is to assume that \( \beta_j > 0 \) for \( j = 1, 2, \ldots, p \) are all fixed and then replace the weight function (8) by \( \tilde{h}(z)^2 = 1. \) This will result in a consistent estimator for the multinomial model with PPI model latent variables when the \( m_j \) have a common bound as \( n \to \infty. \) This model has roughly similar properties to the logistic normal multinomial model mentioned in the Introduction. That is, the latent variables model general location and covariance structures and the latent variable density is zero on the boundary of the simplex. This represents a major computational advance for multinomial latent variable models because the marginal log-likelihood is not tractable for the logistic normal multinomial model.

### 6.3. Simulation Based Estimators of the \( \beta_j \)’s.

The points made in Section 4.2 concerning the treatment of the \( \beta_j \)’s are relevant in the multinomial PPI latent variable model. In particular, it is often reasonable to treat some of the \( \beta_j \)’s as tuning parameters and not necessary to estimate all of them. Estimation of the \( \beta_j \)’s via score matching is challenging in the multinomial latent variable model; see the comments after Theorem 3 and the discussion of Model 16 in Appendix A.11 (supplementary material). It is often easier to use simulation based methods to estimate the \( \beta_j \)’s; see Appendix A.9 (supplementary material) and Section 7 for further details.

### 7. Data Application

We analysed a subset of the longitudinal microbiome dataset obtained from a study carried out in a helminth-endemic area in Indonesia (Martin et al. 2018). Stool samples were collected from 150 subjects in the years 2008 (pretreatment) and in 2010 (post-treatment). The 16s rRNA gene from the stool samples was processed and resulted in counts of 18 bacterial phyla. Whether or not an individual was infected by helmint was also determined at both time points (see Martin et al. 2018, fig. 4).

Here, we restricted the analysis to the year 2008 for individuals infected by helmints which resulted in a sample size of \( n = 94 \), and we treated these individuals as being independent. Martin et al. (2018) analyzed the five most prevalent phyla and pooled the remaining categories. We analysed a different set of phyla and for demonstrative purposes, we deliberately included two with a high number of zeros. In our analysis, we have five categories (in this order): TM7, Cyanobacteria/Chloroplast, Actinobacteria, Proteobacteria and pooled; pooled contains mostly Firmicutes which is the most abundant category. The relative abundances of Actinobacteria, Proteobacteria, TM7 and Cyanobacteria/Chloroplast are 12%, 10%, 0.1%, and 0.3%, respectively. The categories TM7 and Cyanobacteria/Chloroplast contained 38% and 42% zeros, respectively. Note that both categories TM7 and Cyanobacteria/Chloroplast also contained outliers and we deleted two observations prior to the analysis resulting in a final sample size of \( n = 92 \). Let \( x_{ij}, i = 1, 2, \ldots, 92 \) and \( j = 1, 2, 3, 4 \), and 5 represent the sample counts with total count \( m_j = 2000 \). We assume that these counts follow the multinomial latent variable model as defined in Section 6.

The sample proportions were calculated as follows: \( \hat{\pi} = \hat{z}_{ij} = x_{ij}/m_j, \) where \( i = 1, 2, \ldots, 92 \) and \( j = 1, 2, 3, 4, 5 \). The marginal distributions of the sample proportions are plotted in Figure 3 which gives strong evidence that \( \beta_1 < 0 \) and \( \beta_2 < 0 \), since there is a large mode at zero for both \( \hat{u}_1 \) and \( \hat{u}_2 \). All four sample proportions \( \hat{u}_{ij}, j = 1, 2, 3, 4 \) are right skewed with mode either at or close to zero and \( \hat{u}_5 \) is distributed close to one. This suggests that \( \beta_1 = 0 \) and we fixed \( \beta_2 = 0 \) prior to estimation.

The estimator \( \hat{\pi} \) defined in Section 6.1 based on the weight function (11) or (12) is not expected to work well here for estimating \( \beta \) because the condition \( a \delta > 1 \) in Theorem 3 does not look credible. [Although we have not proved \( a \delta > 1 \) is a necessary condition for part (III) of Theorem 3 to hold, we believe this condition is not far from being necessary.] Also, since \( p \) is not small (8) is expected to lead to an inefficient estimator for \( \beta \). [Tables 15 and 16 in Appendix A.11 (supplementary material) show that the RMSE’s based on (8) can be very inflated compared with the moment and maximum likelihood estimator when \( p > 3 \) - To estimate \( \beta \) we applied the methodology.
First we chose a set of \( \beta \) values on small grid ranges for \( \beta_1, \beta_2, \) and \( \beta_4 \) and we preset \( \beta_3 = 0 \) and \( \beta_5 = 0 \) (note that from Figure 3 \( \hat{u}_{13} \) looks close to a truncated Gaussian type and \( \hat{u}_{15} \) is distributed close to one and \( \beta_5 \) needs to be fixed). Then for each grid combination of \( \beta_1, \beta_2, \) and \( \beta_4, \) treating these as fixed, we then calculated the estimate of \( \beta_1 \) using \( \hat{r}^T \) with weight function (12). The final estimates of \( \beta_1, \beta_2, \) and \( \beta_4 \) corresponded to the cases which gave the smallest Kolmogorov–Smirnov distances based on comparing the marginal distributions of the sample data with rounded simulated data under the fitted PPI models. The rounded simulated data were calculated as follows: \( \hat{u}_{ij} = \text{round}(\hat{u}_{ij}/m_i) \) where \( \hat{u}_{ij} \) denotes the simulated proportion under the fitted PPI model for \( i = 1, 2, \ldots, 100,000. \) Rounding to the nearest integer on the count scale was needed in order to help mimic the discreteness in the sample data. The final estimates are \( \hat{\beta}_1 = -0.80, \hat{\beta}_2 = -0.85, \) and \( \hat{\beta}_4 = -0.2. \)

Table 3 contains the estimates of \( A_L \) using \( \hat{r}^T \) based on weight function (12) with \( a_c = 0.01 \) and with \( b_1 \) omitted. This choice of \( a_c \) corresponded to approximately the 62th percentile of the \( h \) function (11) under the fitted model; see Model 1 values in Table 11 in Appendix A.11 (supplementary material). The standard errors (SE) were calculated using (20). All the diagonal terms in \( A_L \) are significant since \( |\text{estimate}/\text{SE}| > 2. \) Most of the off diagonal terms in \( A_L \) are not significant, with the exception of \( a_{13}. \) This is not surprising because there does appear to be a positive (Pearson) correlation between TM7 and Actinobacteria (see Figure 4 (e) and (f)). For comparative purposes, we also fitted the model including the \( b_1 \) term which is given in Table 4. This shows that the \( b_1 \) term is not significant, as expected.

For the purpose of model diagnostics, we simulated a large sample of size \( n = 100,000 \) from the fitted model with parameters set equal to the estimates in Table 3. As in the definition of the \( \hat{u}_{ij} \), we rounded the simulated proportions to the nearest integer on the count scale. Figure 4 (a)–(d) contains qq-plots of the simulated proportions TM7, Cyanobacteria/Chloroplast, Actinobacteria, Proteobacteria versus the true sample proportions (compares the four sets of marginal distributions). The qq-plots are close to the \( y = x \) line demonstrating that the model fits reasonably well marginally. Figure 4 (e) is a plot of the sample Actinobacteria proportions versus the sample TM7 proportions and Figure 4(f) is the corresponding simulated version (the first 1000 simulated values were plotted), which shows a similar pattern to the true sample ones. The fitted model is clearly able to model positive (Pearson) correlations between variables, unlike the less flexible Dirichlet distribution.

For comparative purposes, we also fitted a 5-dimensional Dirichlet distribution to the sample proportions based on the
moment estimator and we simulated a large sample of size $n=100,000$ from this fitted model and rounded them as in the definition of the $\hat{u}'_j$. Table 5 contains the results of Kolmogorov–Smirnov tests based on comparing the simulated proportions to the true sample proportions for each category separately. The new model is clearly a much better fit than the Dirichlet model based on the Kolmogorov–Smirnov test statistic values. The $p$-value for Proteobacteria in the new model is small (0.025), but this is not surprising since the tail is a bit curved in Figure 4 (d). The Dirichlet fits the first category very well, at the expense of the four other categories where it gets the scale and shape of the marginal distributions completely wrong. Table 6 compares the true sample moments to the simulated sample moments under both the new PPI model and the Dirichlet model. The new model does a very good job overall of getting both the first- and second-order moments approximately correct. The Dirichlet model gets the means correct (this is not surprising because the moment estimator was applied), but the standard deviations are grossly underestimated in most of the categories which implies that the Dirichlet distribution is getting the scale wrong. Clearly, the sample proportions are over-dispersed when compared with a Dirichlet distribution.

### 8. Conclusion

In 1986 John Aitchison published a seminal book on compositional data analysis. Since then many articles have been written on the topic, as this continues to be a challenging problem in the age of complex modern biological and ecological data applications. We proposed a new flexible model for compositional data and we derived score matching estimators for the parameters in the model. These estimators were scalable and computationally efficient and handled many zeros in the data. The simulation study demonstrated that with an appropriate choice of weight function, the score matching estimators had small bias and were generally reasonably efficient. Our analysis of a real microbiome dataset highlighted the deficiencies of current approaches for summarising microbiome data. The Dirichlet distribution was not adequate because the data were over-dispersed and the logistic normal distribution could not be applied directly to analyse the sample proportions due to the large number of zeros in the data. In comparison, our new model was able to accommodate proportions concentrated near zero with high right skewness.

### Acknowledgments

JLS was supported by an Australian Research Council Discovery Early Career Researcher Award. We thank Jeanine Houwing-Duistermaat for...
useful discussions regarding methodology; Jeanine and Ivonne Martin for their help with the data; and an AE and three referees for their constructive criticisms which have helped to improve the structure and presentation of the article.

Supplementary Material

Supplementary material online includes proof of theorems and lemmas, detailed sampling algorithms, additional numerical results and discussion for simulations, and codes for reproducing the results in this article.

References

Aitchison, J. (1986), *The Statistical Analysis of Compositional Data*, Monographs on Statistics and Applied Probability, Vol 25. London: Chapman & Hall. [1811]

Bear, J., and Billheimer, D. (2016), "A Logistic Normal Mixture Model for Compositional Data Allowing Essential Zeros," *Austrian Journal of Statistics*, 45, 3–23. [1811]

Billheimer, D., Guttorp, P. and Fagan, W. F. (2001), Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96, 1205–1214. [1812]

Butler, A., and Glasbey, C. (2008), "A Latent Gaussian Model for Compositional Data With Zeros," *Applied Statistics*, 57, 505–520; correction, 58 (2009), 141. [1811]

Hyvarinen, A. (2005), "Estimation of Non-Normalised Statistical Models by Score Matching," *Journal of Machine Learning Research*, 6, 695–709. [1812,1813]

Hyvarinen, A. (2007), "Some Extensions of Score Matching," *Computational Statistics and Data Analysis*, 51, 2499–2512. [1812,1813]

Krzyztofowicz, R., and Reese, S. (1993), "Stochastic Bifurcation Processes and Distributions of Fractions," *Journal of the American Statistical Association*, 88, 345–354. [1811]

Leininger, T. J., Gelfand, A. E., Allen, J. M., and Silander Jr., J. A. (2013), "Spatial Regression Modeling for Compositional Data With Many Zeros," *Journal of Agricultural, Biological, and Environmental Statistics*, 18, 314–334. [1811]

Li, H. (2015), "Microbiome, Metagenomics, and High-Dimensional Compositional Data Analysis," *The Annual Review of Statistics and Its Application*, 2, 73–94. [1812]

Liu, S., Kanamori, T., and Williams, D. J. (2020), "Estimating Density Models with Truncation Boundaries." Available at [https://arxiv.org/abs/1910.03834]. [1812,1813]

Mardia, K. V. (2018), "A New Estimation Methodology for Standard Directional Distributions," in 2018 21st International Conference on Information Fusion (FUSION), New York: IEEE, 724–729. [1812]

Mardia, K. V., Kent, J. T. and Laha, A. K. (2016), "Score Matching Estimators for Directional Distributions." Available at: [https://arxiv.org/abs/1604.08470]. [1812,1813,1816,1818]

Martin, I., Uh, H-W., Supali, T., Mitreva, M., and Houwing-Duistermaat, J. J. (2018), “The Mixed Model for the Analysis of a Repeated-Measurement Multivariate Count Data,” *Statistics in Medicine*, 38, 2248–2268. [1812,1820]

Ongaro, A., Migliorati, S., and Ascarì, R. (2020), "A New Mixture Model on the Simplex," *Statistics and Computing*, 30, 749–770. [1811]

Scealy, J. L., and Welsh, A. H. (2011), "Regression for Compositional Data by Using Distributions Defined on the Hypersphere," *Journal of the Royal Statistical Society*, Series B, 73, 351–375. [1811,1814]

——— (2014), "Fitting Kent Models to Compositional Data With Small Concentration," *Statistics and Computing*, 24, 165–179. [1811,1814]

——— (2017), "A Directional Mixed Effects Model for Compositional Expenditure Data," *Journal of the American Statistical Association*, 112, 24–36. [1814]

Stewart, C., and Field, C. A. (2010), "Managing the Essential Zeros in Quantitative Fatty Acid Signature Analysis," *Journal of Agricultural, Biological, and Environmental Statistics*, 16, 45–69. [1811]

Takasu, Y., Yano, K., and Komaki, F. (2018), "Scoring Rules for Statistical Models on Spheres," *Statistics and Probability Letters*, 138, 111–115. [1812]

Tsagris, M., and Stewart, C. (2020), "A Folded Model for Compositional Data Analysis," *Australian and New Zealand Journal of Statistics*, 62, 249–277. [1811]

Yu, S., Drton, M., and Shojaie, A. (2019), "Generalised Score Matching for Non-Negative Data," *Journal of Machine Learning Research*, 20, 1–70. [1812,1814,1815,1817]

Yu, S., Drton, M., and Shojaie, A. (2020), "Generalised Score Matching for General Domains." Available at [https://arxiv.org/abs/2009.11428]. [1812]

Zhang, J., and Lin, W. (2019), "Scalable Estimation and Regularization for the Logistic Normal Multinomial Model," *Biometrics*, 75, 1098–1108. [1812]