The High-Dimensional Simplex: Open Problems in Compositional Data Analysis

VT CoDA Reading Group

March 2026

1 Background

Potential Lead: Charlie Keglovitz

Compositional data arise whenever observations represent parts of a whole; they are often expressed as proportions, relative abundances, or probability vectors constrained to the positive simplex

$$\Delta^{K-1} = \{(\pi_1, \pi_2, \ldots, \pi_K) : \pi_i \geq 0,\ \sum_{i=1}^{K} \pi_i = 1\}. \tag{1}$$

Such data are increasingly prevalent across scientific disciplines, spanning settings that range from time use surveys [Chastin et al., 2015, Dumuid et al., 2018] to microbiome profiles [Gloor et al., 2017, Li, 2015]. In ecology, species or taxon relative abundances naturally form a composition [Douma and Weedon, 2019]. In genomics, sequencing data yield read count tables that, after normalization, live on the simplex [Quinn et al., 2018]. In text analysis, latent Dirichlet allocation describes each document by a vector of topic proportions that sums to one [Blei, 2012, Blei et al., 2003], and in mixture modeling more broadly, the vector of component weights is itself simplex-valued [McLachlan et al., 2019]. Some fields refer to multivariate count observations with a fixed total as compositions, whereas others consider the specification in (1) as compositions; we refer to the former as "count compositional" and the latter as "continuous compositional" data where distinction is needed. Although compositional data are common across many scientific disciplines, the distinctive challenges of the high-dimensional regime remain comparatively underexplored. In this manuscript, we review and motivate the modeling of compositional data in the high-dimensional simplex (i.e., $K \gg 100$), where there may be hundreds to thousands of components.

What distinguishes compositional data from standard multivariate data is the *sum-to-one constraint*: the components π_i are not independent, and their total is fixed. This seemingly simple constraint has deep consequences for both the geometry of the sample space and for statistical inference. Conventional multivariate methods developed for unconstrained data in $\mathbb{R}^K$ are not directly applicable, and naïve application leads to spurious correlations and misleading conclusions [Aitchison, 1986]. The modern study of compositional data analysis (CoDA) was developed in response to this difficulty and was largely systematized by Aitchison [1986, 1982], who emphasized the role of log-ratio transformations in mapping

simplex-valued data to the unconstrained Euclidean space before applying standard tools such as multivariate normals.

Today, log-ratio approaches remain incredibly popular because of their ease of use, but their convenience comes at a cost in sparse high-dimensional regimes; log-ratio transformations are undefined when components are zero-valued, which is inevitable in high-dimensional settings. Zero-replacement methods attempt to address this by substituting a small positive value for each zero, but such methods introduce bias in high-dimensional compositions where zeros are increasingly prevalent [Martín-Fernández et al., 2003].

The Dirichlet distribution is the natural starting point for modeling compositional data because it is defined directly on the simplex Δ^{K-1} and is the conjugate prior for the multinomial [Ongaro and Migliorati, 2013]. These properties have made it the dominant choice for prior modeling in compositional data analysis. More recent work has developed methods that remain native to the simplex, drawing on constructive definitions of the Dirichlet distribution to build flexible prior families directly on Δ^{K-1}. However, the Dirichlet distribution induces a restrictive covariance structure, limiting its ability to represent the realistic correlation patterns often observed in high-dimensional compositional data [Aitchison, 1986, Tang and Chen, 2019]. It also lacks mechanisms for adapting to sparsity, often spreading mass over many components even when the true signal is concentrated [Bhattacharya et al., 2015, Chen and Li, 2013].

An alternative geometric approach is to model compositions as directional observations on the surface of the unit hypersphere [Scealy and Welsh, 2011, Schwob et al., 2024]. This approach offers two main advantages. First, it can flexibly model covariance structure, as opposed to the Dirichlet and its extensions. Second, unlike log-ratio transformations, the square-root transformation that maps compositions to the surface of the nonnegative orthant of the unit hypersphere is defined at zero, a substantial benefit under sparsity. The principle disadvantage is that directional distributions are not natively generative for compositional data; the two remedies are folding, which introduces noisy missing sign imputation [Figueiredo and Figueiredo, 2026, Scealy and Welsh, 2011], and truncation, which is computationally prohibitive in high dimensions [Schwob et al., 2024]. Directional models are used less frequently, as they lack conjugacy and most datasets are recorded in counts rather than proportions.

The problems that motivate this review are specifically high-dimensional (i.e., when the number of categories K is large, often comparable to or exceeding the sample size). This regime introduces challenges that are qualitatively different from the classical low-dimensional setting. A central problem is the presence of a large number of zero or near-zero components, particularly as K grows. Such zeros fall into two distinct types: (i) sampling zeros, which arise when a component has a very low but nonzero value and is simply not observed due to finite sampling or sampling error, and (ii) structural zeros, which reflect a true absence of that component in the system. Distinguishing between the two is difficult but critical, as they require fundamentally different modeling approaches. An additional challenge in the high-dimensional simplex is that, generally, only a small number of categories account for most of the compositional signal. Log-ratio approaches struggle with zero-valued components, the standard Dirichlet prior and its classical extensions are poorly equipped to handle either of these challenges at scale, and truncated directional approaches are computationally prohibitive for the high-dimensional simplex.

To the best of our knowledge, this is the first review to focus specifically on high-dimensional compositional data analysis from a Bayesian perspective. We survey current problems and open gaps in five key areas: model evaluation, compositional-on-compositional regression, sparsity-inducing priors, positive correlation structure, and prior sensitivity.

2 Motivating Applications

Lead: Michael

(3/31: discuss potential applications that we will use to reveal problems in existing methods)

(i) Microbiome

- EMP; the largest microbiome dataset that Michael knows; easy to benchmark against other methods because this is very commonly analyzed
- TARA Oceans; also commonly studied; marine microbiomes are notoriously trickier; data doesn't seem as accessible, though
- Brian Badgley/Bryan Brown might have some local ones, too

(ii) Genomics

- 10x Genomics PBMC data set; gene expressions on the simplex with $K \approx 20,000$ genes and zero-inflation (around 85% zeros)

(iii) Nutritional Epidemiology

- NHANES data; dietary recall data is expressed as compositions; **many** compositional analyses are 24-hour type data, as well

(iv) Topic Modeling

- 20 newsgroups text data set; non-biological example; 20000 documents across 20 topics; with vocabulary as K, there are over 100000 unique words; with latent topics as K (as in LDA), K could be smaller (50-100); either approach results in incredible sparsity and positive correlations

3 Current Problems

(3/17: please claim a subsection below and fill out the information by our next meeting on 3/31; feel free to add more, claim multiple, and share sections, but we need one lead for each problem)

3.1 Transformations for Compositional Data

One of the defining features of compositional data is that the information is relative. The parts are meaningful through their ratios to one another, not through their absolute magnitudes [Aitchison, 1986, 1982]. The unit-sum constraint is therefore not just a technical restriction, but a way of representing data that is invariant to scale. If every component of a positive vector is multiplied by the same positive constant before closure, the resulting composition is unchanged. This scale invariance is what distinguishes compositional

data from ordinary multivariate data, and it explains why applying standard methods directly to raw proportions can be misleading [Aitchison, 1986, Pawlowsky-Glahn et al., 2015]. Because the parts must sum to one, variation in any single component is necessarily expressed relative to variation in the others. As a result, standard methods applied to raw proportions can produce spurious correlations and can violate subcompositional coherence, since relationships among a fixed set of parts may appear to change when other parts are removed from the analysis [Aitchison, 1982, Gloor et al., 2017].

3.1.1 Aitchison Geometry

Aitchison geometry provides the standard mathematical framework for compositional data analysis, treating compositions according to their relative structure, rather than as ordinary vectors in $\mathbb{R}^K$ [Egozcue et al., 2003, Pawlowsky-Glahn and Egozcue, 2001, Pawlowsky-Glahn et al., 2015]. In this geometry, peturbation plays the role of addition by combining compositions, while powering plays the role of scalar multiplication by rescaling a composition. Together with the Aitchison inner product, these operations allow the simplex Δ^{K-1} to be viewed as having the structure of a $K-1$ dimensional Euclidean space.

This Euclidean structure is not the usual geometry inherited from $\mathbb{R}^K$. Because it is built from log-ratios, distances measure relative differences among parts rather than absolute differences between raw proportions. For two compositions $\pi^{(1)}, \pi^{(2)} \in \Delta^{K-1}$, the Aitchison distance is

$$d_A(\pi^{(1)}, \pi^{(2)}) = \sqrt{\frac{1}{2K} \sum_{i=1}^{K} \sum_{j=1}^{K} \left[\log\left(\frac{\pi_i^{(1)}}{\pi_j^{(1)}}\right) - \log\left(\frac{\pi_i^{(2)}}{\pi_j^{(2)}}\right) \right]^2}$$

The Aitchison distance is scale invariant, permutation invariant, and subcompositionally dominant, in the sense that a subcomposition is never rather apart than the full composition containing it [Aitchison et al., 2000, Quinn et al., 2018]. Few competing dissimilarities satisfy all four; Aitchison et al. [2000] note that none of the metrics in routine use for hierarchical clustering does. Because the distance is defined on the open simplex, it diverges as any part approaches zero, a property that becomes the central difficulty in Section 3.1.6.These properties make Aitchison geometry well suited for statistical modeling, but they also impose a clear requirement. Methods for compositional data should respect the fact that the information lies in ratios. Transformation-based approaches address this by mapping compositions into Euclidean coordinates where standard statistical tools apply, and then mapping results back to the simplex when needed [Aitchison, 1986, Pawlowsky-Glahn et al., 2015].

3.1.2 The Additive Log-Ratio Transformation

The additive log-ratio (alr) transformation is one of the simplest ways to represent a composition in ordinary Euclidean coordinates [Aitchison, 1986, 1982]. It chooses one part as a reference and expressing every other part relative to it. If π_K is used as the reference part, then

$$\text{alr}(\pi) = \left[\log\left(\frac{\pi_1}{\pi_K}\right), \ldots, \log\left(\frac{\pi_{K-1}}{\pi_K}\right) \right]^\top$$

The alr transformation maps the simplex Δ^{K-1} one-to-one onto $\mathbb{R}^{K-1}$, which makes it convenient for modeling [Aitchison, 1986]. It is especially natural in settings where one

component has a clear role as the reference category, so that each coordinate reads directly as the log-abundance of a part relative to a fixed and scientifically meaningful baseline. It is worth noting that the applied time-use literature, where a natural reference part might seem available, has largely declined to use alr for exactly the reason given below: Chastin et al. [2015] and Dumuid et al. [2018] both analyze the 24-hour composition of sleep, sedentary time, light activity, and moderate-to-vigorous activity in isometric coordinates, with Dumuid et al. [2018] rejecting alr on the grounds that its non-isometry "violates distances and angles in the Aitchison geometry."

The main limitation is that alr coordinates are oblique rather than orthonormal with respect to Aitchison geometry, so the ordinary Euclidean metric on alr space is not the Aitchison metric [Egozcue et al., 2003]. The consequence is not that alr discards information: Aitchison et al. [2000] show that equipping alr space with the quadratic form $\Delta_R(y, Y) = \{(y - Y)^\top H^{-1}(y - Y)\}^{1/2}$, where $H_{ij} = 2$ for $i = j$ and 1 otherwise, recovers the Aitchison distance exactly and renders the choice of denominator immaterial. The difficulty is that this correction is not standard practice [Egozcue et al., 2003], so that in applied work Euclidean distances, model coefficients , and interpretations computed naively in alr coordinates do depend on which part is placed in the denominator. This dependence may be acceptable when the reference component has a clear meaning. It becomes more problematic in sparse, high-dimensional settings, where no single part can be seen naturally as the reference and where a rare or unstable reference part can introduce substantial noise [Li, 2015].

3.1.3 The Centered Log-Ratio Transformation

The centered log-ratio (clr) transformation avoids choosing a single reference part, comparing each component to the geometric mean of the full composition [Aitchison, 1986, 1982]. For a composition $\pi \in \Delta^{K-1}$,

$$\mathrm{clr}(\pi)_i = \log\left(\frac{\pi_i}{g(\pi)}\right), \quad i = 1, \ldots, K, \quad g(\pi) = \left(\Pi_{i=1}^K \pi_i\right)^{1/K}$$

Equivalently,

$$\mathrm{clr}(\pi) = \left(I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top\right)\log(\pi)$$

where $\log(\pi)$ is applied componentwise. The clr transformation treats all parts symmetrically and preserves Aitchison distances. That is, ordinary Euclidean distances between clr-transformed compositions are equal to Aitchison distances on the simplex [Aitchison, 1986, Pawlowsky-Glahn et al., 2015].

This symmetry comes with a constraint. The clr-transformed vector satisfies

$$\sum_{i=1}^K \mathrm{clr}(\pi)_i = 0$$

so clr-transformed observations lie in a $K - 1$ dimensional hyperplane of $\mathbb{R}^K$ rather than in all of $\mathbb{R}^K$, and the clr covariance matrix is consequently singular [Aitchison, 1986, Egozcue et al., 2003].

Despite this singularity, the clr transformation is useful in high-dimensional settings because each coordinate has a simple interpretation as the log of one part relative to the geometric

mean of all parts [Gloor et al., 2017]. The difficulty is that many standard methods assume unconstrained Euclidean coordinates, while clr coordinates always satisfy a zero-sum constraint. In penalized regression, this means that sparsity is imposed on coordinates that cannot vary independently, so variable selection and coefficient interpretation must be handled carefully [Li, 2015]. In graphical models, precision-matrix estimation is not directly defined because the clr covariance matrix has no ordinary inverse. Methods applied to clr coordinates must therefore account for this constrained geometry or treat the resulting estimates as approximations made outside the usual Euclidean assumptions.

3.1.4 The Isometric Log-Ratio Transformation and Balances

The isometric log-ratio transformation addresses the singularity of clr coordinates by representing a composition in an orthonormal coordinate system on the simplex [Egozcue et al., 2003]. Let $V \in \mathbb{R}^{K\times(K-1)}$ be a contrast matrix satisfying

$$V^\top V = I_{K-1}, \quad VV^\top = I_K - \frac{1}{K}\mathbf{1}\mathbf{1}^\top$$

The ilr transformation is defined as

$$\operatorname{ilr}(\pi) = V^\top \log(\pi) = V^\top \operatorname{clr}(\pi)$$

For any two compositions $\pi^{(1)}$ and $\pi^{(2)}$ it is an exact isometry,

$$||\operatorname{ilr}(\pi^{(1)}) - \operatorname{ilr}(\pi^{(2)})||_2 = d_A(\pi^{(1)}, \pi^{(2)})$$

The ilr transformation therefore produces $K-1$ unconstrained Euclidean coordinates while preserving the geometry of the simplex, which makes it attractive for methods that require ordinary Euclidean coordinates, such as distance-based analysis, regression, and dimensionality reduction [Egozcue et al., 2003]. Whether this theoretical advantage translates into practical benefit is contested though. Greenacre and Grunsky [2019] give a three-part example in which a relationship that is significant and visually evident in the ternary diagram is nullified when the same contrast is expressed as an ilr coordinate, because the geometric mean in the denominator varies with the internal composition of the group it summarizes.

The main limitation is interpretability. The ilr coordinates are not individual parts; they are balances, meaning log-ratios between groups of parts [Egozcue and Pawlowsky-Glahn, 2005]. Different choices of the orthonormal basis V produce different sets of balances, even though all such choices preserve the same Aitchison geometry [Egozcue and Pawlowsky-Glahn, 2005, Greenacre and Grunsky, 2019]. When there is meaningful domain structure, this basis dependence can be useful. A sequential binary partition can encode interpretable balances, such as contrasts associated with internal nodes of a phylogenetic tree in microbiome studies or levels of a hierarchical taxonomy in ecology [Egozcue and Pawlowsky-Glahn, 2005, Silverman et al., 2017]. Pivot coordinates provide a related construction that compares one part to the remaining parts in an orthonormal coordinate system [Fišerová and Hron, 2011]. In generic high-dimensional settings, however, there may be no natural basis to choose. The number of distinct sequential binary partitions of a K-part composition is $(2K-2)!/(2^{K-1}(K-1)!)$, already exceeding 3.4×10^7 at $K=10$, so an automatic choice is unavoidable and is rarely scientifically meaningful [Greenacre and Grunsky, 2019].

Greenacre and Grunsky [2019] argue further that a balance should not be read as a contrast between groups of parts at all, since its value depends on the relative composition within each geometric mean; they show that including or excluding a single rare part shifts the balance substantially while leaving the corresponding amalgamation log-ratio almost unchanged.

3.1.5 Alternative and Partial Transformations

The alr, clr, and ilr transformations are the classical log-ratio tools for compositional data analysis, but they are not always easy to use in high-dimensional sparse settings [Li, 2015, Quinn et al., 2018]. Several alternatives have therefore been developed to address sparsity, interpretability, and other challenges.

Pairwise log-ratio methods avoid choosing a single reference part by using contrasts of the form $\log(\pi_i/\pi_j)$ for many or all pairs of components [Lovell et al., 2015, Quinn et al., 2017]. These methods preserve the ratio-based logic of compositional data analysis and are widely used in proportionality-based association screening and differential abundance analysis [Quinn et al., 2017, 2018]. Their main limitation is scale with the number of distinct pairwise ratios being $K(K-1)/2$, which becomes very large when the number of parts is high [Lovell et al., 2015].

Other approaches modify the log-ratio transformation itself. Robust clr variances replace or adjust the geometric mean to reduce sensitivity to zeros and extremely rare components, while power and Box-Cox type transformations provide a broader class of alternatives that move between raw proportions and log-ratio coordinates [Rayens and Srinivasan, 1991]. The α-transformation used by Tsagris [2025] is one such family, indexed so that the log-ratio transformation arises as a limiting case and $\alpha \neq 0$ admits zeros. These methods can be useful when zeros are too common for strict log-ratio analysis to be stable, although they usually relax some of the exact geometric properties of Aitchison analysis [Quinn et al., 2018].

A common non-log-ratio alternative is the Hellinger transformation [Legendre and Gallagher, 2001]. For a composition π, define

$$h(\pi)_i = \sqrt{\pi_i}, \quad i = 1, \dots K$$

The transformed vector lies on the positive orthant of the unit hypersphere because

$$\sum_{i=1}^{K} h(\pi)_i^2 = 1,$$

a representation that also connects compositional data to spherical and directional models [Scealy and Welsh, 2011]. The corresponding Hellinger distance between two compositions is

$$d_H(\pi^{(1)}, \pi^{(2)}) = \left\{ \sum_{i=1}^{K} \left(\sqrt{\pi_i^{(1)}} - \sqrt{\pi_i^{(2)}} \right)^2 \right\}^{1/2}$$

in the form used by Legendre and Gallagher [2001], which attains its maximum at $\sqrt{2}$ when two compositions share no support; the probabilistic literature oftentimes carries a factor $1/\sqrt{2}$ so that the distance is bounded by one. The choice is immaterial for model

ranking but must be stated, since the two differ by a constant that propagates into any squared-error loss build from them. The Hellinger transformation is common in ecology and microbiome analysis because it is defined at zero, and because it occupies a middle ground on the weighting of rare parts: it gives them more weight than raw Euclidean distance but far less than the chi-squared metric preserved by correspondence analysis [Legendre and Gallagher, 2001, Table 2]. In their simulated gradient it produced the highest coefficient of determination among the distances compared, with a monotone distance gradient and little horseshoe effect [Legendre and Gallagher, 2001]. Its limitation is that it does not preserve Aitchison geometry and does not represent the scale-invariant ratio structure that defines compositional information [Aitchison, 1986]. For this reason, the Hellinger transformation is frequently viewed as a useful abundance-based transformation rather than a truly compositional one [Quinn et al., 2018].

3.1.6 Zeros and the Boundary of the Simplex

Zeros are one of the main obstacles for transformation-based compositional data analysis. Classical log-ratio transformations require strictly positive components, but in high-dimensional sparse settings, zeros are often a central feature of the data rather than a rare exception [Martín-Fernández et al., 2003]. In sequencing applications, zeros may arise from limited sampling depth, detection limits, or genuine structural absence, and these mechanisms carry different statistical meanings [Gloor et al., 2017, Silverman et al., 2020]. A sampling zero represents a component that is present but not observed, whereas a structural zero represents a component that is truly absent from the underlying system. Treating these two cases as the same can distort estimation, prediction, and interpretation [Martín-Fernández et al., 2003, Silverman et al., 2020].

Because log-ratio methods are undefined when any component is zero, one common response is to replace zeros with small positive values before transformation [Martín-Fernández et al., 2003]. A widely used approach is multiplicative replacement, in which zero components are assigned small positive values and the nonzero components are rescaled to preserve closure [Martín-Fernández et al., 2003]. If $Z = \{j : \pi_j = 0\}$ denotes the zero set, one version is

$$\pi_i^* = \begin{cases} \delta_i, & i \in Z \\ \left(1 - \sum_{j \in Z} \delta_j\right) \pi_i, & i \notin Z \end{cases}$$

where each δ_i is a small replacement mass chosen by an empirical or Bayesian rule [Martín-Fernández et al., 2003, Martín-Fernández et al., 2014]. This procedure preserves ratios among the nonzero parts, but it necessarily inserts artificial mass into the zero components. When zeros are rare, the resulting distortion may be small; when zeros are frequent, as in microbiome, single-cell, text, and other high-dimensional abundance data, the replacement values can strongly influence the conclusions drown from the data [Li, 2015, Silverman et al., 2020].

As a result, the transformation problem in high-dimensional compositional data analysis is not just a matter of preprocessing. It determines the geometry in which models are trained, regularized, interpreted, and evaluated. The alr transformation is interpretable when the reference part is meaningful, but it depends on the choice of reference. The clr transformation treats all parts symmetrically, but its covariance matrix is singular. The ilr transformation preserves Aitchison geometry in ordinary Euclidean coordinates, but its

coordinates depend on the chosen basis and can be hard to interpret. Hellinger and power transformations handle zeros more naturally, but they relax many traditional assumptions of compositional data analysis. Zero replacement makes log-ratio analyses possible, but it can introduce substantial distortion when the data are sparse. These tradeoffs explain why transformation choice remains a central methodological issue in high-dimensional compositional data analysis [Pawlowsky-Glahn et al., 2015, Quinn et al., 2018].

3.2 Model Evaluation

Model evaluation remains less settled than model construction in compositional data analysis. A large literature has developed regression models, covariance and network estimators, differential-abundance methods, latent-factor, and generative models for simplex-valued or count-derived compositional data [Chen and Li, 2013, Li, 2015, Tang and Chen, 2019]. There is much less agreement on how these models should be compared after fitting, especially when the number of parts is large, zeros are common, and scientific interest includes both dominant and rare components. Evaluation is not just a final implementation step. Instead, the choice of model validation determines what kind of predictive accuracy is being measured and can substantially change the ranking of competing models [Gelman et al., 2014].

The central difficulty is that there is no single loss function that is appropriate for every compositional problem. A composition can be treated as a point in Aitchison geometry, a normalized count vector generated by a sampling process, a probability vector for a multinomial distribution, or an abundance profile in which the presence or absence of rare parts has scientific meaning [Aitchison, 1986, Gloor et al., 2017]. Each view leads to a different evaluation target. This problem becomes sharper in high dimensions because many components lie near the boundary of the simplex. A small absolute error in a rare component can produce a large log-ratio error, while a larger absolute error in a dominant component may have only a modest log-ratio effect. For this reason, evaluation criteria must be explicit about what they prioritize, such as relative accuracy, absolute abundance, predictive probability, support recovery, or scientific interpretability.

3.2.1 Transformation-Based Evaluation

One major class of evaluation methods is transformation-based. The observed and predicted compositions are first mapped into a chosen coordinate system, and prediction error is then measured in that transformed space [Aitchison, 1986]. This approach is useful because it allows standard loss functions, such as squared error, to be applied to compositional data. However, the transformation is not neutral. It determines the geometry in which prediction error is measured and therefore determines what kind of model performance is rewarded.

The most principled version of this approach uses log-ratio coordinates that preserve Aitchison geometry [Egozcue et al., 2003]. If $\hat{\pi}_i$ is the predicted composition and π_i is the observed composition for observation i, an ilr squared-error loss is

$$L_{\text{ilr}} = \sum_{i=1}^{N} || \operatorname{ilr}(\pi_i) - \operatorname{ilr}(\hat{\pi})_i)||_2^2$$

Because the ilr transformation is isometric, this loss is equal to the squared Aitchison

distance,

$$L_{\text{ilr}} = \sum_{i=1}^{N} d_A^2(\pi_i, \hat{\pi}_i)$$

so ilr-based squared error evaluates prediction accuracy in the intrinsic geometry of the simplex rather than the geometry of raw proportions [Egozcue et al., 2003]. The same aggregate loss can also be computed using clr coordinates. For positive compositions,

$$||\operatorname{clr}(\pi_i) - \operatorname{clr}(\hat{\pi})_i)||_2^2 = ||\operatorname{ilr}(\pi_i) - \operatorname{ilr}(\hat{\pi})_i)||_2^2 = d_A^2(\pi_i, \hat{\pi}_i)$$

when the ilr basis is orthonormal [Aitchison, 1986, Egozcue et al., 2003]. The difference is not the total distance but the representation. The clr transformation uses K coordinates that satisfy a zero-sum constraint, while the ilr transformation uses $K - 1$ unconstrained orthonormal coordinates. For aggregate distance-based evaluation, these two choices are equivalent. For coordinate-wise diagnostics, however, they can lead to different interpretations because ilr coordinates depend on the chosen basis and clr coordinates are tied together by the zero-sum constraint.

Other log-ratio losses evaluate different targets. An alr squared-error loss can be useful when one part has a clear role as a reference category, but it is not invariant to the choice of reference. Changing the reference part changes the transformed coordinates and can change the resulting evaluation. For this reason, alr-based evaluation is most defensible when the reference component has a clear scientific meaning, and it is less suitable as a generic metric in settings where no single part is naturally privileged.

The main difficulty for log-ratio evaluation is zero handling. If either π_i or $\hat{\pi}_i$ contains zeros, then alr, clr, and ilr losses are undefined unless a replacement rule or model-based smoothing step is introduced [Martín-Fernández et al., 2003]. In sparse high-dimensional data, the resulting loss can depend as much on the zero-handling rule as on the models itself [Silverman et al., 2020]. This is especially important when comparing models that produce strictly positive predictions with models that allow exact zeros. A sparse model may correctly predict structural absence, yet a log-ratio loss after zero replacement can still penalize those exact zeros heavily. Conversely, a model that assigns small positive mass to every component may score well under log-ratio loss while failing to recover the support of the composition.

This limitation motivates transformation-based losses that are not built from log-ratios. One common example is the Hellinger transformation, which maps a composition to the square-root scale as described in Section 3.1.5 [Legendre and Gallagher, 2001]. A corresponding squared Hellinger loss is

$$L_H = \sum_{i=1}^{N} d_H^2(\pi_i, \hat{\pi}_i) = \frac{1}{2} \sum_{i=1}^{N} \sum_{k=1}^{K} \left(\sqrt{\pi_{ik}} - \sqrt{\hat{\pi}_{ik}} \right)^2$$

This loss is attractive in sparse abundance data because it is well defined when components are zero and is less dominated by large components than raw Euclidean error. It evaluates similarity on the square-root abundance scale rather than in Aitchison geometry, so Hellinger-based evaluation may be useful for ecological or microbial prediction but should not be interpreted as a strict compositional accuracy in the log-ratio sense [Legendre and Gallagher, 2001, Quinn et al., 2018].

Transformation-based evaluation is therefore useful, but it must be interpreted through the transformation used to define the loss. Aitchison based losses such as clr and ilr measure relative accuracy among parts; alr losses measure accuracy relative to a chosen reference component; and Hellinger losses measure similarity on the square-root abundance scale and handle zeros more naturally but do not preserve strict log-ratio geometry. In high-dimensional compositional data analysis, the evaluation criterion should therefore be reported alongside the transformation and any zero-handling rule used to compute it.

3.2.2 Likelihood-Based Criteria

A second class of evaluation methods is likelihood-based. Rather than measuring a distance between an observed composition and a predicted composition, these criteria score the predictive distribution assigned to held-out data [Gelman et al., 2014, Vehtari et al., 2015]. Let D_{train} denote the training data and let $D_{\text{test}} = \{y_1^*, \dots, y_N^*\}$ denote the held-out observations. The log predictive density is

$$\text{LPD}(D_{\text{test}}) = \sum_{i=1}^{N} \log p(y_i^* \mid D_{\text{train}}) = \sum_{i=1}^{N} \log \int p(y_i^* \mid \theta) p(\theta \mid D_{\text{train}})\, d\theta$$

This criterion evaluates the full predictive distribution rather than only a point prediction, and as a strictly proper scoring rule it rewards models that assign high probability to held-out observations and penalizes models that are overconfident or poorly calibrated [Gneiting and Raftery, 2007].

In Bayesian models, leave-one-out cross-validation estimates expected out-of-sample predictive accuracy by scoring each observation under a model fitted without that observation [Vehtari et al., 2015]. The expected log predictive density can be written as

$$\widehat{\text{elpd}}_{\text{loo}} = \sum_{i=1}^{N} \log p(y_i \mid y_{-i}) = \sum_{i=1}^{N} \log \int p(y_i \mid \theta) p(\theta \mid y_{-i})\, d\theta$$

where y_{-i} denotes the data with observations i removed. When exact leave-one-out refitting is too expensive, Pareto-smoothed importance sampling leave-one-out-validation (PSIS-LOO) validation and the widely applicable information criterion (WAIC) are common approximations.

For posterior draws $\theta^{(1)}, \dots, \theta^{(S)}$, WAIC can be written as

$$\text{WAIC} = -2 \left[\sum_{i=1}^{N} \log \left\{ \frac{1}{S} \sum_{s=1}^{S} p(y_i \mid \theta^{(s)}) \right\} - \sum_{i=1}^{N} \text{Var}_{s=1}^{S} \left\{ \log p(y_i \mid \theta^{(s)} \right\} \right]$$

where the variance is taken across posterior draws [Gelman et al., 2014, Watanabe, 2010].

Likelihood-based evaluation has several advantages in high-dimensional compositional data analysis. It avoids choosing a log-ratio coordinate system, can be applied directly to counts, and can distinguish models with similar point predictions but different uncertainty quantification [Gelman et al., 2014]. It is also natural for generative models that represent the sampling process explicitly, including multinomial, Dirichlet-multinomial, and logistic-normal multinomial models [Aitchison and Shen, 1980, Chen and Li, 2013, Tang and Chen, 2019]. This is important when observed compositions arise from finite sequencing depth.

In that setting, the observed data are not the latent composition itself but a noisy count realization. Scoring the probability of held-out counts may therefore be more appropriate than scoring the distance between normalized proportions [Gloor et al., 2017, Silverman et al., 2020].

The main limitation is that likelihood scores evaluate the entire assumed data-generating distribution, not just the predicted mean composition. A model may estimate the average composition accurately but still receive a poor predictive score if it assigns too little probability to the observed variability, overdispersion, zero inflation, or tail behavior in the held-out data. Conversely, a highly flexible likelihood may assign high probability to the observed counts by capturing these sampling features, while the resulting latent composition is less stable. Likelihood-based evaluation therefore rewards distributional calibration, which is valuable, but it does not always coincide with the accuracy of the underlying composition.

3.2.3 Ecological and Distributional Dissimilarities

A third class of evaluation methods uses ecological, distributional, or domain-specific dissimilarities. These criteria do not necessarily treat a composition as a point in Aitchison geometry. Instead, they often treat it as an abundance profile, a probability distribution, or a structured object whose parts may have domain-specific relationships [Legendre and Gallagher, 2001]. This makes them useful when prediction quality is not fully captured by log-ratio accuracy. Their limitation is that they usually evaluate a different target from Aitchison distance and are not always coherent under the formal rules of compositional data.

One common way to evaluate an abundance profile is Bray-Curtis dissimilarity [Bray and Curtis, 1957]. For an observed composition π and its predicted vector $\hat{\pi}$, it is defined as

$$d_{\mathrm{BC}}(\pi, \hat{\pi}) = \frac{\sum_{k=1}^{K} |\pi_k - \hat{\pi}_k|}{\sum_{k=1}^{K} (\pi_k + \hat{\pi}_k)}$$

When both vectors are closed compositions, they sum to one, so the denominator equals two and this becomes

$$d_{\mathrm{BC}}(\pi, \hat{\pi}) = \frac{1}{2} \sum_{i=1}^{K} |\pi_i - \hat{\pi}_i|,$$

the usual form used for compositional predictions. It measures absolute discrepancy in component proportions rather than relative discrepancy in log-ratios. Because closed compositions sum to one, this simplified Bray-Curtis metric is mathematically equivalent to the total variation distance between two probability vectors. Two qualifications follow. First, on general abundance vectors Bray-Curtis is neither Euclidean nor metric [Legendre and Gallagher, 2001], which is why it cannot be obtained by transforming the data and computing Euclidean distances, and why ordination requires principal coordination analysis with a correction for negative eigenvalues. Restricted to closed compositions the simplified form above coincides with total variation distance and is therefore a metric, but this is a property of the closure, not of the coefficient. Second, and more consequentially for compositional work, the coefficient is not subcompositionally dominant in the sense of Aitchison et al. [2000]. Letting $S \subset \{1, \ldots, K\}$ index a subset of components with re-normalized subcompositions

$$\pi_S = \mathcal{C}((\pi_j)_{j \in S}), \quad \hat{\pi}_S = \mathcal{C}((\hat{\pi}_j)_{j \in S})$$

one may construct pairs for which $d_{BC}(\pi_S, \hat{\pi}_S) > d_{BC}(\pi, \hat{\pi})$, so that discarding parts increases the apparent discrepancy. Aitchison et al. [2000] exhibit exactly this failure for ordinary Euclidean distance on the simplex, which the closed Bray-Curtis coefficient is a rescaled ℓ_1 analogue. As a result, model rankings based on Bray-Curtis can change when rare or scientifically irrelevant categories are included or excluded. This is not necessarily a flaw in every application but it means that Bray-Curtis should be interpreted as an abundance-profile metric rather than a strict compositional loss.

The Jensen-Shannon divergence is another probability-based criterion. It compares two compositions through their average distribution [Lin, 1991]. Let $m = \frac{1}{2}(\pi + \hat{\pi})$. Then

$$d_{JS}(\pi, \hat{\pi}) = \frac{1}{2} \sum_{k=1}^{K} \pi_k \log\left(\frac{\pi_k}{m_k}\right) + \frac{1}{2} \sum_{k=1}^{K} \hat{\pi}_k \log\left(\frac{\hat{\pi}_k}{m_k}\right),$$

with the convention that $0 \log(0) = 0$. This divergence is symmetric and finite in the presence of zeros, which makes it more stable near the boundary of the simplex than the Kullback-Leibler divergence [Lin, 1991]. It is useful when the goal is to compare predicted and observed probability distributions, but it is not an Aitchison-geometric loss.

Some applications require distances that account for relationships among the parts. Wasserstein distances, also called earth-mover distances, are appropriate when the components have a meaningful ground metric. This occurs when parts represent ordered categories, phylogenetic positions, or other structured features [Rubner et al., 2000]. Unlike Bray-Curtis or Jensen-Shannon divergence, a Wasserstein distance does not treat all discrepancies equally. Moving mass between nearby components is penalized less than moving mass between distant components. In microbiome analysis, the UniFrac family of dissimilarities follows this logic by measuring differences along a phylogenetic tree [Lozupone and Knight, 2005]. Such distances can be scientifically meaningful when the structure among parts is central to the problem.

These dissimilarities are valuable because they often align closely with applied scientific questions. Bray-Curtis emphasizes absolute abundance error and has a clear ecological interpretation [Bray and Curtis, 1957]. Jensen-Shannon divergence compares compositions as probability distributions while remaining finite at zeros [Lin, 1991]; and Wasserstein and UniFrac distances use additional structure among the parts [Lozupone and Knight, 2005, Rubner et al., 2000]. None of these criteria is universally superior; they answer different evaluative questions from Aitchison loss. For this reason, ecological and distributional dissimilarities should be treated as domain-specific evaluation criteria, and their use should be justified by the scientific target of the analysis.

3.3 Comp-on-Comp/Compositional Covariates

Regression with compositional data splits into three related problems. In the first, the composition appears on the right-hand side, where a non-compositional response is regressed upon by covariates lying in Δ^{K-1}. In the second, the composition appears on the left, where a simplex-valued response is regressed on non-compositional covariates. In the third, which is referred to as compositional-on-compositional regression, both sides are constrained to the simplex. The literature is unevenly distributed across the three. The first is comparatively mature, driven largely by microbiome applications [Lin et al., 2014,

Lu et al., 2019, Shi et al., 2016]. The second is well developed in low dimensions through Dirichlet regresison [Hijazi and Jernigan, 2009] and through regression models built on the logistic-normal class introduced by Aitchison and Shen [1980]. The third remains sparsely populated [Egozcue et al., 2012, Fiksel et al., 2022, Tsagris, 2025], and almost empty in the high dimensional setting.

These three settings differ because a constraint on the response and a constraint on the design play different statistical roles. A constraint on the response restricts the support of the conditional distribution and is, as a result, a question of likelihood specification. This means the modeler must either choose a family supported on the simplex or transform the response into a space where another family applies. A constraint on the design restricts the support of nothing but instead leaves the regression coefficients unidentified. Because closure forces an exact linear dependence among the columns of the transformed design, the coefficient vector is identified only up to that dependence. The usual reading of a coefficient as the effect of increasing part j while holding all other parts fixed then becomes irrelevant, since no part can move while the others are held fixed. Methods for compositional covariates are best understood not as competing estimators of a common target but instead as different answers to questions of which functions of the coefficient vector are identified and scientifically interpretable.

3.3.1 The Linear Log-Contrast Model

The linear log-contrast model of Aitchison and Bacon-Shone [1984] is the standard resolution of the identifiability problem in the compositional-covariate setting. It has a second and independent justification: Aitchison and Bacon-Shone [1984] show that the linear log-contrast form is not merely sufficient but necessary for complete additivity meaning additivity of the expected response with respect to every partition of the parts and every permutation of them. Among response surfaces on the simplex it is therefore the unique completely additive one, which is the mixture-experiment analogue of the characterization of the Dirichlet class by complete neutrality. For observations $i = 1, \ldots, n$ with covariate composition $x_i \in \Delta^{K-1}$ and unconstrained response $y_i \in \mathbb{R}$, the model is

$$y_i = \beta_0 + \sum_{j=1}^{K} \beta_j \log(x_{ij}) + \varepsilon_i, \quad \sum_{j=1}^{K} \beta_j = 0$$

Here, the coefficients are required to sum to zero. This appears to be a technical restriction, but it is what makes the model estimable in the first place.

Suppose the quantity of real scientific interest is a vector of absolute abundances $w_i = (w_{i1}, \ldots, w_{iK})$, such as the actual number of each bacterial species present in a sample. What the data record is only the composition $x_i = \mathcal{C}(w_i)$, which keeps the proportions and discards the total $T_i = \sum_j w_{ij}$. Since $w_{ij} = T_i x_{ij}$, a model written on the abundances splits into two pieces,

$$\underbrace{\sum_{j=1}^{K} \beta_j \log(w_{ij})}_{\text{what we want}} = \underbrace{\sum_{j=1}^{K} \beta_j \log(x_{ij})}_{\text{what we can compute}} + \underbrace{\log(T_i) \sum_{j=1}^{K} \beta_j}_{\text{what we cannot}}$$

The first piece uses only the observed proportions. The second involves the total, which is often not known. Forcing the coefficients to sum to zero eliminates this second term entirely,

making it the exact condition under which a model about absolute abundances can be fit to proportions without ever needing to know the totals.

The constraint also settles what an individual coefficient means. Written on abundances, the log contrast $\sum_j \beta_j \log(w_{ij})$ is unchanged when every w_{ij} is multiplied by a common factor, and $\sum_j \beta_j = 0$ is exactly the condition for that invariance. One consequence is that the linear predictor takes no notice of the renormalization back to Δ^{K-1}, so the parts may be perturbed in whatever way is convenient and closure is applied afterward without changing anything. This is what makes the coefficients readable.

Multiply part j by a factor $t > 0$, leave the other parts as they are, and close the results back to the simplex. The mean response changes by $\beta_j \log(t)$, so β_j measures the effect of inflating part j against the remainder of the composition taken as a single block. Trade two parts against one another instead, multiplying part j by t and dividing part k by t with the remaining parts untouched, and the mean response changes by $(\beta_j - \beta_k)\log(t)$, so the contrast $\beta_j - \beta_k$ is the effect of moving material from k to j. What has no counterpart here is the usual reading of the regression coefficient as the effect of raising one covariate while the others are held fixed. That experiment does not exist on the simplex, and every statement about β is a statement about some reallocation.

In the coordinates of Section 3.1, the model is one we have already seen, and seen twice at that. Since $\operatorname{clr}(x_i) = \log(x_i) - \overline{\log(x_i)}\mathbf{1}$, multiplying by $\beta^\top$ makes the centering contribute $-\overline{\log(x_i)}\sum_j \beta_j = 0$, so that

$$\sum_{j=1}^{K} \beta_j \log(x_{ij}) = \beta^\top \operatorname{clr}(x_i)$$

The centering that defines clr therefore has no effect on a zero-sum coefficient vector. Substituting $\beta_K = -\sum_{j<k} \beta_j$ into the same expression instead gathers the terms into log-ratios against part K,

$$\sum_{j=1}^{K} \beta_j \log(x_{ij}) = \sum_{j=1}^{K-1} \beta_j \log\left(\frac{x_{ij}}{x_{iK}}\right)$$

and the constrained model is an alr regression as well.

That both readings hold at once is not a coincidence. It is what the constraint is for. The two coordinate systems fail in complementary ways as regression bases, and the zero-sum constraint is the point at which the failures cancel. Alr coordinates are identified but reference-dependent, carrying $K - 1$ free coefficients with no redundancy among them but attached to a part K that the analyst had to choose. The log-contrast form shows the choice to be immaterial, since the fitted linear predictor is determined by β alone and any other reference yields the same fit. Clr coordinates are reference-free but over-parametrized. Because $\mathbf{1}^\top \operatorname{clr}(x_i) = 0$, replacing β by $\beta + c\mathbf{1}$ leaves $\beta^\top \operatorname{clr}(x_i)$ unchanged, so a whole line of coefficient vectors produce exactly the same fit and no amount of data can single out one point on it. This is the singularity of the clr covariance appearing again on the coefficient side. Requiring $\sum_j \beta_j = 0$ picks out one point on that line, and it is the point at which the coefficients regain the abundance reading of the previous paragraphs. The log-contrast model inherits the symmetry of clr and the identifiability of alr, at the price of a linear contrast that every method below must respect. Each of those methods begins from this model and differs only in what it adds to it.

3.3.2 Regularization and Selection under Linear Constraints

When K is comparable to or larger than n, the log-contrast model must be regularized. Lin et al. [2014] proposed the constrained ℓ_1 estimator

$$\hat{\beta} = \underset{\mathbf{1}^\top \beta = 0}{\arg\min} \left\{ \frac{1}{2n} \|y - Z\beta\|_2^2 + \lambda \|\beta\|_1 \right\}, \quad Z_{ij} = \log(x_{ij})$$

solved by a coordinate descent method of multipliers, and established estimation and selection consistency under conditions analogous to those for the ordinary lasso. This estimator is the reference point for most subsequent work, and the extensions that followed can be read as relaxations of its three fixed components, namely the single constraint, the squared error loss, and the Gaussian response. Shi et al. [2016] replaced the single constraint by one sum-to-zero constraint within each subcomposition, which accommodates covariates measured at several taxonomic resolutions and makes the fitted model subcompositionally coherent by construction, and they developed a debiased procedure for inference. Lu et al. [2019] carried the construction to generalized linear models, minimizing a penalized negative log-likelihood subject to the same group of linear constraints by a generalized accelerated proximal gradient method, and accompanied it with a debiased procedure yielding asymptotically normal estimates and valid confidence intervals for binary and count responses. Combettes and Müller [2021] unified these proposals in a single convex program admitting a general linear constraint matrix and a free choice of data-fitting term, including Huber and other robust losses. Their formulation belongs to the class of perspective M-estimation models, which permits the noise scale to be estimated jointly with the regression vector while preserving the convexity of the whole problem, and thereby yields a theoretically derived regularization parameter in place of cross-validation. They solve the resulting constrained non-smooth program exactly by a Douglas-Rachford algorithm with convergence guarantees on the iterates. What the whole line shares is that constitutionality enters only as a linear equality on the parameter, so the problem remains convex and the computational price of respecting the simplex is small.

The interaction between sparsity and the constraint deserves more attention than it usually receives. Suppose $\hat{\beta}$ has support S. The coefficients outside S vanish, so the constraint reduces to $\sum_{j \in S} \hat{\beta}_j = 0$ and the fitted linear predictor is a zero-sum combination of $\log(x_{ij})$ over $j \in S$. By the invariance established above, that combination is unaffected if the parts in S are reclosed to sum to one, so the fitted model is a log-contrast on the subcomposition $\mathcal{C}(x_{iS})$ and on nothing else. This is a genuine strength, making the selected model automatically coherent, and its coefficients keep the reallocation reading of the previous subsection once attention is restricted to S. The corresponding weakness is that selection is not invariant to the resolution at which the parts are defined. Amalgamating two rare parts replaces the columns $\log(x_{ij})$ and $\log(x_{ik})$ by $\log(x_{ij} + x_{ik})$, which no linear operation can recover from the two columns it displaces, and splitting a coarse part does the same in reverse. Either operation moves the design and the constraint set together, and there is no general guarantee that the selected support is stable under such a re-partitioning. Since the resolution in high-dimensional applications is usually a processing choice rather than a scientific fact, the concern is not a negligible one.

Uncertainty quantification is comparatively underdeveloped. Srinivasan et al. [2021] constructed a two-step compositional knockoff filter: a screening step that reduces dimension while retaining a sum-to-zero constraint and attains a sure screening property, followed

by a fixed-X knockoff filter on the reduced model. The knockoff and threshold pair to deliver finite-sample false discovery rate control conditional on the screening event; the unmodified threshold controls only a modified false discovery rate. Cao et al. [2018] derived two-sample tests for high-dimensional compositional means by way of the clr transformation, and Liu et al. [2022] developed debiased tests for multivariate log-contrast models with sub-compositional predictors under a scaled composite nuclear norm penalty. The Bayesian treatments differ from the preceding in a way worth marking: the constraint ceases to be an exact equality. Scott et al. [2023] fit a hierarchical linear log-contrast model by structured mean-field Monte Carlo coordinate ascent variational inference, with spike-and-slab priors constructed to respect the constrained parameter space and the differing orders of magnitude across taxon abundances, and place the alr transformation inside the log-contrast model precisely so that no reference category need be chosen. Zhang et al. [2024] develop Bayesian compositional generalized linear mixed models in which a structured regularized horseshoe prior selects moderate phylogenetically related effects while a random effect absorbs the accumulated contribution of many small ones, and in which the sum-to-zero restriction is imposed softly, through the prior, rather than as a hard constraint on the parameter space. Whether a soft constraint preserves the abundance interpretation of Section **??** is not addressed in that literature and appears to be open. These are encouraging advances, but the available theory is asymptotic in a regime in which K grows while the design remains well conditioned, and that is the assumption which fails first in practice. Sparse compositions produce many zero and near-zero proportions, and once a pseudocount is imputed the corresponding columns of Z are nearly constant and nearly identical to one another, which is exactly the configuration that restricted eigenvalue conditions are there to exclude.

3.4 Sparsity-Inducing Priors

Potential Lead: Jyotishka Datta, and ...

In the high-dimensional regime where the number of categories K is large, compositional data typically exhibit what can be called *quasi-sparsity*, loosely defined as a regime where most components π_i are negligibly small while a few dominant categories account for the bulk of the composition. This pattern arises naturally in count data, e.g., microbiome and ecological data (species abundance), topic models (most topics carry negligible weight for a given document), and overfitted mixture models (most components are empty or near-empty). A prior for the simplex should reflect this structure by concentrating mass near the boundary of Δ^{K-1}, shrinking small components toward zero, and remaining robust enough to let genuinely large components stand out.

Strategies for Modeling Sparse Compositions

Several approaches have been proposed to represent sparse probability vectors. Although these methods are often classified according to their mathematical construction, it is more useful from the perspective of high-dimensional compositional data analysis to classify them according to how they attempt to induce sparsity. Broadly speaking, existing approaches fall into four categories: (i) transforming the composition into Euclidean coordinates and applying sparse methods there, (ii) concentrating prior mass near the boundary of the simplex, (iii) introducing latent selection or mixture mechanisms, and (iv) employing global-local shrinkage directly on the simplex.

1. Transform and Shrink One approach is to map the simplex into a Euclidean space and exploit the extensive machinery developed for sparse estimation in unconstrained settings. The most widely used reparametrization approach is the transformed normal family, which maps Δ^{K-1} to $\mathbb{R}^{K-1}$ via a log-ratio transform before placing a Gaussian prior on the result. The *additive log-ratio* (ALR) transform uses a reference category π_K and maps $\pi \mapsto (\log(\pi_1/\pi_K), \ldots, \log(\pi_{K-1}/\pi_K))$; the *centered log-ratio* (CLR) divides by the geometric mean; and the *isometric log-ratio* (ILR) [Egozcue et al., 2003] provides an isometric embedding into $\mathbb{R}^{K-1}$ with respect to the Aitchison inner product. The logistic normal distribution [Aitchison and Shen, 1980] results from placing a multivariate normal prior on the ALR-transformed coordinates, allows for a flexible covariance structure and, in principle, positive correlations. However, this generality comes at a cost: the ILR basis is not unique and the resulting inferences are sensitive to the choice of reference [Greenacre and Grunsky, 2019], and interpretability in terms of the original category probabilities is lost.

A second, more recent, reparametrization approach maps the simplex to the unit hypersphere via the square-root transform $\mathbf{y} = \sqrt{\boldsymbol{\pi}}$, so that $\sum_i y_i^2 = 1$. The elliptically symmetric angular Gaussian (ESAG) distribution of Paine et al. [2020] and its extension ESAG$^+$ [Schwob et al., 2024], which allows for zero components and positive correlations, exploit this geometry.

The advantage of these approaches is that they inherit the rich theory and computational tools available for sparse Euclidean models. However, sparsity in the transformed coordinates does not necessarily translate into sparsity of the original composition, and the resulting inferences may depend on the choice of basis or reference component.

2. Concentration near the boundary A second strategy attempts to induce sparsity by concentrating prior mass near the vertices and lower-dimensional faces of the simplex. Classical examples include the Dirichlet distribution and its numerous extensions. This, second broad class of priors used in compositional data analysis, are priors native to the simplex, where we model $\boldsymbol{\pi}$ directly, using one of two constructive definitions of the Dirichlet, as described below.

2a. Normalized-basis constructions obtain $\pi_i = W_i / \sum_j W_j$ where W_i are independent positive random variables. The standard Dirichlet corresponds to Gamma$(\alpha_i, 1)$ basis variables. Favaro et al. [2011] generalize this by allowing infinitely divisible laws for the W_i. The flexible Dirichlet (FD) of Ongaro and Migliorati [2013] and its extension [Ongaro et al., 2020] use a mixture-of-Dirichlet construction to allow for multimodality and more flexible marginal and dependence structure. The generalized Liouville distributions [Rayens and Srinivasan, 1994] extend Dirichlet by allowing the basis variables to have different scale parameters, relaxing the independence structure while staying on the simplex. While these families offer genuine improvements in flexibility and dependence structure, none is specifically designed around the goal of inducing quasi-sparsity in high dimensions.

2b. Stick-breaking constructions Write $\pi_i = Z_i \prod_{\ell<i}(1 - Z_\ell)$ where $Z_i \sim \mathcal{B}eta(\alpha_i, \beta_i)$ independently, with $Z_K = 1$. The standard Dirichlet corresponds to $Z_i \sim \mathcal{B}eta(\alpha_i, \sum_{\ell>i} \alpha_\ell)$ for appropriate choices. The generalized Dirichlet (GD) of Connor and Mosimann [1969] allows fully free (α_i, β_i) pairs, producing a more flexible covariance structure and conjugacy to the multinomial [Wong, 1998], but is not designed for sparsity.

Although these families provide greater flexibility than the standard Dirichlet, they were primarily developed to enrich dependence structures rather than to mimic the adaptive shrinkage behavior familiar from sparse Euclidean models.

3. Mixture and latent-selection approaches Another strategy introduces latent indicators to distinguish active and inactive components.

The scaled Dirichlet [Aitchison, 1985] provides yet another stick-breaking variant. Heiner et al. [2019] introduce two sparsity-focused constructions: a sparse Dirichlet mixture (SDM), which uses a Dirichlet-mixture prior on $\boldsymbol{\pi}$, and a stick-breaking mixture (SBM) prior, which places a mixture of three Beta densities on each Z_k. While these are among the first Bayesian priors explicitly targeting simplex sparsity, the SBM prior suffers from a truncation-error issue when the number of components is large. The zero-inflated multinomial Dirichlet regression of Tang and Chen [2019] places a zero-augmented Beta on each stick-breaking weight, but the zero-inflation approach introduces mixing and identifiability difficulties. The spike-and-slab stick-breaking model of Zeng et al. [2023] targets consistent estimation of the number of clusters in a mixture model rather than the quasi-sparsity of the probability vector per se.

These methods are conceptually appealing because they explicitly separate large and small components. However, they often introduce difficult posterior geometries, label-switching phenomena, truncation errors, and computational challenges that become increasingly severe as K grows.

4. Global-local shrinkage on the simplex Inspired by the success of global-local priors in sparse Euclidean problems [Carvalho et al., 2009; Polson and Scott, 2010], recent work has attempted to separate the overall level of concentration from local component-specific shrinkage.

The Pochhammer/half-horseshoe prior A recent and conceptually distinct approach is the Pochhammer prior of Wang and Polson [2024], which places a heavy-tailed, mass-near-zero prior on the concentration parameter α of a symmetric Dirichlet$(\alpha, \ldots, \alpha)$. In the same spirit as global-local shrinkage priors for real-valued parameters [Carvalho et al., 2009], the recommended default (the "half-horseshoe" (HHS) prior on α) uses heavy tails to allow some categories to escape shrinkage while pulling most toward zero. This construction maintains conjugacy with the multinomial likelihood and allows tractable posterior computation. However, it inherits a fundamental limitation of the symmetric Dirichlet: because a single scalar parameter α governs both the global sparsity level and the precision of each component, the prior cannot decouple sparse structure from concentration, a critical limitation in settings where the dominant categories require substantially different treatment from the rare ones.

More generally, extending the global-local paradigm to probability vectors remains largely unexplored.

Table 1 provides a systematic catalog of the main prior families for compositional data, organized by their construction. We briefly describe the major categories.

Category	Prior	Source
Transform (CLR)	Logistic Normal	Aitchison and Shen [1980]
Transform (ILR)	ILR-Normal	Egozcue et al. [2003]
Transform (sqrt)	ESAG	Paine et al. [2020]
Transform (sqrt)	ESAG$^+$	Schwob et al. [2024]
Simplex (stick-breaking)	Generalized Dirichlet	Connor and Mosimann [1969]
Simplex (stick-breaking)	Scaled Dirichlet	Aitchison [1985]
Simplex (stick-breaking)	Stick-Breaking Mixture	Heiner et al. [2019]
Simplex (stick-breaking)	Zero-Inflated MND Regression	Tang and Chen [2019]
Simplex (stick-breaking)	Spike-and-Slab Stick-Breaking	Zeng et al. [2023]
Simplex (stick-breaking)	Pochhammer Prior	Wang and Polson [2024]
Simplex (normalized basis)	Generalized Liouville	Rayens and Srinivasan [1994]
Simplex (normalized basis)	Normalized Infinitely Divisible	Favaro et al. [2011]
Simplex (normalized basis)	Flexible Dirichlet	Ongaro and Migliorati [2013]
Simplex (normalized basis)	Extended Flexible Dirichlet	Ongaro et al. [2020]
Simplex (normalized basis)	Sparse Dirichlet Mixture	Heiner et al. [2019]

Table 1: A catalog of prior distributions for compositional data, classified by construction method.

Remaining challenges

Several interconnected limitations of current methods emerge when K is large:

(i) **Inability to decouple sparsity from precision.** The symmetric Dirichlet and its Pochhammer extension rely on a single shape parameter α that simultaneously controls how close π_i is to zero and how concentrated the distribution is overall. In truly sparse settings, we want to shrink many components aggressively while allowing a few to escape to their true non-zero values, a behavior more naturally captured by the global-local shrinkage paradigm [Datta and Dunson, 2016, Polson and Scott, 2010].

(ii) **Strong negative correlations.** As noted by Aitchison [1985], the Dirichlet distribution has a completely negative correlation structure: $\mathrm{Corr}(\pi_i, \pi_j) < 0$ for all $i \neq j$. This rules out positive co-abundance patterns that are common in microbiome and ecological data, where groups of functionally related species rise and fall together.

(iii) **Computational difficulties.** Mixture-based constructions and spike-and-slab formulations often lead to challenging posterior landscapes and can become computationally demanding when K is large.

(iv) **Lack of theoretical guarantees.** For real-valued sparse estimation, optimal minimax rates and conditions for posterior contraction at the minimax rate are well established [Castillo et al., 2015, van der Pas et al., 2014, 2017], or the Bayes risk for the compound decision problem [Bogdan et al., 2011, Datta and Ghosh, 2013]. No analogous results exist for simplex-valued estimation, and it is not even clear what the minimax or Bayes oracle rates are for estimating a quasi-sparse probability vector.

3.5 Positive Correlations

Potential Lead: Andrew Pope

Why this problem is worth exploring: It is conceivable that some real-life compositional

data has sizable positive correlations between observations. For example, a pair of microbes may have a symbiotic relationship with each other; thus if one is flourishing in a given site, we'd expect the other to also be flourishing and vice versa.

Brief review of current approaches and why they fail to solve the problem:

- Typical distributions used, such as the Dirichlet distribution, don't allow for positive correlations at all.
- Transformed versions of correlated variables (such as Logistic Normal (Aitchison + Shen, 1980)) are harder to interpret due to not working directly on the simplex.
- Flexible Dirichlet (Ongaro + Migliorati, 2013) Flexible dependence; no idea about it; same for SBM Prior (Heiner et al., 2019) Truncation error issue
- Some methods (such as the Generalized Dirichlet) allow for positive correlations, but only as an artifact; the large positive correlations we'd hope to potentially model are not supported.

Potential applications that reveal the problem:

Future directions to potentially fix the problem:

3.6 Prior Sensitivity (Improved Priors in ILR Space)

Potential Lead: Dylan Steberg (due June 1st)

Also show Dir prior is not so great. For example, generate comp data from a different process and show how it is limiting. - Dir prior can be restricting.

posterior predictive checks? - note that it can be hard to do a sensitivity analysis

Why this problem is worth exploring: The big issue here is that compositional data lives on the simplex which can cause modeling challenges. It would be great if we could map the simplex to Euclidean space and employ our plethora of standard statistical tool. The isometric log-ratio transform does just that, but this does not come without issues. When we place an uninformative prior in ILR space, we are implicitly inducing a prior on the simplex which tends to be quite informative. The behavior from the induced prior can bias posterior inference by pushing mass to the edges or corners of the simplex. These priors are also basis dependent (maybe a different research problem/area?) so it is difficult to construct an "objective" prior when said prior depends on the coordinate system.

Brief review of current approaches and why they fail to solve the problem: Very common approach is to assume an "uninformative" normal prior which induces a logit-normal distribution on the simplex. This logit-normal distribution induces strong structure on the simplex. - Some work on careful construction of the covariance matrix of a normal as an attempt at invariance. - Some work on Jeffery's priors and reference priors but seems lacking and difficult as they are basis dependent.

Potential applications that reveal the problem:

Future directions to potentially fix the problem:

4 Conclusion

Lead: Michael

References

J. Aitchison. *The statistical analysis of compositional data*. Monographs on Statistics and Applied Probability. Chapman & Hall, London, 1986. ISBN 0-412-28060-4. doi: 10.1007/978-94-009-4109-0. URL https://doi.org/10.1007/978-94-009-4109-0.

J Aitchison and Sheng M Shen. Logistic-normal distributions: Some properties and uses. *Biometrika*, 67(2):261–272, 1980.

J. Aitchison, Carles Vidal, Josep Martín-Fernández, and Vera Pawlowsky-Glahn. Logratio analysis and compositional distance. *Mathematical Geology*, 32:271–275, 01 2000. doi: 10.1023/A:1007529726302.

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

John Aitchison. A general class of distributions on the simplex. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 47(1):136–146, 1985.

John Aitchison and John Bacon-Shone. Log contrast models for experiments with mixtures. *Biometrika*, 71(2):323–330, 1984. doi: 10.1093/biomet/71.2.323.

Anirban Bhattacharya, Debdeep Pati, Natesh S. Pillai, and David B. Dunson. Dirichlet-Laplace priors for optimal shrinkage. *Journal of the American Statistical Association*, 110:1479–1490, 2015.

David M Blei. Probabilistic topic models. *Communications of the ACM*, 55(4):77–84, 2012.

David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

Małgorzata Bogdan, Arijit Chakrabarti, Florian Frommlet, and Jayanta K Ghosh. Asymptotic Bayes-optimality under sparsity of some multiple testing procedures. *The Annals of Statistics*, 39(3):1551–1579, 2011.

J Roger Bray and John T Curtis. An ordination of the upland forest communities of southern wisconsin. *Ecological Monographs*, 27(4):325–349, 1957. doi: 10.2307/1942268.

Yuanpei Cao, Wei Lin, and Hongzhe Li. Two-sample tests of high-dimensional means for compositional data. *Biometrika*, 105(1):115–132, 2018. doi: 10.1093/biomet/asx060.

C. M. Carvalho, N. G. Polson, and J. G. Scott. Handling sparsity via the horseshoe. *Journal of Machine Learning Research W&CP*, 5:73–80, 2009.

Ismail Castillo, Johannes Schmidt-Hieber, and Aad van der Vaart. Bayesian linear regression with sparse priors. *The Annals of Statistics*, 43(5):1986–2018, October 2015. URL http://dx.doi.org/10.1214/15-AOS1334.

Sebastien F. M. Chastin, Javier Palarea-Albaladejo, Manon L. Dontje, and Dawn A. Skelton. Combined effects of time spent in physical activity, sedentary behaviors and sleep on

obesity and cardio-metabolic health markers: A novel compositional data analysis approach. *PLOS ONE*, 10(10):1–37, October 2015. doi: 10.1371/journal.pone.0139984. URL https://doi.org/10.1371/journal.pone.0139984.

Jun Chen and Hongzhe Li. Variable selection for sparse Dirichlet-multinomial regression with an application to microbiome data analysis. *The Annals of Applied Statistics*, 7(1): 418–442, 2013. doi: 10.1214/12-AOAS592.

Patrick L. Combettes and Christian L. Müller. Regression models for compositional data: General log-contrast formulations, proximal optimization, and microbiome data applications. *Statistics in Biosciences*, 13(2):217–242, 2021. doi: 10.1007/s12561-020-09283-2.

Robert J Connor and James E Mosimann. Concepts of independence for proportions with a generalization of the dirichlet distribution. *Journal of the American Statistical Association*, 64(325):194–206, 1969.

Jyotishka Datta and David B Dunson. Bayesian inference on quasi-sparse count data. *Biometrika*, 103(4):971–983, 2016.

Jyotishka Datta and Jayanta K Ghosh. Asymptotic properties of Bayes risk for the horseshoe prior. *Bayesian Analysis*, 8(1):111–132, 2013.

Jacob C. Douma and James T. Weedon. Analysing continuous proportions in ecology and evolution: A practical introduction to beta and dirichlet regression. *Methods in Ecology and Evolution*, 10(9):1412–1430, 2019. doi: https://doi.org/10.1111/2041-210X.13234. URL https://besjournals.onlinelibrary.wiley.com/doi/abs/10.1111/2041-210X.13234.

Dorothea Dumuid, Tyman E Stanford, Josep-Antoni Martin-Fernández, Željko Pedišić, Carol A Maher, Lucy K Lewis, Karel Hron, Peter T Katzmarzyk, Jean-Philippe Chaput, Mikael Fogelholm, et al. Compositional data analysis for physical activity, sedentary time and sleep research. *Statistical Methods in Medical Research*, 27(12):3726–3738, 2018.

J. J. Egozcue, J. Daunis-i Estadella, V. Pawlowsky-Glahn, K. Hron, and P. Filzmoser. Simplicial regression. The normal model. *Journal of Applied Probability and Statistics*, 6(1):87–108, 2012.

Juan J. Egozcue, Vera Pawlowsky-Glahn, Guillem Mateu-Figueras, and Carlos Barceló-Vidal. Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3):279–300, 2003. doi: 10.1023/A:1023818214614.

Juan Jose Egozcue and Vera Pawlowsky-Glahn. Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37:795–828, 10 2005. doi: 10.1007/s11004-005-7381-9.

Stefano Favaro, Georgia Hadjicharalambous, and Igor Prünster. On a class of distributions on the simplex. *Journal of Statistical Planning and Inference*, 141(9):2987–3004, 2011.

Adelaide Figueiredo and Fernanda Figueiredo. Discriminant analysis for a folded Watson distribution. *Journal of Applied Statistics*, 53(4):555–573, 2026.

Jacob Fiksel, Scott Zeger, and Abhirup Datta. A transformation-free linear regression for compositional outcomes and predictors. *Biometrics*, 78(3):974–987, 2022. doi: 10.1111/biom.13465.

Eva Fišerová and Karel Hron. On the interpretation of orthonormal coordinates for compositional data. *Mathematical Geosciences*, 43:455–468, 04 2011. doi: 10.1007/s11004-011-9333-x.

Andrew Gelman, Jessica Hwang, and Aki Vehtari. Understanding predictive information criteria for bayesian models. *Statistics and computing*, 24(6):997–1016, 2014.

Gregory B Gloor, Jean M Macklaim, Vera Pawlowsky-Glahn, and Juan J Egozcue. Microbiome datasets are compositional: And this is not optional. *Frontiers in Microbiology*, 8: 2224, 2017.

Tilmann Gneiting and Adrian Raftery. Strictly proper scoring rules, prediction, and estimation. *Journal of the American Statistical Association*, 102:359–378, 03 2007. doi: 10.1198/016214506000001437.

Michael Greenacre and Eric Grunsky. The isometric logratio transformation in compositional data analysis: a practical evaluation. 2019.

Matthew Heiner, Athanasios Kottas, and Stephan Munch. Structured priors for sparse probability vectors with application to model selection in Markov chains. *Statistics and Computing*, 29(5):1077–1093, Sep 2019. ISSN 1573-1375. doi: 10.1007/s11222-019-09856-2. URL https://doi.org/10.1007/s11222-019-09856-2.

Rafiq H. Hijazi and Robert W. Jernigan. Modelling compositional data using Dirichlet regression models. *Journal of Applied Probability and Statistics*, 4(1):77–91, 2009.

Pierre Legendre and Eugene Gallagher. Ecologically meaningful transformations for ordination of species data. *Oecologia*, 129:271–280, 10 2001. doi: 10.1007/s004420100716.

Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2:73–94, 2015.

Jianhua Lin. Divergence measures based on the shannon entropy. *IEEE Transactions on Information Theory*, 37(1):145–151, 1991. doi: 10.1109/18.61115.

Wei Lin, Pixu Shi, Rui Feng, and Hongzhe Li. Variable selection in regression with compositional covariates. *Biometrika*, 101(4):785–797, 2014. doi: 10.1093/biomet/asu031.

Xiaokang Liu, Xiaomei Cong, Gen Li, Kendra Maas, and Kun Chen. Multivariate log-contrast regression with sub-compositional predictors: Testing the association between preterm infants' gut microbiome and neurobehavioral outcomes. *Statistics in Medicine*, 41 (3):580–594, 2022. doi: 10.1002/sim.9273.

David Lovell, Vera Pawlowsky-Glahn, Juan José Egozcue, Samuel Marguerat, and Jürg Bähler. Proportionality: a valid alternative to correlation for relative data. *PLoS Computational Biology*, 11(3):e1004075, 2015. doi: 10.1371/journal.pcbi.1004075.

Catherine Lozupone and Rob Knight. Unifrac: a new phylogenetic method for comparing microbial communities. *Applied and Environmental Microbiology*, 71(12):8228–8235, 2005. doi: 10.1128/AEM.71.12.8228-8235.2005.

Jiarui Lu, Pixu Shi, and Hongzhe Li. Generalized linear models with linear constraints for microbiome compositional data. *Biometrics*, 75(1):235–244, 2019. doi: 10.1111/biom.12956.

Josep A Martín-Fernández, Carles Barceló-Vidal, and Vera Pawlowsky-Glahn. Dealing with

zeros and missing values in compositional data sets using nonparametric imputation. *Mathematical Geology*, 35(3):253–278, 2003.

Josep Martín-Fernández, Karel Hron, Matthias Templ, Peter Filzmoser, and Javier Palarea-Albaladejo. Bayesian-multiplicative treatment of count zeros in compositional data sets. *Statistical Modelling*, 15:134–158, 09 2014. doi: 10.1177/1471082X14535524.

Geoffrey J McLachlan, Sharon X Lee, and Suren I Rathnayake. Finite mixture models. *Annual Review of Statistics and Its Application*, 6(1):355–378, 2019.

A. Ongaro and S. Migliorati. A generalization of the Dirichlet distribution. *Journal of Multivariate Analysis*, 114:412–426, 2013. ISSN 0047-259X. doi: https://doi.org/10.1016/j.jmva.2012.07.007. URL https://www.sciencedirect.com/science/article/pii/S0047259X12001753.

Andrea Ongaro, Sonia Migliorati, and Roberto Ascari. A new mixture model on the simplex. *Statistics and Computing*, 30:749–770, 2020. URL https://doi.org/10.1007/s11222-019-09920-x.

Phillip J Paine, SP Preston, Michail Tsagris, and Andrew TA Wood. Spherical regression models with general covariates and anisotropic errors. *Statistics and Computing*, 30(1): 153–165, 2020.

Vera Pawlowsky-Glahn and Juan Jose Egozcue. Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15:384–398, 10 2001. doi: 10.1007/s004770100077.

Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana-Delgado. Modeling and analysis of compositional data. 2015.

Nicholas G Polson and James G Scott. Shrink globally, act locally: Sparse Bayesian regularization and prediction. *Bayesian Statistics*, 9:501–538, 2010.

Thomas P Quinn, Mark F Richardson, David Lovell, and Tamsyn M Crowley. propr: An r-package for identifying proportionally abundant features using compositional data analysis. *Scientific reports*, 7(1):16252, 2017.

Thomas P Quinn, Ionas Erb, Mark F Richardson, and Tamsyn M Crowley. Understanding sequencing data as compositions: an outlook and review. *Bioinformatics*, 34(16):2870–2878, 2018.

William S Rayens and Cidambi Srinivasan. Box–cox transformations in the analysis of compositional data. *Journal of Chemometrics*, 5(3):227–239, 1991.

William S Rayens and Cidambi Srinivasan. Dependence properties of generalized liouville distributions on the simplex. *Journal of the American Statistical Association*, 89(428):1465–1470, 1994.

Yossi Rubner, Carlo Tomassi, and Leonidas J Guibas. The earth mover's distance as a metric for image retrieval. *International Journal of Computer Vision*, 40(2):99–121, 2000. doi: 10.1023/A:1026543900054.

Janice L Scealy and Alan H Welsh. Regression for compositional data by using distributions defined on the hypersphere. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 73(3):351–375, 2011.

Michael R Schwob, Mevin B Hooten, Nicholas M Calzada, and Timothy H Keitt. Spatial hyperspheric models for compositional data. *arXiv preprint arXiv:2410.03648*, 2024.

Darren A. V. Scott, Ernest Benavente, Julian Libiseller-Egger, Dmitry Fedorov, Jody Phelan, Elena Ilina, Polina Tikhonova, Alexander Kudryavtsev, Julia Galeeva, Taane Clark, and Alex Lewin. Bayesian compositional regression with microbiome features via variational inference. *BMC Bioinformatics*, 24:210, 2023. doi: 10.1186/s12859-023-05219-x.

Pixu Shi, Anru Zhang, and Hongzhe Li. Regression analysis for microbiome compositional data. *The Annals of Applied Statistics*, 10(2):1019–1040, 2016. doi: 10.1214/16-AOAS928.

Justin D Silverman, Alex D Washburne, Sayan Mukherjee, and Lawrence A David. A phylogenetic transform enhances analysis of compositional microbiota data. *eLife*, 6: e21887, 2017. doi: 10.7554/eLife.21887.

Justin D Silverman, Kimberly Roche, Sayan Mukherjee, and Lawrence A David. Naught all zeros in sequence count data are the same. *Computational and Structural Biotechnology Journal*, 18:2789–2798, 2020. doi: 10.1016/j.csbj.2020.09.014.

Arun Srinivasan, Lingzhou Xue, and Xiang Zhan. Compositional knockoff filter for high-dimensional regression analysis of microbiome data. *Biometrics*, 77(3):984–995, 2021. doi: 10.1111/biom.13336.

Zheng-Zheng Tang and Guanhua Chen. Zero-inflated generalized dirichlet multinomial regression model for microbiome compositional data analysis. *Biostatistics*, 20(4):698–713, 2019.

Michail Tsagris. Constrained least squares simplicial-simplicial regression. *Statistics and Computing*, 35(2):27, 2025. doi: 10.1007/s11222-024-10560-z.

SL van der Pas, BJK Kleijn, and AW van der Vaart. The horseshoe estimator: Posterior concentration around nearly black vectors. *Electronic Journal of Statistics*, 8:2585–2618, 2014.

Stéphanie van der Pas, Botond Szabó, and Aad van der Vaart. Adaptive posterior contraction rates for the horseshoe. *arXiv:1702.03698*, 2017.

Aki Vehtari, Andrew Gelman, and Jonah Gabry. Practical bayesian model evaluation using leave-one-out cross-validation and waic. *Statistics and Computing*, 27:1413–1432, 2015. doi: 10.1007/s11222-016-9696-4.

Yuexi Wang and Nicholas G Polson. Pochhammer priors for sparse count models. *arXiv preprint arXiv:2402.09583*, 2024.

Sumio Watanabe. Asymptotic equivalence of bayes cross validation and widely applicable information criterion in singular learning theory. *Journal of Machine Learning Research*, 11 (116):3571–3594, 2010. URL http://jmlr.org/papers/v11/watanabe10a.html.

Tzu-Tsung Wong. Generalized dirichlet distribution in bayesian analysis. *Applied Mathematics and Computation*, 97(2-3):165–181, 1998.

Cheng Zeng, Jeffrey W Miller, and Leo L Duan. Consistent model-based clustering using the quasi-bernoulli stick-breaking process. *Journal of Machine Learning Research*, 24(153): 1–32, 2023. URL http://jmlr.org/papers/v24/22-0436.html.

Li Zhang, Xinyan Zhang, and Nengjun Yi. Bayesian compositional generalized linear models for analyzing microbiome data. *Statistics in Medicine*, 43(1):141–155, 2024. doi: 10.1002/sim.9946.