Interaction Models and Generalized Score Matching for Compositional Data

Shiqing Yu
University of Washington, Seattle
syu.phd@gmail.com

Mathias Drton
Technical University of Munich
mathias.drton@tum.de

Ali Shojaie
University of Washington, Seattle
ashojaie@uw.edu

Abstract

Applications such as the analysis of microbiome data have led to renewed interest in statistical methods for compositional data, i.e., data in the form of relative proportions. In particular, there is considerable interest in modelling interactions among such proportions. To this end we propose a class of exponential family models that accommodate arbitrary patterns of pairwise interaction. Special cases include Dirichlet distributions as well as Aitchison's additive logistic normal distributions. Generally, the distributions we consider have a density that features a difficult-to-compute normalizing constant. To circumvent this issue, we design effective estimation methods based on generalized versions of score matching.

1 Introduction

Compositional data is comprised of data points that are elements of the probability simplex. Nonnegative and adding up to one, such data points naturally represent proportions or the event probabilities in multinomial distributions. Formally, when observing compositions comprised of m relative proportions, the data points belong to the $(m-1)$-dimensional probability simplex

$$\Delta \equiv \Delta_{m-1} = \left\{ \boldsymbol{x} \in \mathbb{R}^m : \boldsymbol{x} \succeq \mathbf{0},\, \mathbf{1}_m^\top \boldsymbol{x} = 1 \right\}, \tag{1.1}$$

where $\boldsymbol{x} \succ 0$ (or $\boldsymbol{x} \succeq 0$) means $x_j > 0$ (or $x_j \geq 0$) for all j, and $\mathbf{1}_m = (1, \ldots, 1) \in \mathbb{R}^m$ is the vector of all ones. Compositional data arise, e.g., as rock composition in geology [1, 2], demographic data [3], concentrations in chemistry [4], and recently, relative abundances of microbiome compositions [5–9], where technological limitations hinder the absolute quantification of microbial abundances.

The seminal work by Aitchison [10] and more recent work on microbiome data (e.g. [11]) indicate that ignoring the compositional nature of the data leads to spurious correlations, which is especially problematic in graphical modeling [12]. The classical approach for analysis of such data is Aitchison's A^{m-1} model [10]. Its distributions can be parameterized as having a density proportional to

$$\exp\left(-0.5 \log \boldsymbol{x}^\top \mathbf{K} \log \boldsymbol{x} + \boldsymbol{\eta}^\top \log \boldsymbol{x}\right), \quad \boldsymbol{x} \in \Delta,$$

where the *interaction matrix* $\mathbf{K} \in \mathbb{R}^{m \times m}$, with $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, and $\boldsymbol{\eta} \in \mathbb{R}^m$ are parameters; cf. [13, Chap. 7]. This class includes Dirichlet distributions (when $\mathbf{K} = \mathbf{0}$) and the widely used additive logistic normal distribution [1] with $\mathbf{1}_m^\top \boldsymbol{\eta} = -m$, positing that $(\log(x_1/x_m), \ldots, \log(x_{m-1}/x_m)) \in \mathbb{R}^{m-1}$ follows a Gaussian graphical model when $\mathbf{K}$ is sparse. We note that for compositional data zero interactions do not correspond to conditional independences among the proportions themselves.

As a flexible extension of Aitchison's class, we propose a *power interaction model* with densities

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}) \propto \exp\left(-(2a)^{-1} \boldsymbol{x}^{a\top} \mathbf{K} \boldsymbol{x}^a + b^{-1} \boldsymbol{\eta}^\top \boldsymbol{x}^b\right), \quad \boldsymbol{x} \in \Delta, \tag{1.2}$$

with $a \geq 0, b \geq 0$. For $a = b = 0$, $\boldsymbol{x} \succ \mathbf{0}$ almost surely and we define $\boldsymbol{x}^{a\top} \mathbf{K} \boldsymbol{x}^a / a \equiv \log \boldsymbol{x}^\top \mathbf{K} \log \boldsymbol{x}$ and $\boldsymbol{x}^b / b \equiv \log \boldsymbol{x}$. This class, which we also term *a-b interaction models*, contains the A^{m-1} models via $a = b = 0$, but allows for many other possibilities such as Gaussian or square root models [14] truncated to the simplex. For the new models, we develop high-dimensionally consistent estimators

S. Yu et al., Interaction Models and Generalized Score Matching for Compositional Data. *Proceedings of the Second Learning on Graphs Conference (LoG 2023)*, PMLR 231, Virtual Event, November 27–30, 2023.

of $\mathbf{K}$ as well as the graph given by the support of $\mathbf{K}$ by suitably adapting a score matching method for full-dimensional domains [15]; independent work of [16] considers a related but different approach focused on low-dimensional problems.

The methods we derive from the models in (1.2) have the strong appeal that, with $a > 0$ and $b > 0$, they can directly handle proportions that are exactly zero. Many modern applications feature such "sparse data," and heuristics like adding small positive numbers are needed for the Aitchison's additive logistic normal model, which relies on logarithms.

Outline. Section 2 reviews generalized score matching for domains of positive measure [15] and proposes a modification of this methodology for simplex domains, concretely, the probability simplex Δ. Section 3 studies identifiability of a-b interaction models and gives conditions for their density kernels to be normalizable to proper densities. Section 4 customizes our estimators to a-b interaction models. High-dimensional consistency of our regularized estimators is established in Section 5, where we show that the previously found rates of convergence also hold for simplex domains. We empirically evaluate the performance of our estimators in Section 6, and illustrate their utility in an application to microbiome data in Section 7. Additional details and proofs are given in the Appendix.

Notation. Random quantities are in upper-case. Boldface font distinguishes vectors from scalars. Matrices are written in upright bold, with constant matrices in upper-case ($\mathbf{K}$) and random data matrices in lower-case ($\mathbf{x}$, $\mathbf{y}$). Super-/subscripts index rows/columns of a data matrix $\mathbf{x}$: $\boldsymbol{X}^{(i)}$ is the i-th row/sample, and $X_j^{(i)}$ is its j-th feature. For $\boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^m$, $\boldsymbol{u} \odot \boldsymbol{v} \equiv (u_1 v_1, \dots, u_m v_m)$ is the Hadamard product, and $\|\boldsymbol{u}\|_a = (\sum_{j=1}^m |u_j|^a)^{1/a}$ is the ℓ_a-norm, with $\|\boldsymbol{u}\|_\infty = \max_{j=1,\dots,m} |u_j|$. For $a \in \mathbb{R}$, write $\boldsymbol{v}^a \equiv (v_1^a, \dots, v_m^a)$. For a vector-valued function $\boldsymbol{f} : \mathbb{R}^m \to \mathbb{R}^m$, $\boldsymbol{x} \mapsto (f_1(\boldsymbol{x}), \dots, f_m(\boldsymbol{x}))$, we write $\boldsymbol{f}^a(\boldsymbol{x}) \equiv (f_1^a(\boldsymbol{x}), \dots, f_m^a(\boldsymbol{x}))$. We also write $\boldsymbol{f}'(\boldsymbol{x}) \equiv (\partial f_1(\boldsymbol{x})/\partial x_1, \dots, \partial f_m(\boldsymbol{x})/\partial x_m)$. Given a matrix $\mathbf{K} = [\kappa_{ij}]_{i,j} \in \mathbb{R}^{n \times m}$, the vectorization $\mathrm{vec}(\mathbf{K}) \in \mathbb{R}^{nm}$ is the stacking of matrix columns. The Frobenius norm of the matrix is $\|\mathbf{K}\|_F = \|\mathrm{vec}(\mathbf{K})\|_2$, its max norm is $\|\mathbf{K}\|_\infty \equiv \|\mathrm{vec}(\mathbf{K})\|_\infty \equiv \max_{i,j} |\kappa_{ij}|$, and its ℓ_a–ℓ_b operator norm is $|||\mathbf{K}|||_{a,b} \equiv \max_{\boldsymbol{x} \neq \mathbf{0}} \|\mathbf{K}\boldsymbol{x}\|_b / \|\boldsymbol{x}\|_a$, with $|||\mathbf{K}|||_a \equiv |||\mathbf{K}|||_{a,a}$. Given a vector $\boldsymbol{x}$, we write $(y; \boldsymbol{x}_{-j})$ to indicate the vector obtained by replacing the j-th coordinate x_j by y. A composition of two functions is denoted $f \circ g$.

2 Generalized Score Matching for General Domains

2.1 Generalized Score Matching and Domains of Positive Measure

Score matching [17, 18] is an effective method for estimation of Lebesgue densities that are defined only up to a finite normalizing constant. A generalized form of score matching [19, 20] allows for more efficient estimation for densities on $\mathbb{R}_+^m$. For a family of distributions $\mathcal{P}(\mathcal{D})$ with twice continuously differentiable densities on a domain $\mathcal{D} \subseteq \mathbb{R}^m$, the main idea behind generalized score matching is to estimate an unknown density p_0 by picking the distribution $P \in \mathcal{P}(\mathcal{D})$ whose density p minimizes a measure of distance between p and p_0 given by

$$\frac{1}{2} \int_{\mathcal{D}} p_0(\boldsymbol{x}) \Big\| \nabla \log p(\boldsymbol{x}) \odot \boldsymbol{h}^{1/2}(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \odot \boldsymbol{h}^{1/2}(\boldsymbol{x}) \Big\|_2^2 \mathrm{d}\boldsymbol{x}. \tag{2.1}$$

This modified Fisher divergence is half a weighted version of the $L_2(P_0)$ distance between the gradients of the log-densities. The weights are given by a function $\boldsymbol{h}(\boldsymbol{x}) = (h_1(x_1), \dots, h_m(x_m))$, which is a pre-selected almost everywhere (a.e.) positive function from $\mathcal{D}$ to $\mathbb{R}_+^m$. The weights give flexibility to efficiently cope with the effects of the boundary of $\mathcal{D}$. For consistent estimation, the divergence ought to be minimized if and only if $p_0 = p$ a.e. An estimator of p_0 is obtained by minimizing the loss resulting from a sample version of the divergence in (2.1). Importantly, this estimator does not depend on normalizing constants, which drop out in the gradient of the log-density. Moreover, for exponential families, the sample loss is quadratic in the canonical parameters.

The divergence in (2.1) involves the full gradient of the log-density and the desideratum of the divergence being minimal if and only if p is a.e. equal to the true density only makes sense for $\mathcal{D}$ with positive Lebesgue measure in $\mathbb{R}^m$. Such domains were treated in [15] (see also [21]) but do not include the case where $\mathcal{D}$ is the probability simplex Δ from (1.1). This paper further extends the generalized score to the simplex case and compositional data. To this end, we first briefly review the method of [15] for general domains of positive measure $\mathcal{D} \subseteq \mathbb{R}^m$. Our focus in this review is on

modeling and estimating the joint distribution P_0 of a random vector $\boldsymbol{X} \in \mathbb{R}^m$, when P_0 has support $\mathcal{D}$. We consider a family $\mathcal{P}(\mathcal{D})$ of distributions with twice continuously differentiable densities on $\mathcal{D}$.

For any $j = 1, \ldots, m$, let $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j}) \equiv \{y \in \mathbb{R} : (y; \boldsymbol{x}_{-j}) \in \mathcal{D}\}$ be the jth section of $\mathcal{D}$ defined by the $(m-1)$-dimensional vector $\boldsymbol{x}_{-j}$ and define the projection $\mathcal{S}_{-j,\mathcal{D}} \equiv \{\boldsymbol{x}_{-j} : \mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j}) \neq \varnothing\} \subseteq \mathbb{R}^{m-1}$. A measurable domain $\mathcal{D}$ is a *component-wise countable union of intervals* if for any j and fixed $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j,\mathcal{D}}$, the section $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j})$ is a union of finite or countably many intervals in $\mathbb{R}$. Let $\boldsymbol{C} \in \mathbb{R}^m$ be a truncation constant with $\boldsymbol{C} \succ \boldsymbol{0}_m$. Define $\boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}} = (\varphi_{1,C_1,\mathcal{D}}, \ldots, \varphi_{m,C_m,\mathcal{D}})$, where

$$\varphi_{j,C_j,\mathcal{D}}(\boldsymbol{x}) \equiv \min\Big\{ \inf_{(y;\,\boldsymbol{x}_{-j}) \in \mathcal{D}} |y - x_j|, C_j \Big\}. \tag{2.2}$$

By assuming a component-wise countable union of intervals, each x_j lies in a unique maximal subinterval of the section $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j})$ and the infimum in (2.2) gives the distance between x_j and the boundary of this interval. The minimum then truncates the distance from above by some $C_j > 0$, in order to maintain bounded weights in the divergence from (2.1). The second ingredient for the weights are transformations that adapt to the decay of densities at the boundary of $\mathcal{D}$. Given a user-specified $\boldsymbol{h} : \mathbb{R}_+^m \to \mathbb{R}_+^m$, $\boldsymbol{x} \mapsto (h_1(x_1), \ldots, h_m(x_m))^\top$ with $h_1, \ldots, h_m : \mathbb{R}_+ \to \mathbb{R}_+$ almost surely positive and absolutely continuous in every bounded sub-interval of $\mathbb{R}_+$, the *generalized* $(\boldsymbol{h}, \mathcal{C}, \mathcal{D})$*-score matching loss* in $P \in \mathcal{P}(\mathcal{D})$ with density p, denoted $L_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P)$, is defined as the divergence

$$\frac{1}{2}\int_{\mathcal{D}} p_0(\boldsymbol{x}) \Big\| \nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})^{1/2}(\boldsymbol{x}) - \nabla \log p_0(\boldsymbol{x}) \odot (\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C},\mathcal{D}})^{1/2}(\boldsymbol{x}) \Big\|_2^2 \mathrm{d}\boldsymbol{x}. \tag{2.3}$$

This population loss is minimized at p if and only if $p = p_0$ a.e. Under the (mild) assumptions (A.1)–(A.3) laid out in Appendix A, one can show that

$$\begin{aligned} L_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P) \equiv \frac{1}{2}\sum_{j=1}^m \int_{\mathcal{D}} p_0(\boldsymbol{x}) \cdot (h_j \circ \varphi_{C_j,\mathcal{D},j})(\boldsymbol{x}) \cdot [\partial_j \log p(\boldsymbol{x})]^2 \,\mathrm{d}\boldsymbol{x} \\ + \sum_{j=1}^m \int_{\mathcal{D}} p_0(\boldsymbol{x}) \cdot \partial_j \left[(h_j \circ \varphi_{C_j,\mathcal{D},j})(\boldsymbol{x}) \cdot \partial_j \log p(\boldsymbol{x}) \right] \mathrm{d}\boldsymbol{x} \end{aligned} \tag{2.4}$$

plus a constant depending on p_0 only (so, independent of p). Let $\boldsymbol{X}^{(i)}$, $1 \leq i \leq n$, be an i.i.d. sample from P_0. Then the associated empirical loss is

$$\begin{aligned} \hat{L}_{\boldsymbol{h},\boldsymbol{C},\mathcal{D}}(P) = \frac{1}{2}\sum_{j=1}^m \sum_{i=1}^n \frac{1}{2}(h_j \circ \varphi_{C_j,\mathcal{D},j})\left(\boldsymbol{X}^{(i)}\right) \cdot \left[\partial_j \log p\left(\boldsymbol{X}^{(i)}\right)\right]^2 + \\ \partial_j \left[(h_j \circ \varphi_{C_j,\mathcal{D},j})\left(\boldsymbol{X}^{(i)}\right) \cdot \partial_j \log p\left(\boldsymbol{X}^{(i)}\right) \right]. \end{aligned} \tag{2.5}$$

Note that the truncation by $\boldsymbol{C}$ is not necessary if each h_j is bounded from above e.g. by some $h_j(C_j)$.

2.2 Extension to Simplices

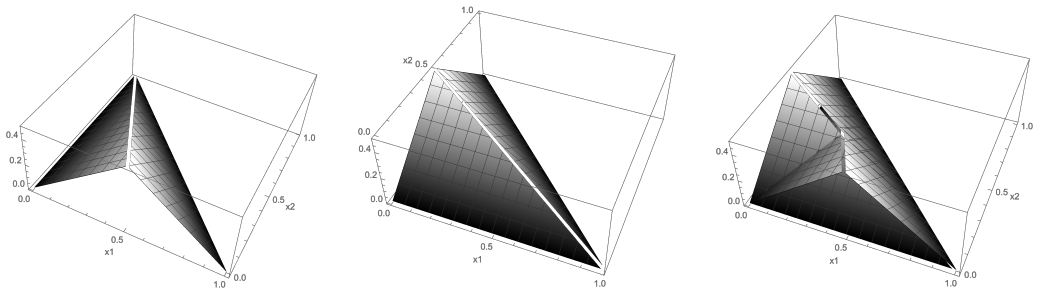
To generalize the approach just presented to domains that are null sets of dimension $k < m$, we propose to transform $\mathcal{D}$ to a full-dimensional subset of $\mathbb{R}^k$. This generalization is particularly tractable for the important case of the probability simplex Δ as the resulting inequality constraints can be efficiently handled in the method of [15]. Indeed, we may drop, say, the last coordinate x_m, substituting it with $1 - \boldsymbol{1}_{m-1}^\top \boldsymbol{x}_{-m}$, and work instead with the full-dimensional simplex

$$\Delta_{-m} \equiv \left\{ \boldsymbol{x}_{-m} \in \mathbb{R}^{m-1} : \boldsymbol{x}_{-m} \succeq \boldsymbol{0},\, \boldsymbol{1}_{m-1}^\top \boldsymbol{x}_{-m} \leq 1 \right\} \subseteq \mathbb{R}^{m-1}. \tag{2.6}$$

Henceforth, we thus consider the domain $\mathcal{D} \equiv \Delta_{-m}$, and remove the dependency of L and $\boldsymbol{\varphi}$ on $\mathcal{D}$.

For $\boldsymbol{x} \in \mathbb{R}^m$, and $j \in \{1, \ldots, m-1\}$, let $\boldsymbol{x}_{-\{j,m\}}$ be the vector in $\mathbb{R}^{m-2}$ obtained by removing x_m and x_j. Then Δ_{-m} has jth section $\mathcal{C}_j(\boldsymbol{x}_{-m}) = \left[0, 1 - \boldsymbol{1}_{m-2}^\top \boldsymbol{x}_{-\{j,m\}}\right]$. Hence, we have the coordinate-wise distance $\varphi_{C_j,j}(\boldsymbol{x}) = \min\left\{C_j, x_j, 1 - \boldsymbol{1}_{m-1}^\top \boldsymbol{x}_{-m}\right\} = \min\left\{C_j, x_j, x_m\right\}$. The role of the truncation constants C_j is to ensure boundedness of the coordinate-wise distances. As the simplex is naturally bounded by the unit cube, it is natural to not use any truncation here and simply use the following coordinate-wise distance, which is depicted in Figure 1,

$$\varphi_j(\boldsymbol{x}) = \min\left\{x_j, 1 - \boldsymbol{1}_{m-1}^\top \boldsymbol{x}_{-m}\right\} = \min\left\{x_j, x_m\right\}.$$



(a) φ_1 **(b)** φ_2 **(c)** $\varphi_1 \vee \varphi_2$

Figure 1: Plots of φ_1 and φ_2 for $\Delta_{-3} \equiv \{(x_1, x_2) : x_1 \geq 0, x_2 \geq 0, x_1 + x_2 \leq 1\}$.

Let $\mathcal{P} \equiv \{P_{\boldsymbol{\theta}} : \boldsymbol{\theta} \in \boldsymbol{\Theta} \subset \mathbb{R}^r\}$ be an r-dimensional exponential family with canonical parameter $\boldsymbol{\theta} \in \mathbb{R}^r$ (note that r may be different from the dimension of the random vector m) and distribution $P_{\boldsymbol{\theta}}$ with continuous support Δ and densities of the form $\log p_{\boldsymbol{\theta}}(\boldsymbol{x}) = \boldsymbol{\theta}^\top \boldsymbol{t}(\boldsymbol{x}) - \psi(\boldsymbol{\theta}) + b(\boldsymbol{x})$ for $\boldsymbol{x} \in \Delta$. The empirical loss $\hat{L}_{\boldsymbol{h},\boldsymbol{C}}$ from (2.5) can be written as a quadratic function in $\boldsymbol{\theta}$, i.e.,

$$\hat{L}_{\boldsymbol{h},\boldsymbol{C}}(P_{\boldsymbol{\theta}}) = \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \text{const.} \tag{2.7}$$

with $\boldsymbol{\Gamma}(\mathbf{x}) \in \mathbb{R}^{r\times r}$ (positive semi-definite) and $\boldsymbol{g}(\mathbf{x}) \in \mathbb{R}^r$ sample averages of known functions in $\mathbf{x}$ only. We detail the derivation of (2.7), $\boldsymbol{\Gamma}$, and $\boldsymbol{g}$ for our specific models of interest in Section 4.

In high-dimensional settings where the number of parameters r is large compared to the sample size n, we add an ℓ_1 regularization on $\boldsymbol{\theta}$ and consider the *regularized generalized score matching loss*

$$\hat{L}_{\boldsymbol{h},\boldsymbol{C},\lambda,\delta}(P_{\boldsymbol{\theta}}) \equiv \frac{1}{2}\boldsymbol{\theta}^\top \boldsymbol{\Gamma}_\delta(\mathbf{x})\boldsymbol{\theta} - \boldsymbol{g}(\mathbf{x})^\top \boldsymbol{\theta} + \lambda\|\boldsymbol{\theta}\|_1. \tag{2.8}$$

Here $\boldsymbol{\Gamma}_\delta(\mathbf{x})$ equals $\boldsymbol{\Gamma}(\mathbf{x})$ except that its diagonals are multiplied by $\delta > 1$. The *diagonal multiplier* δ is introduced to avoid possible unboundedness of the loss when $\boldsymbol{\Gamma}(\mathbf{x})$ is singular (due to high dimension) and the regularization parameter λ is small; cf. Section 4 of [20]. When estimating the interaction matrix $\mathbf{K}$, we typically only penalize its off-diagonal entries, i.e., the penalty is $\|\text{vec}(\mathbf{K}_{\text{off}})\|_1$.

The loss in (2.8) is no longer symmetric as it depends on the choice of the removed coordinate x_m. The asymmetry can be mitigated by calculating the loss $\hat{L}_j$ for each removed coordinate x_j and optimizing the average loss. In high dimensions, averaging over all j is costly, but we may nevertheless average over a set of (say 10 randomly chosen) coordinates $\mathcal{J} \subseteq \{1, \ldots, m\}$. Importantly, the averaged loss is still a quadratic form as in (2.8), just with $\boldsymbol{\Gamma}_\delta(\mathbf{x})$ and $\boldsymbol{g}(\mathbf{x})$ replaced by averages.

3 Power Interaction Models on the Probability Simplex

Reasoning as in Theorem 4.1 of [15], we obtain the following conditions on a and b for the density in (1.2) to be proper in the simplex case. Refined conditions for $a = b = 0$ are obtained in Section B.2.

Theorem 1 (Finite normalizing constant)**.** *If one of the following conditions holds, then the right-hand side of* (1.2) *is integrable over* Δ *and defines a proper density:*

(CC1) $a > 0$, $b > 0$*;*

(CC2) $a > 0$, $b = 0$, $\eta_j > -1$ *for all* j*;*

(CC3) $a = 0$, $b = 0$, $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) > 0\ \forall \boldsymbol{x} \in \Delta$*;*

(CC4) $a = 0$, $b > 0$, $\log(\boldsymbol{x})^\top \mathbf{K} \log(\boldsymbol{x}) \geq 0\ \forall \boldsymbol{x} \in \Delta$.

In the case where $\boldsymbol{\eta} = \mathbf{0}$ *is known, the conditions on* b *and* $\boldsymbol{\eta}$ *can be ignored.*

As a prerequisite for our subsequent discussion of estimation, the following theorem gives the conditions for the identifiability of $\mathbf{K}$ and $\boldsymbol{\eta}$ from a given a-b density on the simplex. In particular, the parameters are identifiable if $a \neq 1$ and we do not have $2a = b > 0$.

Theorem 2 (Identifiability). *Suppose there exist* $\mathbf{K}_1$, $\mathbf{K}_2$, $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ *such that*

$$\exp\left(-(2a)^{-1}\boldsymbol{x}^{a\top}\mathbf{K}_1\boldsymbol{x}^a + b^{-1}\boldsymbol{\eta}_1^\top\boldsymbol{x}^b\right) = \exp\left(-(2a)^{-1}\boldsymbol{x}^{a\top}\mathbf{K}_2\boldsymbol{x}^a + b^{-1}\boldsymbol{\eta}_2^\top\boldsymbol{x}^b\right)$$

for all $\boldsymbol{x} \in \Delta$, *where* $\boldsymbol{x}^0 \equiv \log(\boldsymbol{x})$ *and* $0^{-1} \equiv 1$. *Then* $\mathbf{K}_1 = \mathbf{K}_2$ *and* $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$, *or else one of the following must hold: (I)* $a = b = 1$, *(II)* $a = 1$, $b = 2$, *(III)* $a = 1$ *and* $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$, *(IV)* $2a = b > 0$ *and* $\mathbf{K}_1 - \mathbf{K}_2 = 2\boldsymbol{\eta}_1 - 2\boldsymbol{\eta}_2$.

The theorem uses that two densities are equal if and only if they are equal up to proportionality if and only if the log densities have equal gradients (up to null sets). We prove the theorem by taking log-gradients and match coefficients in the resulting expressions; the details are deferred to Section D.

Our approach to estimation by score matching is to profile out the last component of $\boldsymbol{x}$ using $x_m = 1 - \sum_{j=1}^{m-1} x_j$. The density in (1.2) becomes

$$p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}_{-m}) \propto \exp\left[-\frac{1}{2a}\boldsymbol{x}_{-m}^{a\top}\mathbf{K}_{-m,-m}\boldsymbol{x}_{-m}^a - \frac{1}{a}\boldsymbol{x}_{-m}^{a\top}\boldsymbol{\kappa}_{-m,m}\Big(1 - \sum_{j=1}^{m-1} x_j\Big)^a \right.$$
$$\left. -\frac{1}{2a}\kappa_{m,m}\Big(1 - \sum_{j=1}^{m-1} x_j\Big)^{2a} + \frac{1}{b}\boldsymbol{\eta}_{-m}^\top\boldsymbol{x}_{-m}^b + \frac{\eta_m}{b}\Big(1 - \sum_{j=1}^{m-1} x_j\Big)^b\right] \quad (3.1)$$

on $\Delta_{-m} \subseteq \mathbb{R}^{m-1}$. The next theorem gives sufficient conditions for the assumptions (A.1)–(A.3) in Appendix A to hold. Under these assumptions, the generalized score matching loss from (2.3) has the equivalent form in (2.4), and the empirical loss stated in (2.5) is valid.

Theorem 3 (Assumptions for score matching). *Suppose one of (CC1) through (CC4) holds, and* $\boldsymbol{h}(\boldsymbol{x}) = (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$, *where*

(I) if $a > 0$ *and* $b > 0$, $\alpha_j > \max\{0, 1-a, 1-b\}$;

(II) if $a > 0$ *and* $b = 0$, $\alpha_j > 1 - \eta_{0,j}$;

(III) if $a = 0$, $\alpha_j \geq 0$.

Then conditions (A.1)–(A.3) in Appendix A are satisfied. In the case with $\boldsymbol{\eta} \equiv \mathbf{0}$ *known, it suffices to have* $a > 0$ *and* $\alpha_j > \max\{0, 1-a\}$, *or* $a = 0$ *and* $\alpha_j \geq 0$.

The proof in Section D treats $a > 0$ vs. $a = 0$ and $b > 0$ vs. $b = 0$ separately. In each case, we bound relevant terms by sums of polynomials and inspect limits as a coordinate x_j approaches its boundary.

4 Estimation for Power Interaction Models

The equations in this section cover the case $a = 0$ (and $b = 0$) by the following convention. If $a = 0$, substitute exponents "a" with "1" and "$(a-1)$" with "-1". As before, $x^a \equiv \log x$ if $a = 0$.

For notational simplicity, we again drop the last coordinate x_m. Having substituted $x_m \equiv 1 - \mathbf{1}_{m-1}^\top\boldsymbol{x}_{-m}$ and working on Δ_{-m}, the partial derivative $\partial_j \log p(\boldsymbol{x}_{-m})$ of the density $p(\boldsymbol{x}_{-m}) \equiv p_{\boldsymbol{\eta},\mathbf{K}}(\boldsymbol{x}_{-m})$ in (3.1) now depends on both $(\boldsymbol{\kappa}_{,j}, \eta_j)$ and $(\boldsymbol{\kappa}_{,m}, \eta_m)$. Thus, unlike in the case of a-b models on domains with positive Lebesgue measure, the subvectors $(\boldsymbol{\kappa}_{,j}, \eta_j)$ and $(\boldsymbol{\kappa}_{,m}, \eta_m)$ are no longer isolated in the score-matching loss. Instead, we have

$$\partial_j \log p(\boldsymbol{x}_{-m}) = -\left(\boldsymbol{\kappa}_{,j}^\top\boldsymbol{x}^a\right)x_j^{a-1} + \left(\boldsymbol{\kappa}_{,m}^\top\boldsymbol{x}^a\right)x_m^{a-1} + \eta_j x_j^{b-1} - \eta_m x_m^{b-1}, \quad (4.1)$$

$$\partial_{jj} \log p(\boldsymbol{x}_{-m}) = -(a-1)\left[\left(\boldsymbol{\kappa}_{,j}^\top\boldsymbol{x}^a\right)x_j^{a-2} + \left(\boldsymbol{\kappa}_{,m}^\top\boldsymbol{x}^a\right)x_m^{a-2}\right] \quad (4.2)$$
$$- a\left[\kappa_{jj}x_j^{2a-2} + \kappa_{mm}x_m^{2a-2} + 2\kappa_{jm}x_j^{a-1}x_m^{a-1}\right] + (b-1)\left[\eta_j x_j^{b-2} + \eta_m x_m^{b-2}\right].$$

These derivatives yield the penalized loss (Eq. (2.8) with $\boldsymbol{\theta} = (\mathrm{vec}(\mathbf{K}), \boldsymbol{\eta})$), which we may write as

$$\frac{1}{2}\left(\mathrm{vec}(\mathbf{K}), \boldsymbol{\eta}\right)^\top\boldsymbol{\Gamma}\left(\mathrm{vec}(\mathbf{K}), \boldsymbol{\eta}\right) - \boldsymbol{g}^\top\left(\mathrm{vec}(\mathbf{K}), \boldsymbol{\eta}\right) + \lambda_{\mathbf{K}}\|\mathrm{vec}(\mathbf{K}_{\mathrm{off}})\|_1 + \lambda_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_1, \quad (4.3)$$

with matrix $\boldsymbol{\Gamma}$ and vector $\boldsymbol{g}$ naturally partitioned as

$$\boldsymbol{\Gamma} \equiv \begin{bmatrix}\boldsymbol{\Gamma}_{\mathbf{K}} & \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \\ \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^\top & \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\end{bmatrix} \in \mathbb{R}^{(m^2+m)\times(m^2+m)}, \quad \boldsymbol{g} \equiv \left(\mathrm{vec}(\mathbf{g}_{\mathbf{K}}), \boldsymbol{g}_{\boldsymbol{\eta}}\right) \in \mathbb{R}^{m^2+m}, \quad (4.4)$$

where the blocks $\mathbf{\Gamma_K} \in \mathbb{R}^{m^2 \times m^2}$, $\mathbf{\Gamma_{K,\eta}} \in \mathbb{R}^{m^2 \times m}$, $\mathbf{\Gamma_\eta} \in \mathbb{R}^{m \times m}$, and $\mathbf{g_K} \in \mathbb{R}^{m \times m}$, $\boldsymbol{g_\eta} \in \mathbb{R}^m$ are detailed in Appendix A. Notably, $\mathbf{\Gamma_K}$, $\mathbf{\Gamma_{K,\eta}}$ and $\mathbf{\Gamma_\eta}$ are (block-wise) sparse. Our estimates $\mathbf{K}$ and $\boldsymbol{\eta}$ are then obtained by minimizing (4.3). Further details on two special cases, Aitchison's A^{m-1} model [1] and the log–log model are given in Appendix B.

Remark 1. As noted in Section 2.2, we may average the losses obtained by removing in turn each one of the coordinates in a set $\mathcal{J} \subseteq \{1, \ldots, m\}$, instead of only m, to mitigate the dependence on the choice of the coordinate removed. This yields a quadratic loss obtained by averaging the respective matrices $\mathbf{\Gamma}$ and vectors $\boldsymbol{g}$. The time complexity of calculating $\mathbf{\Gamma}$ and $\boldsymbol{g}$ becomes linear in $|\mathcal{J}|$. However, the matrices $\mathbf{\Gamma_K}$, $\mathbf{\Gamma_{K,\eta}}$ and $\mathbf{\Gamma_\eta}$ would lose block-diagonal structure; all blocks corresponding to indices in $\mathcal{J}$ become non-zero, which makes the time complexity of coordinate descent methods for computing the loss minimizers $\hat{\mathbf{K}}$ and $\hat{\boldsymbol{\eta}}$ also linear in $|\mathcal{J}|$. Instead of attempting to make the loss independent of the choice of removed coordinate by choosing $\mathcal{J} = \{1, \ldots, m\}$, randomly sampling 5-10 coordinates is more practical for high-dimensional problems.

5 Theoretical Properties

We next present theoretical guarantees for our generalized score matching estimators when applied to the pairwise interaction power a-b models on the simplex. We consider high-dimensional settings under ℓ_1 regularization and derive bounds on the deviation of our estimates $\hat{\mathbf{K}}$ and $\hat{\boldsymbol{\eta}}$ (minimizer of (4.3)) from their true values $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ that hold with high probability.

We begin by restating Definition 12 from Yu et al. [20].

Definition 1. Let $\mathbf{\Gamma}_0 \equiv \mathbb{E}_0 \mathbf{\Gamma}(\mathbf{x})$ and $\boldsymbol{g}_0 \equiv \mathbb{E}_0 \boldsymbol{g}(\mathbf{x})$ be the expectations of $\mathbf{\Gamma}(\mathbf{x})$ and $\boldsymbol{g}(\mathbf{x})$ under the distribution given by a true parameter matrix $\mathbf{\Psi}_0 \equiv [\mathbf{K}_0, \boldsymbol{\eta}_0]^\top \in \mathbb{R}^{m(m+1)}$, or $\mathbf{\Psi}_0 \equiv \mathbf{K}_0 \in \mathbb{R}^{m^2}$ in the "centered" case with $\boldsymbol{\eta}_0 \equiv \mathbf{0}$. The support of a matrix $\mathbf{\Psi} = (\psi_{ij})$ is $S(\mathbf{\Psi}) \equiv \{(i,j) : \psi_{ij} \neq 0\}$, and we let $S_0 = S(\mathbf{\Psi}_0)$. Furthermore, let $d_{\mathbf{\Psi}_0}$ be the maximum number of non-zero entries in any column of $\mathbf{\Psi}_0$, and let $c_{\mathbf{\Psi}_0} \equiv |||\mathbf{\Psi}_0|||_{\infty,\infty}$. Writing $\mathbf{\Gamma}_{0,AB}$ for the $A \times B$ submatrix of $\mathbf{\Gamma}_0$, we define $c_{\mathbf{\Gamma}_0} \equiv |||(\mathbf{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty}$. Then $\mathbf{\Gamma}_0$ satisfies *the irrepresentability condition with incoherence parameter* $\omega \in (0,1]$ *and support set* S_0 if

$$|||\mathbf{\Gamma}_{0,S_0^c S_0}(\mathbf{\Gamma}_{0,S_0S_0})^{-1}|||_{\infty,\infty} \leq (1-\omega). \tag{5.1}$$

For simplicity, the proofs of the results in this section assume the last coordinate x_m is removed. The arguments also generalize to the case where we average $\mathbf{\Gamma}(\mathbf{x})$ and $\boldsymbol{g}(\mathbf{x})$ over multiple coordinates, following from the triangle inequality as the theorems all follow from probabilistic bounds on the deviation of $\mathbf{\Gamma}$ from $\mathbf{\Gamma}_0$ and $\boldsymbol{g}$ from $\boldsymbol{g}_0$.

5.1 Models on the Standard Simplex

For models with $a > 0$ on the simplex, the fact that each coordinate is in $[0,1]$ allows us to derive the following result, which extends Theorem 5.3 in Yu et al. [15].

Theorem 4. *Suppose* $a > 0$ *and* $b \geq 0$. *Suppose further that the true parameters* $\mathbf{K}_0$ *and* $\boldsymbol{\eta}_0$ *satisfy the conditions in Theorem 1 (so that the density is proper). Assume* $\boldsymbol{h}(\boldsymbol{x}) \equiv (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$ *with* $\alpha_1, \ldots, \alpha_m \geq \max\{1, 2-a, 2-b\}$. *Define* $\varsigma_{\boldsymbol{g}} \equiv \max_{j=1,\ldots,m} \alpha_j + \max\{|a-1|+2a, |b-1|\}$ *and* $\varsigma_{\mathbf{\Gamma}} \equiv 1$. *Suppose, without loss of generality, that* $\lambda \equiv \lambda_{\boldsymbol{K}} = \lambda_{\boldsymbol{\eta}}$; *otherwise replace* $\boldsymbol{\eta}$ *by* $(\lambda_{\boldsymbol{\eta}}/\lambda_{\mathbf{K}})\boldsymbol{\eta}$. *Suppose that* $\mathbf{\Gamma}_{0,S_0S_0}$ *is invertible and satisfies the irrepresentability condition in Equation 5.1 with* $\omega \in (0,1]$. *Suppose for* $\tau > 0$, *the sample size, the regularization parameter and the diagonal multiplier* δ *from Section 2.2 satisfy*

$$n > 72 c_{\mathbf{\Gamma}_0}^2 d_{\mathbf{\Psi}_0}^2 \varsigma_{\mathbf{\Gamma}}^2 (\tau \log m + \log 4)/\omega^2, \tag{5.2}$$

$$\lambda > \frac{3(2-\omega)}{\omega} \max\left\{ c_{\mathbf{\Psi}_0} \varsigma_{\mathbf{\Gamma}} \sqrt{2(\tau \log m + \log 4)/n}, \varsigma_{\boldsymbol{g}} \sqrt{(\tau \log m + \log 4)/(2n)} \right\}, \tag{5.3}$$

$$1 < \delta < C_{\textit{bounded}}(n, m, \tau) \equiv 1 + \sqrt{(\tau \log m + \log 4)/(2n)}. \tag{5.4}$$

Then the following statements hold with probability $1 - m^{-\tau}$:

(a) The regularized generalized $\boldsymbol{h}$*-score matching estimator* $\hat{\mathbf{\Psi}}$ *that minimizes Equation (2.8) is unique, has its support included in the true support,* $\hat{S} \equiv S(\hat{\mathbf{\Psi}}) \subseteq S_0$, *and satisfies*

$$\max\left\{ \|\hat{\mathbf{K}} - \mathbf{K}_0\|_\infty, \|\hat{\boldsymbol{\eta}} - \boldsymbol{\eta}_0\|_\infty \right\} \leq \frac{c_{\mathbf{\Gamma}_0}}{2-\omega} \lambda = \mathcal{O}\left(c_{\mathbf{\Psi}_0} \sqrt{\log m / n} \right),$$

$$\max\left\{\left|\!\left|\!\left|\hat{\mathbf{K}}-\mathbf{K}_0\right|\!\right|\!\right|_F, |||\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0|||_F\right\} \le \frac{c_{\boldsymbol{\Gamma}_0}}{2-\omega}\lambda\sqrt{|S_0|},$$
$$\max\left\{\left|\!\left|\!\left|\hat{\mathbf{K}}-\mathbf{K}_0\right|\!\right|\!\right|_2, |||\hat{\boldsymbol{\eta}}-\boldsymbol{\eta}_0|||_2\right\} \le \frac{c_{\boldsymbol{\Gamma}_0}}{2-\omega}\lambda\min\left(\sqrt{|S_0|}, d_{\boldsymbol{\Psi}_0}\right).$$

(b) Moreover, if $\min_{j,k:(j,k)\in S_0}|\kappa_{0,jk}| > \frac{c_{\boldsymbol{\Gamma}_0}}{2-\omega}\lambda$ *and* $\min_{j:(m+1,j)\in S_0}|\eta_{0,j}| > \frac{c_{\boldsymbol{\Gamma}_0}}{2-\omega}\lambda$*, then* $\hat{S} = S_0$ *and* $\operatorname{sign}(\hat{\kappa}_{jk}) = \operatorname{sign}(\kappa_{0,jk})$ *for all* $(j,k) \in S_0$ *and* $\operatorname{sign}(\hat{\eta}_j) = \operatorname{sign}(\eta_{0j})$ *for* $(m+1,j) \in S_0$.

The proof in Section D draws on the Theorem in [22], which gives the stated conclusions if estimation errors in $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ are sufficiently small. In order to show that small errors are achieved with high probability under the assumptions made, we establish bounds on $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ (which hold as long as $\alpha_1, \ldots, \alpha_m$ are large enough) and draw on arguments that invoke Hoeffding's inequality.

The trivial constant $\varsigma_{\boldsymbol{\Gamma}} \equiv 1$ is tracked for comparison to cases with $a = 0$ such as the A^{m-1} log-log models presented in Appendix B. The requirement on $\alpha_j \ge 1$ is only used for bounding the two $\partial_j(h_j \circ \varphi_j)$ terms in $\boldsymbol{g}(\mathbf{x})$. Our simulations indicate that the method also works for smaller α_j and that it might not be necessary to enforce the constraint $\alpha_j \ge 1$ in practice. The proof of Theorem 4 reveals tighter constant bounds $\varsigma_{\boldsymbol{\Gamma}}$ and $\varsigma_{\boldsymbol{g}}$ but these have rather complicated forms.

6 Numerical Experiments for A^{m-1} Models on the Simplex

This section summarizes simulation results for our regularized generalized score matching estimator for simplices. Many of the details are deferred to Appendix C.

6.1 Choices of h and C and Experimental Setup

We remove coordinate x_m. Recall that multiplication of $\nabla \log p(\boldsymbol{x})$ with $(\boldsymbol{h} \circ \boldsymbol{\varphi}_{\boldsymbol{C}})^{1/2}(\boldsymbol{x})$ is key to our method; here, the j-th component of $\boldsymbol{\varphi}_{\boldsymbol{C}}(\boldsymbol{x}) = (\varphi_{C_1,1}(\boldsymbol{x}), \ldots, \varphi_{C_{m-1},m-1}(\boldsymbol{x}))$ is the truncated distance of x_j to the boundary of its domain holding $\boldsymbol{x}_{-j}$ fixed. Thus, $\varphi_{C_j,j} = \min\{C_j, x_j, x_m\}$.

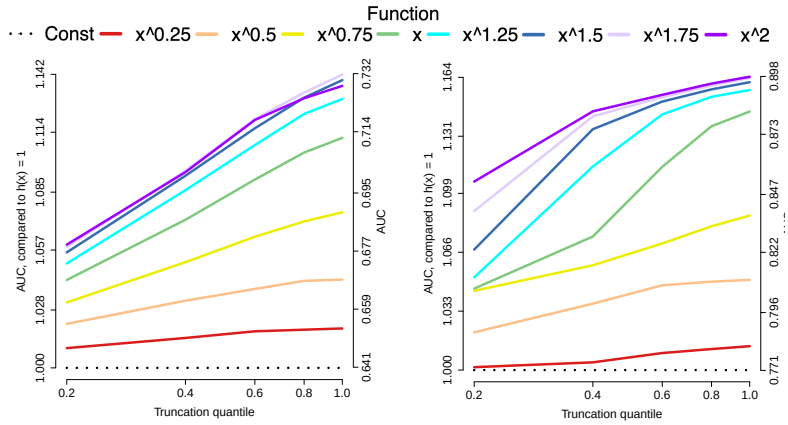
We use the same function h for all components of $\boldsymbol{h}(\boldsymbol{x}) = (h(x_1), \ldots, h(x_m))$ and compare the performance of different choices $h(x) = x^c$ for powers $c \ge 0$ along with various truncation points $\boldsymbol{C}$. Specifically, we choose $c = i/4$ for $i = 0, 1, \ldots, 8$. Instead of pre-specifying constants $\boldsymbol{C}$, we choose a probability $\pi \in (0, 1]$ and set each C_j to be the π sample quantile of $\varphi_{1,j}$ applied to each row of the data matrix $\mathbf{x}$, namely $\{\varphi_{1,j}(\boldsymbol{x}^{(1)}), \ldots, \varphi_{1,j}(\boldsymbol{x}^{(n)})\}$, assuming there are n samples in the data. We choose $\pi \in \{0.2, 0.4, 0.6, 0.8, 1\}$, where $\pi = 1$ means no truncation for all finite φ_j values.

As a prominent model, we consider the A^{m-1} models of Section B, i.e., with $a = b = 0$ and $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ with $\mathbf{K} = \mathbf{K}^\top$. We consider dimension $m = 100$, sample sizes $n = 80$ and $n = 1000$, and assume $\boldsymbol{\eta}_0 \equiv \mathbf{0}$ is known for simplicity. The density is then proportional to $\exp\left(-\log \boldsymbol{x}^\top \mathbf{K} \log \boldsymbol{x}/2\right)$. The true interaction matrix $\mathbf{K}_0$ is a banded matrix with bandwidths $s = 7$ for $n = 1000$ and $s = 2$ for $n = 80$; the bandwidth, defined as $\max\{|i-j| : \kappa_{0,i,j} > 0\}$, is chosen so that $n/(d^2_{\mathbf{K}_0}\log m)$ is roughly constant, where d is the maximum node degree; this quantity is suggested to be linked to the probability of successful support recovery by our consistency theory in Section 5. We set $\kappa_{0,i,j}$ to $1 - |i-j|/(s+1)$ for $1 \le |i-j| \le s$, and set the diagonals so that $\mathbf{K}_0\mathbf{1}_m = \mathbf{0}_m$. Each n is thus associated with only one graph, for which we run 50 trials.

We investigate recovery of the interaction pattern given by the support of $\mathbf{K}_0$, as well as estimation of the entries of $\mathbf{K}_0$. The off-diagonal entries in the support of $\mathbf{K}_0$ naturally define the edges of a graph. In the sequel, we will thus refer to edge recovery and plot estimation results in terms of such graphs.

6.2 AUCs and Estimation Errors

To investigate the recovery of graphical interaction patterns, we consider the *areas under the ROC curve* (AUCs); the ROC curve plots the *true positive rate (TPR)* against the *false positive rate (FPR)*, defined as $\mathrm{TPR} \equiv |\hat{S}_{\text{off}} \cap S_{0,\text{off}}|/|S_{0,\text{off}}|$ and $\mathrm{FPR} \equiv |\hat{S}_{\text{off}} \backslash S_{0,\text{off}}|/(m(m-1) - |S_{0,\text{off}}|)$, where $\hat{S}_{\text{off}}$ and $S_{0,\text{off}}$ are the sets of estimated and true edges, i.e., pairs of distinct indices (i,j) in the support of the estimated and the true interaction matrix, respectively. We conduct 50 trials and compute average AUCs as functions of $\pi \in \{0.2, 0.4, 0.6, 0.8, 1\}$, which correspond to the column-wise sample quantiles used as the truncation points $\boldsymbol{C}$ for $\boldsymbol{\varphi}_{\boldsymbol{C}}$ (c.f. Section 6.1). In doing so, we fix



(a) $n = 80$, $|\mathcal{J}| = 5$ **(b)** $n = 1000$, $|\mathcal{J}| = 5$

Figure 2: AUCs averaged over 50 trials for edge recovery for the A^{m-1} models on the simplex.

the diagonal multiplier to the upper bound in (5.4); compare Appendix C.2 . We then form curves representing $h(x) = 1$, or $h(x) = x^c$ with $c = 1/4, 1/2, \ldots, 2$.

As discussed in Sections 2.2 and 4, the effect of the choice of the removed coordinate may be reduced via a loss obtained by averaging with respect to a randomly sampled set of coordinates $\mathcal{J}$. The results for edge recovery with $|\mathcal{J}| = 5$ are shown in Figure 4, where $h(x) = x^2$ is among the best performers, with $C_j \equiv 1$ (no truncation) being a safe choice. This conforms to our previous conclusion of the choice of $h(x) = x^{\max\{2-a,0\}}$ for a-b type models on full-dimensional domains [15, 20].

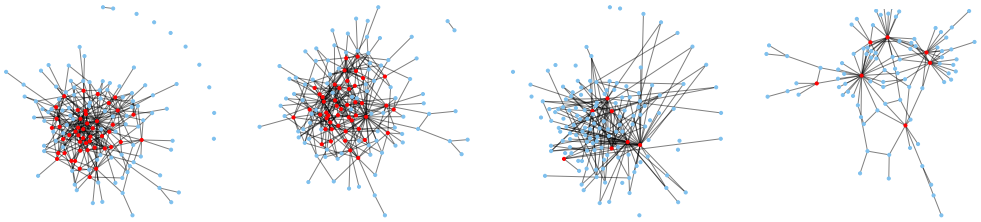
We also investigated how the choice of $|\mathcal{J}| = 5$ compares to using $|\mathcal{J}| = 1$ only. The conclusion here is that while using multiple removed coordinates $\mathcal{J}$ is beneficial for the high-dimensional case, the improvement may not justify the added computational burden in the low-dimensional case. For space reasons the figures supporting this conclusion are deferred to Appendix C.1, where we also report on further results on the estimation error in spectral and Frobenius norms, with similar conclusions.

7 Analysis of Microbiome Data

We illustrate our method by analyzing the human gut microbiome dataset studied in Yatsunenko et al. [23] and Wang et al. [24]. The dataset yields relative abundance measurements for $m = 149$ microbes for $n = 100$ healthy children and adults. As in [24], we split the 100 samples into two age groups, with $n_1 = 67$ individuals of age < 3 years as the first group, and $n_2 = 33$ of age ≥ 3 as the second. We then conduct a differential analysis of their microbial interaction networks [25], i.e., we estimate key differences between the interaction matrices $\mathbf{K}_1, \mathbf{K}_2 \in \mathbb{R}^{149\times 149}$ for the two groups.

We apply two versions of a-b models (Equation 1.2) on Δ_{148}: (i) the A^{148} model of Aitchison [10] with $a = b = 0$ (see Section B.1), and (ii) the simplex-restricted exponential square-root model [14] with $a = b = 1/2$ (see Section 4). For each setting, we use permutation tests with $B = 500$ trials, where each trial randomly partitions the n data points into two subsets of size n_1 and n_2. In the estimation of the respective interaction matrices $\mathbf{K}_1 = [\kappa_{1,jk}]$ and $\mathbf{K}_2 = [\kappa_{2,jk}]$, we choose the tuning parameter λ by cross validation and set the diagonal multipliers to the upper bound in (5.4). As discussed in Section 6.2, since the data is high-dimensional, it is beneficial to use $|\mathcal{J}| = 5$, where we choose $\mathcal{J} = \{29, 58, 87, 116, 145\}$. The h functions are chosen as $h(x) = x^{2-a}$.

Let $(\widehat{\mathbf{K}}_1, \widehat{\mathbf{K}}_2)$ be the estimates for the original data, and let $(\widehat{\mathbf{K}}_{1,(b)}, \widehat{\mathbf{K}}_{2,(b)})$, $b = 1, \ldots, B$, be those for the resampled data. To discern differences in $\mathbf{K}_1$ and $\mathbf{K}_2$, we test all (local) hypotheses $\kappa_{1,jk} = \kappa_{2,jk}$, for $1 < j < k < 149$, with p-values $\frac{1}{B}\sum_{i=1}^{B} \mathbb{1}(|\hat{\kappa}_{1,jk} - \hat{\kappa}_{2,jk}| \leq |\hat{\kappa}_{1,(b),jk} - \hat{\kappa}_{2,(b),jk}|)$ and controlling the overall false discovery rate at level 0.05 using Benjamini and Yekutieli [26]. This leads to the differential graphs in Figure 3, where an edge $j - k$ indicates rejection of hypothesis $\kappa_{1,jk} = \kappa_{2,jk}$. The figure highlights hub nodes of degree at least 5. These correspond to microbes that in their interaction with other microbes behave very differently in the two age groups.



(a) A^{148} model, layout optimized

(b) A^{148} model, no isolated nodes, layout optimized

(c) Exp model, same layout as (a)

(d) Exp model, no isolated nodes, layout optimized

Figure 3: Differential graphs estimated by regularized generalized score matching estimator with permutation tests assuming the A^{m-1} model (a, b) and the exponential square-root model (c, d). Red points indicate nodes with degree at least 5 ("hub nodes").

8 Discussion

Building on the ideas from [17, 18], the method of Yu et al. [15] estimates densities supported on general domains using a generalized score matching loss. However, the domains considered by Yu et al. [15] are required to have positive Lebesgue measure in order to guarantee consistent estimation. In this paper, we demonstrate how to extend their method to the case of compositional data on a probability simplex of general dimension. Specifically, we show how profiling out the last component of $\boldsymbol{x}$ yields an effective methodology for a flexible class of interaction models for compositional data.

We focus on a-b pairwise interaction models with density proportional to

$$\exp\{-\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a/(2a) + \boldsymbol{\eta}^\top\boldsymbol{x}^b/b\},$$

where for $a = 0$ we let $\boldsymbol{x}^{a\top}\mathbf{K}\boldsymbol{x}^a/(2a) \equiv \log\boldsymbol{x}^\top\mathbf{K}\log\boldsymbol{x}/2$ and for $b = 0$, $\boldsymbol{\eta}^\top\boldsymbol{x}^b/b \equiv \boldsymbol{\eta}^\top\log\boldsymbol{x}$. For this class, our results detail the construction of estimators for simplex domains, with the appendix giving additional details for the well-known special case of A^{m-1} models [10].

In our theoretical treatment, we show that for general a-b models on $(m-1)$-dimensional simplex domains and with $a > 0$, the graph encoding the sparsity pattern of the interaction matrix $\mathbf{K}$ may be recovered successfully when the sample size is of order $n = \Omega(\log m)$. This directly parallels similar results obtained in prior work for unconstrained domains. In the case of $a = 0$, we require an additional multiplicative factor that may weakly depend on m.

In order to account for boundary effects, our method introduces a set of weights in the score matching loss. Through simulation studies, we confirm that weights derived from the choice of a function $\boldsymbol{h}(\boldsymbol{x}) = (x_1^c, \ldots, x_m^c)$ with $c = \max\{2-a, 0\}$ perform the best in most settings in terms of edge recovery, generalizing the conclusion in Yu et al. [15].

Two problems naturally emerge as topics for future work. On the one hand, it would be interesting to extend our theoretical results on a-b models with $a = 0$ in order to get a full understanding of the sample complexity of our estimators; see the discussion after Corollary 8. On the other hand, it would be interesting to develop a more systematic way to deal with Lebesgue-null sets beyond simplices.

Acknowledgements

This project has received funding from the European Research Council (ERC) under the European Union's Horizon 2020 research and innovation programme (grant agreement No 883818), as well as funding from the U.S. National Institutes of Health under grant R01GM133848.

References

[1] John Aitchison. The statistical analysis of compositional data. *J. Roy. Statist. Soc. Ser. B*, 44(2): 139–177, 1982. 1, 6, 13, 15

[2] Vera Pawlowsky-Glahn and Juan José Egozcue. Compositional data and their analysis: An introduction. *Geological Society, London, Special Publications*, 264(1):1–10, 2006. 1

[3] Christopher D Lloyd, Vera Pawlowsky-Glahn, and Juan José Egozcue. Compositional data analysis in population studies. *Annals of the Association of American Geographers*, 102(6): 1251–1266, 2012. 1

[4] H. R. Rollinson. Another look at the constant sum problem in geochemistry. *Mineralogical Magazine*, 56(385):469–475, 1992. 1

[5] Julia Fukuyama, Paul J. McMurdie, Les Dethlefsen, David A. Relman, and Susan Holmes. Comparisons of distance methods for combining covariates and abundances in microbiome studies. In *Biocomputing 2012*, pages 213–224. World Scientific, 2012. 1

[6] Fan Xia, Jun Chen, Wing Kam Fung, and Hongzhe Li. A logistic normal multinomial regression model for microbiome compositional data analysis. *Biometrics*, 69(4):1053–1063, 2013.

[7] Hongzhe Li. Microbiome, metagenomics, and high-dimensional compositional data analysis. *Annual Review of Statistics and Its Application*, 2(1):73–94, 2015.

[8] Pixu Shi, Anru Zhang, and Hongzhe Li. Regression analysis for microbiome compositional data. *Ann. Appl. Stat.*, 10(2):1019–1040, 2016.

[9] Timothy W. Randolph, Sen Zhao, Wade Copeland, Meredith Hullar, and Ali Shojaie. Kernel-penalized regression for analysis of microbiome data. *Ann. Appl. Stat.*, 12(1):540–566, 2018. 1

[10] John Aitchison. A general class of distributions on the simplex. *J. Roy. Statist. Soc. Ser. B*, 47 (1):136–146, 1985. 1, 8, 9, 13, 14, 15

[11] 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. 1

[12] Mathias Drton and Marloes H. Maathuis. Structure learning in graphical modeling. *Annual Review of Statistics and Its Application*, 4:365–393, 2017. 1

[13] Vera Pawlowsky-Glahn, Juan José Egozcue, and Raimon Tolosana Delgado. Lecture notes on compositional data analysis, 2007. 1

[14] David Inouye, Pradeep Ravikumar, and Inderjit Dhillon. Square root graphical models: Multivariate generalizations of univariate exponential families that permit positive dependencies. In *ICML*, pages 2445–2453, 2016. 1, 8

[15] Shiqing Yu, Mathias Drton, and Ali Shojaie. Generalized score matching for general domains. *Inf. Inference*, 11(2):739–780, 2022. 2, 3, 4, 6, 8, 9, 12, 23, 25

[16] Janice L. Scealy and Andrew T. A. Wood. Score matching for compositional distributions. *J. Amer. Statist. Assoc.*, 118(543):1811–1823, 2023. 2

[17] Aapo Hyvärinen. Estimation of non-normalized statistical models by score matching. *J. Mach. Learn. Res.*, 6:695–709, 2005. 2, 9, 15, 17

[18] Aapo Hyvärinen. Some extensions of score matching. *Comput. Statist. Data Anal.*, 51(5): 2499–2512, 2007. 2, 9

[19] Shiqing Yu, Mathias Drton, and Ali Shojaie. Graphical models for non-negative data using generalized score matching. In *AISTATS*, pages 1781–1790, 2018. 2

[20] Shiqing Yu, Mathias Drton, and Ali Shojaie. Generalized score matching for non-negative data. *J. Mach. Learn. Res.*, 20(76):1–70, 2019. 2, 4, 6, 8, 12, 25

[21] Song Liu, Takafumi Kanamori, and Daniel J. Williams. Estimating density models with truncation boundaries using score matching. *J. Mach. Learn. Res.*, 23(186):1–38, 2022. 2

[22] Lina Lin, Mathias Drton, and Ali Shojaie. Estimation of high-dimensional graphical models using regularized score matching. *Electron. J. Stat.*, 10(1):806–854, 2016. 7, 22

[23] Tanya Yatsunenko, Federico E Rey, Mark J Manary, Indi Trehan, Maria Gloria Dominguez-Bello, Monica Contreras, Magda Magris, Glida Hidalgo, Robert N Baldassano, Andrey P Anokhin, et al. Human gut microbiome viewed across age and geography. *Nature*, 486(7402): 222–227, 2012. 8

[24] Yue Wang, Timothy W Randolph, Ali Shojaie, and Jing Ma. The generalized matrix decomposition biplot and its application to microbiome data. *Msystems*, 4(6), 2019. 8

[25] Ali Shojaie. Differential network analysis: A statistical perspective. *Wiley Interdisciplinary Reviews: Computational Statistics*, 13(2):e1508, 2021. 8

[26] Yoav Benjamini and Daniel Yekutieli. The control of the false discovery rate in multiple testing under dependency. *Ann. Statist.*, 29(4):1165–1188, 2001. 8

[27] Martin J. Wainwright. *High-Dimensional Statistics: A Non-Asymptotic Viewpoint*. Cambridge University Press, 2019. 25

[28] Roman Vershynin. Introduction to the non-asymptotic analysis of random matrices. In *Compressed Sensing*, pages 210–268. Cambridge University Press, 2012. 25

A Derivation Details

In this section, we provide technical details for some of the derivations in the main paper. In particular, this section collects the assumptions needed for obtaining a practical sample loss, as well as details of block matrices from Section 4.

Assumptions. The following are assumptions needed to derive a practical sample loss.

(A.1) $p_0(x_j;\boldsymbol{x}_{-j})h_j(\varphi_{C_j,\mathcal{D},j}(\boldsymbol{x}))\partial_j \log p(x_j;\boldsymbol{x}_{-j}) \Big|_{x_j\nearrow b_k(\boldsymbol{x}_{-j})^-}^{x_j\searrow a_k(\boldsymbol{x}_{-j})^+} = 0$
for all $k = 1,\dots,K_j(\boldsymbol{x}_{-j})$ and $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j,\mathcal{D}}$ for all j;

(A.2) $\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\|\nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h}\circ\boldsymbol{\varphi}_{\boldsymbol{C},\boldsymbol{\mathcal{D}}})^{1/2}(\boldsymbol{x})\right\|_2^2 \mathrm{d}\boldsymbol{x} < +\infty$,
$\int_{\mathcal{D}} p_0(\boldsymbol{x}) \left\|[\nabla \log p(\boldsymbol{x}) \odot (\boldsymbol{h}\circ\boldsymbol{\varphi}_{\boldsymbol{C},\boldsymbol{\mathcal{D}}})(\boldsymbol{x})]'\right\|_1 \mathrm{d}\boldsymbol{x} < +\infty$.

(A.3) $\forall j = 1,\dots,m$ and a.e. $\boldsymbol{x}_{-j} \in \mathcal{S}_{-j,\mathcal{D}}$, the component function h_j of $\boldsymbol{h}$ is absolutely continuous in any bounded sub-interval of the section $\mathcal{C}_{j,\mathcal{D}}(\boldsymbol{x}_{-j})$.

Block Matrices. The following are block matrices used in Section 4. We note that in comparison to the full-dimensional setting of [15, 20], the blocks $\boldsymbol{\Gamma}_{\mathbf{K}} \in \mathbb{R}^{m^2\times m^2}$, $\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \in \mathbb{R}^{m(m+1)}$ and $\boldsymbol{\Gamma}_{\boldsymbol{\eta}} \in \mathbb{R}^{m\times m}$ are no longer block-diagonal with m blocks, due to the substitution of x_m. Instead,

$$\boldsymbol{\Gamma}_{\mathbf{K}} \equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},(1,m)} \\ \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},2} & \cdots & \mathbf{0} & \boldsymbol{\Gamma}_{\mathbf{K},(2,m)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\Gamma}_{\mathbf{K},m-1} & \boldsymbol{\Gamma}_{\mathbf{K},(m-1,m)} \\ \boldsymbol{\Gamma}_{\mathbf{K},(1,m)}^\top & \boldsymbol{\Gamma}_{\mathbf{K},(2,m)}^\top & \cdots & \boldsymbol{\Gamma}_{\mathbf{K},(m-1,m)}^\top & \boldsymbol{\Gamma}_{\mathbf{K},m} \end{bmatrix} \in \mathbb{R}^{m^2\times m^2},$$

with each block of size $m \times m$, and

$$\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \equiv \begin{bmatrix} \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},1} & \mathbf{0} & \cdots & \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(1,m)} \\ \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},2} & \cdots & \mathbf{0} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(2,m)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ \mathbf{0} & \mathbf{0} & \cdots & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m-1} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m-1,m)} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,1)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,2)} & \cdots & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,m-1)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m} \end{bmatrix} \in \mathbb{R}^{m^2\times m},$$

with each block a vector of size m, and

$$\boldsymbol{\Gamma}_{\boldsymbol{\eta}} \equiv \begin{bmatrix} \gamma_{\boldsymbol{\eta},1} & 0 & \cdots & 0 & \gamma_{\boldsymbol{\eta},(1,m)} \\ 0 & \gamma_{\boldsymbol{\eta},2} & \cdots & 0 & \gamma_{\boldsymbol{\eta},(2,m)} \\ \vdots & \vdots & \ddots & \vdots & \vdots \\ 0 & 0 & \cdots & \gamma_{\boldsymbol{\eta},m-1} & \gamma_{\boldsymbol{\eta},(m-1,m)} \\ \gamma_{\boldsymbol{\eta},(1,m)} & \gamma_{\boldsymbol{\eta},(2,m)} & \cdots & \gamma_{\boldsymbol{\eta},(m-1,m)} & \gamma_{\boldsymbol{\eta},m} \end{bmatrix} \in \mathbb{R}^{m\times m}.$$

The specific form of the blocks appearing in the preceding displays is as follows. Using the shorthand $\tilde{h}_j \equiv h_j \circ \varphi_j$, we have for $j = 1,\dots,m-1$,

$$\begin{aligned} \boldsymbol{\Gamma}_j &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},j} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},j} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},j}^\top & \gamma_{\boldsymbol{\eta},j} \end{bmatrix} \\ &\equiv \frac{1}{n}\sum_{i=1}^n \tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_j^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_j^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_j^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_j^{(i)}}^{b-1} \end{bmatrix}^\top, \\ \boldsymbol{\Gamma}_m &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},m} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},m}^\top & \gamma_{\boldsymbol{\eta},m} \end{bmatrix} \\ &\equiv \frac{1}{n}\sum_{i=1}^n \sum_{k=1}^{m-1} \tilde{h}_k\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_m^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_m^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_m^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_m^{(i)}}^{b-1} \end{bmatrix}^\top, \end{aligned}$$

$$\begin{aligned}\boldsymbol{\Gamma}_{(j,m)} &\equiv \begin{bmatrix} \boldsymbol{\Gamma}_{\mathbf{K},(j,m)} & \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(j,m)} \\ \boldsymbol{\gamma}_{\mathbf{K},\boldsymbol{\eta},(m,j)} & \gamma_{\boldsymbol{\eta},(j,m)} \end{bmatrix} \\ &\equiv -\frac{1}{n}\sum_{i=1}^{n} \tilde{h}_j\left(\boldsymbol{X}^{(i)}\right) \begin{bmatrix} {X_j^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_j^{(i)}}^{b-1} \end{bmatrix} \begin{bmatrix} {X_m^{(i)}}^{a-1}\boldsymbol{X}^{(i)a} \\ -{X_m^{(i)}}^{b-1} \end{bmatrix}^\top.\end{aligned}$$

In addition,

$$\begin{aligned}\boldsymbol{g}_{\mathbf{K},j} &\equiv \frac{1}{n}\sum_{i=1}^{n}\left[\partial_j\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{a-1} + (a-1)\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{a-2}\right]\boldsymbol{X}^{(i)a} \\ &\quad + a\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{2a-2}\boldsymbol{e}_{j,m} - a\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{a-1}{X_m^{(i)}}^{a-1}\boldsymbol{e}_{m,m}, \\ \boldsymbol{g}_{\mathbf{K},m} &\equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m-1}\left[-\partial_k\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_m^{(i)}}^{a-1} + (a-1)\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_m^{(i)}}^{a-2}\right]\boldsymbol{X}^{(i)a} \\ &\quad + a\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_m^{(i)}}^{2a-2}\boldsymbol{e}_{m,m} - a\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_k^{(i)}}^{a-1}{X_m^{(i)}}^{a-1}\boldsymbol{e}_{k,m}, \\ \boldsymbol{g}_{\boldsymbol{\eta},j} &\equiv \frac{1}{n}\sum_{i=1}^{n} -\partial_j\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{b-1} - (b-1)\tilde{h}_j\left(\boldsymbol{X}^{(i)}\right){X_j^{(i)}}^{b-2}, \\ \boldsymbol{g}_{\boldsymbol{\eta},m} &\equiv \frac{1}{n}\sum_{i=1}^{n}\sum_{k=1}^{m-1}\partial_k\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_m^{(i)}}^{b-1} - (b-1)\tilde{h}_k\left(\boldsymbol{X}^{(i)}\right){X_m^{(i)}}^{b-2}.\end{aligned}$$

B Other Interaction Models on Simplex

In this section, we detail two special cases of the general models in (1.2), namely, Aitchison's A^{m-1} model [1], as well as the log–log model on the standard simplex.

B.1 Estimation for A^{m-1} Models

In Section 4, we described how to estimate to form generalized score matching estimators of $\mathbf{K}$ and $\boldsymbol{\eta}$. The discussion there applies to log–log models ($a = b = 0$), but only under the setting of assumption (I) from Theorem 6, where $\mathbf{K}$ is unconstrained except for positive definiteness. For the A^{m-1} models [10] which impose the additional constraint that $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K} = \mathbf{K}^\top$ as in (II) and (III) of Theorem 6, we need the following modification. We marginalize out the diagonals of $\mathbf{K}$ with $\kappa_{jj} = -\boldsymbol{\kappa}_{-j,j}^\top\mathbf{1}_{m-1}$ and estimate all off-diagonal elements $\mathbf{K}_{\text{off}} \equiv [\boldsymbol{\kappa}_{-1,1}, \ldots, \boldsymbol{\kappa}_{-m,m}]$. Under the additional constraint, for matrices $\mathbf{A}$ with m rows and $\mathbf{B}$ with m columns, we can write

$$\begin{aligned}\boldsymbol{\kappa}_{\cdot,j}^\top\mathbf{A} &= \boldsymbol{\kappa}_{-j,j}^\top\mathbf{A}_{-j,\cdot} + \kappa_{jj}\boldsymbol{a}_{j,\cdot}^\top = \boldsymbol{\kappa}_{-j,j}^\top\left(\mathbf{A}_{-j,\cdot} - \mathbf{1}_{m-1}\boldsymbol{a}_{j,\cdot}^\top\right) = \boldsymbol{\kappa}_{-j,j}^\top\left(\mathbf{C}(j)\mathbf{A}\right), \\ \mathbf{B}\boldsymbol{\kappa}_{\cdot,j} &= \mathbf{B}_{\cdot,-j}\boldsymbol{\kappa}_{-j,j} + \boldsymbol{b}_{\cdot,j}\kappa_{jj} = \left(\mathbf{B}_{\cdot,-j} - \boldsymbol{b}_{\cdot,j}\mathbf{1}_{m-1}^\top\right)\boldsymbol{\kappa}_{-j,j} = \left(\mathbf{B}\mathbf{C}(j)^\top\right)\boldsymbol{\kappa}_{-j,j},\end{aligned}$$

where $\mathbf{C}(j) \in \mathbb{R}^{(m-1)\times m}$ has its j-th column all equal to -1, and the entries $(1,1), \ldots, (j-1, j-1), (j, j+1), \ldots, (m-1, m)$ equal to 1, and all other entries zero. Let $\mathbf{C} \in \mathbb{R}^{(m-1)m\times m^2}$ be the block-diagonal matrix with blocks $\mathbf{C}(1), \ldots, \mathbf{C}(m)$. The unpenalized generalized score-matching loss given by the first two terms in (4.3) thus becomes

$$\frac{1}{2}\text{vec}\left(\begin{bmatrix}\mathbf{K}_{\text{off}} \\ \boldsymbol{\eta}^\top\end{bmatrix}\right)^\top \underbrace{\begin{bmatrix}\mathbf{C}\boldsymbol{\Gamma}_{\mathbf{K}}\mathbf{C}^\top & \mathbf{C}\boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}} \\ \boldsymbol{\Gamma}_{\mathbf{K},\boldsymbol{\eta}}^\top\mathbf{C}^\top & \boldsymbol{\Gamma}_{\boldsymbol{\eta}}\end{bmatrix}}_{\equiv\tilde{\boldsymbol{\Gamma}}\in\mathbb{R}^{m^2\times m^2}} \text{vec}\left(\begin{bmatrix}\mathbf{K}_{\text{off}} \\ \boldsymbol{\eta}^\top\end{bmatrix}\right) \quad - \quad \underbrace{\begin{bmatrix}\mathbf{C}\text{vec}(\boldsymbol{g}_{\mathbf{K}}) \\ \boldsymbol{g}_{\boldsymbol{\eta}}\end{bmatrix}^\top}_{\equiv\tilde{\boldsymbol{g}}\in\mathbb{R}^{m^2}} \text{vec}\left(\begin{bmatrix}\mathbf{K}_{\text{off}} \\ \boldsymbol{\eta}^\top\end{bmatrix}\right).$$

It holds that $\tilde{\boldsymbol{\Gamma}}$ is positive (semi-)definite if and only if $\boldsymbol{\Gamma}$ from (4.4) is positive (semi-)definite.

When applying to the diagonal multiplication operation $(\cdot)_\delta$ to form a loss as in (2.8), we simply operate directly on the matrix $\tilde{\boldsymbol{\Gamma}}$, rather than the matrix $\boldsymbol{\Gamma}$. Since $\tilde{\boldsymbol{\Gamma}}$ is merely a linear transformation of $\boldsymbol{\Gamma}$, later high probability bounds on are not affected, except in constants. The penalized generalized score-matching loss $\hat{L}_{\boldsymbol{h},\boldsymbol{C},\lambda,\delta}(p_{\mathbf{K},\boldsymbol{\eta}})$ for the A^{m-1} models is thus defined as

$$\frac{1}{2}\mathrm{vec}\left(\begin{bmatrix}\mathbf{K}_{\mathrm{off}}\\ \boldsymbol{\eta}^\top\end{bmatrix}\right)^\top \tilde{\boldsymbol{\Gamma}}_\delta(\mathbf{x})\mathrm{vec}\left(\begin{bmatrix}\mathbf{K}_{\mathrm{off}}\\ \boldsymbol{\eta}^\top\end{bmatrix}\right) - \tilde{\boldsymbol{g}}(\mathbf{x})^\top \mathrm{vec}\left(\begin{bmatrix}\mathbf{K}_{\mathrm{off}}\\ \boldsymbol{\eta}^\top\end{bmatrix}\right) + \lambda_{\mathbf{K}}\|\mathrm{vec}(\mathbf{K}_{\mathrm{off}})\|_1 + \lambda_{\boldsymbol{\eta}}\|\boldsymbol{\eta}\|_1,$$

where $\tilde{\boldsymbol{\Gamma}}_\delta$ is simply $\tilde{\boldsymbol{\Gamma}}$ with $\mathbf{C}\boldsymbol{\Gamma}_{\mathbf{K}}\mathbf{C}^\top$ replaced by $\left(\mathbf{C}\boldsymbol{\Gamma}_{\mathbf{K}}\mathbf{C}^\top\right)_\delta$.

For models with $a = 0$ on simplex domains, including the A^{m-1} models discussed in Section B.2, we next derive the following lemma to bound $\log X_j$ with high probability.

Lemma 5. *Suppose $\boldsymbol{X}$ has the density from (1.2) on Δ with true parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ satisfying the conditions in Theorem 1 for $a > 0$ or $b > 0$, or in Theorem 6 for $a = b = 0$. Then for all $j = 1, \ldots, m$, X_j^{2a} is sub-exponential for $a > 0$, and $\log X_j$ is sub-exponential for $a = 0$.*

B.2 Estimation for log–log Models on the Standard Simplex

In this section, we discuss the special case of $a = 0$ and $b = 0$, namely, models with density proportional to

$$\exp\left(-\frac{1}{2}\log \boldsymbol{x}^\top \mathbf{K} \log \boldsymbol{x} + \boldsymbol{\eta}^\top \log \boldsymbol{x}\right) \tag{B.1}$$

supported on the $(m-1)$-dimensional standard simplex Δ. This class encompasses the A^{m-1} class of distributions in Equation (2.7) of Aitchison [10], which have parameters $\boldsymbol{\beta} \equiv (\beta_j)_{j=1,\ldots,m}$ and $(\gamma_{jk})_{1\le j\ne k\le m}$, $\gamma_{jk} = \gamma_{kj}$, and density proportional to

$$\exp\left(-\frac{1}{2}\sum_{j=1}^m\sum_{k\ne j}\gamma_{jk}(\log x_j - \log x_k)^2 + (\boldsymbol{\beta} - \mathbf{1}_m)^\top \log \boldsymbol{x}\right)\mathbb{1}_\Delta(\boldsymbol{x}). \tag{B.2}$$

Indeed, expanding the exponent, the last display can be rewritten as

$$\exp\left(-\sum_{j=1}^m (\log x_j)^2\left(\sum_{k\ne j}\gamma_{jk}\right) + \sum_{j=1}^m\sum_{k\ne j}\gamma_{jk}\log x_j \log x_k + (\boldsymbol{\beta} - \mathbf{1}_m)^\top \log \boldsymbol{x}\right).$$

Letting $\boldsymbol{\eta} \equiv \boldsymbol{\beta} - \mathbf{1}_m$ and taking

$$\kappa_{jj} = 2\sum_{i\ne j}\gamma_{ji}, \quad \kappa_{kj} = \kappa_{jk} = -2\gamma_{kj}, \quad 1 \le j \ne k \le m,$$

the A^{m-1} model with densities as in (B.2) translates to the a-b model from (B.1) for $a = b = 0$ and under the constraint that $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K} = \mathbf{K}^\top$.

Next, we show that under simple conditions the density is again proper and that the population version of the generalized score matching loss can still be rewritten in the form given in (2.4). The proof is given in the Appendix D.

Theorem 6. *Suppose $\mathbf{K}$ is symmetric, and one of the following holds:*

(I) $\mathbf{K}$ *is positive definite, or*

(II) $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, $\mathbf{K}_{-k,-k}$ *is positive definite for some* $k = 1, \ldots, m$, *and* $\mathbf{1}_m^\top\boldsymbol{\eta} + m \ge 0$, *or*

(III) $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, $\mathbf{K}$ *is positive semi-definite, and* $\boldsymbol{\eta} \succ -\mathbf{1}_m$.

Then the density in (B.1) has a finite normalizing constant. Note that (II) implies that $\mathbf{K}$ is positive semi-definite and (III) implies that for all k the submatrix $\mathbf{K}_{-k,-k}$ is positive semi-definite (but not necessarily positive definite).

For all $j = 1, \ldots, m-1$, let $h_j(x) = x^{\alpha_j}$ with $\alpha_j > 0$. If (I) or (II) hold, or if (III) holds with $\alpha_j > \max\{1 - \eta_{0,j}, 1 - \eta_{0,m}\}$, then conditions (A.1)–(A.3) in Appendix A are satisfied.

We highlight in the log-log models obtained from $a = b = 0$, the parameters $\mathbf{K}$ and $\boldsymbol{\eta}$ are exactly identifiable from the density (whether assuming $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ or not). This follows in our context as a corollary of Theorem 2.

Corollary 7. *Suppose there exist $\mathbf{K}_1$, $\mathbf{K}_2$, $\boldsymbol{\eta}_1$, $\boldsymbol{\eta}_2$ such that*

$$\exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_1 \log(\boldsymbol{x}) + \boldsymbol{\eta}_1^\top \log(\boldsymbol{x})\right) = \exp\left(-\frac{1}{2}\log(\boldsymbol{x})^\top \mathbf{K}_2 \log(\boldsymbol{x}) + \boldsymbol{\eta}_2^\top \log(\boldsymbol{x})\right)$$

for all $\boldsymbol{x} \in \Delta$. Then $\mathbf{K}_1 = \mathbf{K}_2$ and $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$.

Now define the *additive log-ratio transformation*

$$\boldsymbol{y}_{-m} \equiv \log \boldsymbol{x}_{-m} - (\log x_m)\mathbf{1}_{m-1} = (\log(x_1/x_m), \ldots, \log(x_{m-1}/x_m)) .$$

The A^{m-1} model, corresponding to (II) and (III) in Theorem 6, is proposed in Aitchison [10] as a generalization of both the Dirichlet distribution and the *additive logistic normal model* [1]. In particular, using the formula (B.1), when $\mathbf{K} = \mathbf{0}$ we have the Dirichlet distribution with parameters $\boldsymbol{\eta} + \mathbf{1}_m$, which belongs to case (III) in Theorem 6. On the other hand, if $\mathbf{1}_m^\top \boldsymbol{\eta} = -m$, we get the normal density in $\boldsymbol{y}_{-m}$ with inverse covariance $\mathbf{K}_{-m,-m}$ and mean $\mathbf{K}_{-m,-m}^{-1}\boldsymbol{\eta}_{-m}$, which belongs to case (II) in Theorem 6. The generalization uses only one additional parameter when compared to the additive logistic normal model, namely $\mathbf{1}_m^\top \boldsymbol{\eta}$ is no longer assumed to be equal to $-m$.

By the nature of the simplex domain, any two proportions X_j and X_k are perfectly conditionally correlated given all other $\boldsymbol{X}_{-j,-k}$. On the other hand, under the additive logistic normal model, $Y_j = \log(X_j/X_m)$ and $Y_k = \log(X_k/X_m)$ are conditionally independent given all other $\log(X_\ell/X_m)$, $\ell \neq j, k, m$ if and only if $\kappa_{jk} = \kappa_{kj} = 0$. As we make clear in the proof of Theorem 6, this is true only for the additive logistic normal model ($\mathbf{1}_m^\top \boldsymbol{\eta} = -m$).

We end this section by presenting a corollary of Lemma 5 in Section 5.1 that establishes the consistency of our estimator for the $a = b = 0$ case.

Corollary 8. *Suppose $a = b = 0$. Also suppose the conditions for $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ in Theorem 6 hold, or $b > 0$ and the condition in Theorem 1 hold. Let $\boldsymbol{h}(\boldsymbol{x}) \equiv (x_1^{\alpha_1}, \ldots, x_m^{\alpha_m})$ with $\alpha_1, \ldots, \alpha_m \geq 2$. Then Theorem 4 holds with $\log 4$ replaced by $\log 6$ in (5.2)-(5.4), and*

$$\varsigma_{\boldsymbol{\Gamma}} \equiv \max\left\{1, c_{\log,\mathbf{K}_0,\boldsymbol{\eta}_0}^2\right\},$$
$$\varsigma_{\boldsymbol{g}} \equiv \max\left\{\left(\max_{j=1,\ldots,m} \alpha_j + 1\right) c_{\log,\mathbf{K}_0,\boldsymbol{\eta}_0} + 2, \max_j \alpha_j + |b-1|\right\},$$

where

$$\begin{aligned} c_{\log,\mathbf{K}_0,\boldsymbol{\eta}_0} \equiv{}& \max_j \mathbb{E}_0 \log X_j \\ &+ \max\left\{2\sqrt{2}e \max_{j=1,\ldots,m} \|\log X_j\|_{\psi_1} \sqrt{\log 3 + \log n + (\tau+1)\log m},\right. \\ &\qquad \left. 4e \max_{j=1,\ldots,m} \|\log X_j\|_{\psi_1} (\log 3 + \log n + (\tau+1)\log m)\right\}, \end{aligned}$$

and $\|\log X_j\|_{\psi_1} \equiv \sup_{q\geq 1} \left(\mathbb{E}_0 |\log X_j|^q\right)^{1/q}/q \geq -\mathbb{E}_0 \log X_j$.

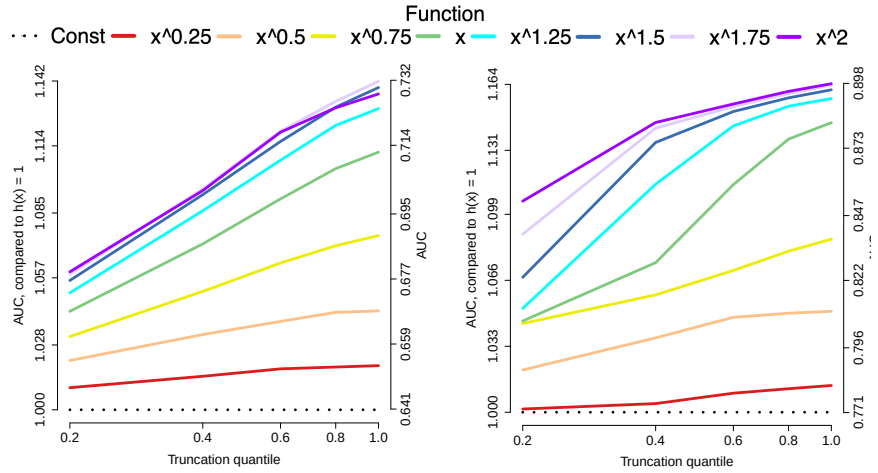
The results are written in terms of the maximum of the sub-exponential norms of $\log X_1, \ldots, \log X_m$, and they indicate the sample size requirement that $n = \Omega(\log m)\, \mathcal{O}\left(\max_j \|\log X_j\|_{\psi_1}^2\right)$. We expect that the maximum of the sub-exponential norms scales as $\Omega\left((\log m)^c\right)$ for some c small, although we cannot currently offer an exact result on this behavior.

C Details of Numerical Studies

C.1 Results for Section 6

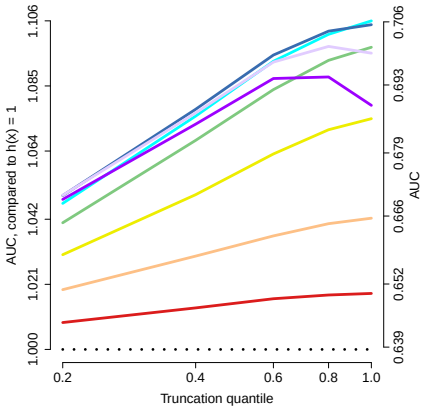
Average AUCs over 50 trials are shown in Figure 4 as functions of $\pi \in \{0.2, 0.4, 0.6, 0.8, 1\}$, which correspond to the column-wise sample quantiles used as the truncation points $\boldsymbol{C}$ for $\boldsymbol{\varphi}_{\boldsymbol{C}}$ (c.f. Section 6.1). Each curve in the figure represents $h(x) = 1$, or $h(x) = x^c$ with $c = 1/4, 1/2, \ldots, 2$. The y-ticks on the right-hand side denote the corresponding AUCs, while those on the left are the AUCs divided by the AUC for $h(x) = 1$ as a reference, measuring the relative performance of each method compared with the estimator for densities on $\mathbb{R}^m$ first given by Hyvärinen [17]; the dotted line corresponds to the AUC for $h(x) = 1$. We fix the diagonal multiplier to the upper bound in (5.4).

As discussed in Sections 2.2 and 4, to reduce the effect of the choice of the removed coordinate, one can randomly sample a set of coordinates $\mathcal{J}$, calculate $\boldsymbol{\Gamma}$ and $\boldsymbol{\eta}$ by removing one coordinate x_j, $j \in \mathcal{J}$ at a time, and take the average. In the first row, we plot the results for such $\boldsymbol{\Gamma}$ and $\boldsymbol{\eta}$ constructed with $\mathcal{J}$ randomly sampled from $\{1, \ldots, m\}$, where $|\mathcal{J}| = 5$. In order to investigate the

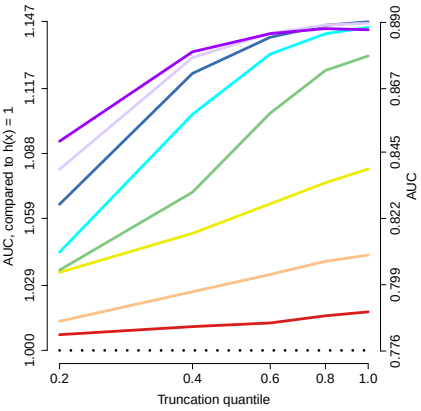


(a) $n = 80$, $|\mathcal{J}| = 5$

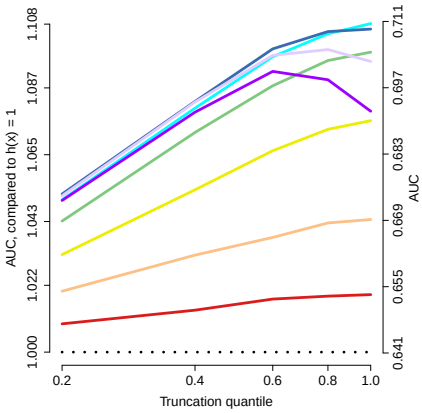
(b) $n = 1000$, $|\mathcal{J}| = 5$



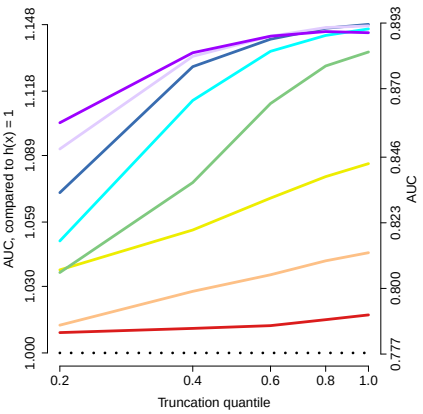
(c) $n = 80$, average of 5 runs with each $j \in |\mathcal{J}|$



(d) $n = 1000$, average of 5 runs



(e) $n = 80$, max of 5 runs

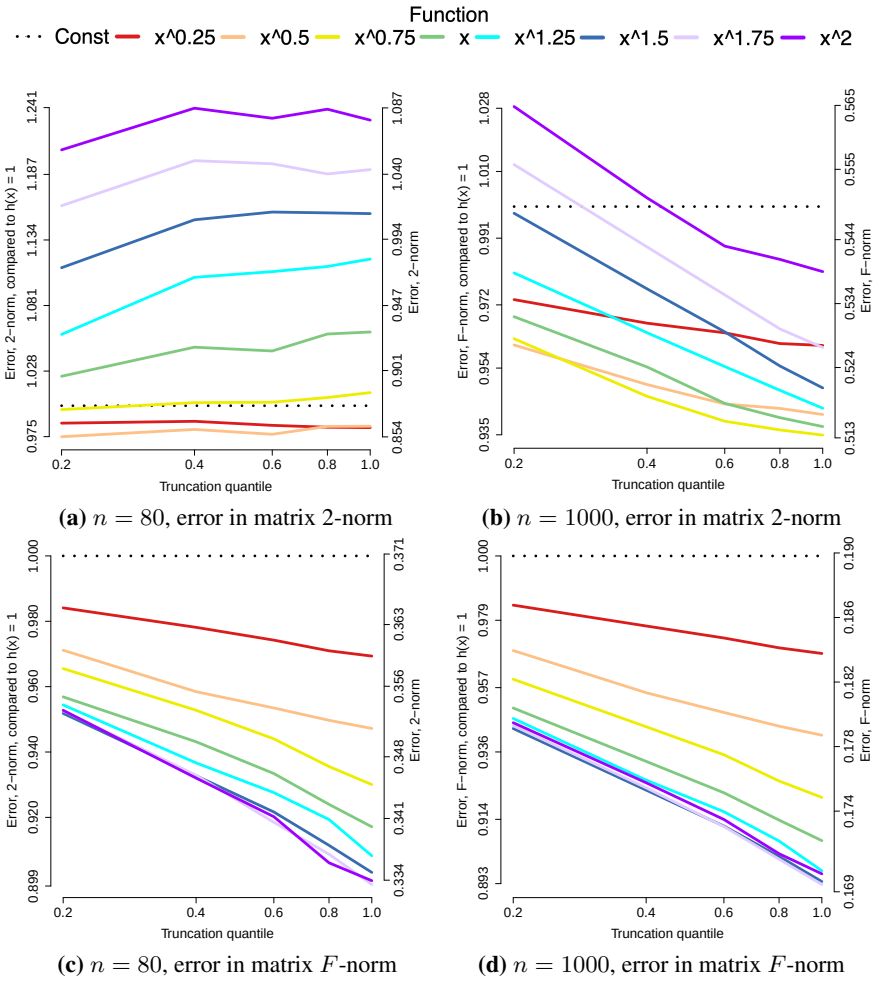


(f) $n = 1000$, max of 5 runs

Figure 4: AUCs averaged over 50 trials for edge recovery for the A^{m-1} models on the simplex.

benefit of using $|\mathcal{J}| > 1$ over $|\mathcal{J}| = 1$ (e.g., removing x_m only), in the second row we also present the average of the 5 AUC curves over 5 separate runs; in each run we construct $\mathbf{\Gamma}$ and $\boldsymbol{\eta}$ by removing one $j \in \mathcal{J}$ only. The third row shows the point-wise maximum of the 5 AUC curves.

In Figure 5 we plot the estimation error in spectral and Frobenius norms, i.e. $\left|\left|\left|\hat{\mathbf{K}} - \mathbf{K}_0\right|\right|\right|_2$ and $\left|\left|\left|\hat{\mathbf{K}} - \mathbf{K}_0\right|\right|\right|_F$, against the quantile probability π. The estimate is chosen by cross validation from the estimates with $|\mathcal{J}| = 5$. The y-ticks on the right-hand side are the errors, and those on the left are the errors divided by the error for $h(x) = 1$, measuring the relative performance of each method compared with Hyvärinen [17]. In contrast to Figure 4, smaller values on the y-axis indicate better performance. As in Figure 4, $h(x) = x^2$ performs the best when considering the Frobenius norm. When the error is measured in the spectral norm, $h(x) = x^2$ has the largest error for $n = 80$ but shows better improvements over other estimators as n increases.



(a) $n = 80$, error in matrix 2-norm

(b) $n = 1000$, error in matrix 2-norm

(c) $n = 80$, error in matrix F-norm

(d) $n = 1000$, error in matrix F-norm

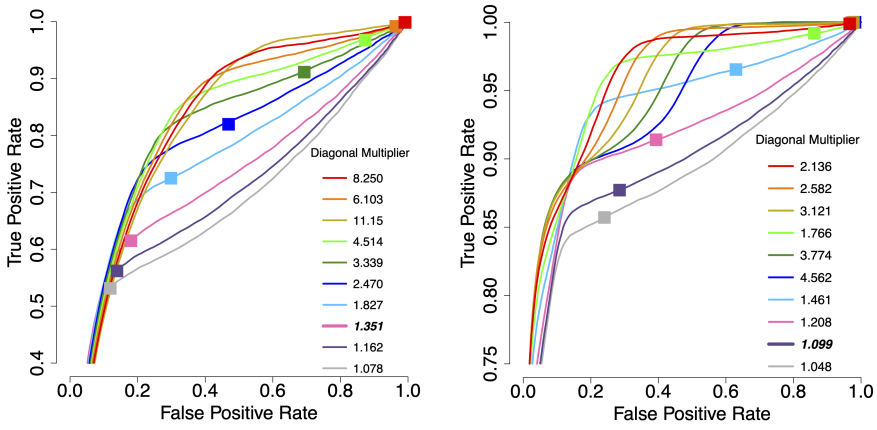
Figure 5: Averaged error in spectral and F norms over 50 trials normalized by the corresponding norms of the true $\mathbf{K}_0$; sparsity chosen by cross validation; $|\mathcal{J}| = 5$.

C.2 Effect of Diagonal Multipliers on ROCs

In our experiments, we set an upper bound on the diagonal multiplier based on our theoretical analysis in (5.4). To investigate whether the AUCs could be significantly improved with very large diagonal multipliers, in Figure 6 we present the average ROC curves over 50 trials for the solution paths with $|\mathcal{J}| = 5$ and varying diagonal multipliers. (The y axes are truncated from below for better visualization.) The upper bound diagonal multipliers (5.4), which we used throughout Section 6.2,

are highlighted in bold and italics in the legend on the right, namely 1.351 for $n = 80$ and 1.099 for $n = 1000$. The legends as well as the colors are sorted by the AUCs in decreasing order.

The results show that the AUCs reach a peak and decrease after some diagonal multiplier much larger than the theoretical upper bound. It is thus tempting to choose a very large diagonal multiplier to achieve high AUC. However, in real applications, instead of focusing on the AUC, one must choose one estimate from the solution path by cross validation and examine the performance of that estimate. For each diagonal multiplier, the square on the corresponding curve with the same color represents the TPR and FPR of the estimate picked by cross validation, averaged over 50 trials. It can be seen that the estimates for the upper bound multiplier chosen by cross validation produce the most reasonable TPRs and FPRs, with the corresponding squares much closer to the upper-left corner than those for larger diagonal multipliers.



(a) $n = 80$, ROCs **(b)** $n = 1000$, ROCs

Figure 6: ROCs averaged over 50 trials for edge recovery, with varying diagonal multipliers; $|\mathcal{J}| = 5$. Squares correspond to the average of 50 TPRs and FPRs for estimates picked by cross validation. Note that the y axes are truncated from below to better separate the curves for visualization.

D Proofs

Proof of Theorem 2. For notational simplicity, denote $\widetilde{\mathbf{K}} \equiv \mathbf{K}_1 - \mathbf{K}_2$ with columns $\widetilde{\boldsymbol{\kappa}}_{,1} \ldots, \widetilde{\boldsymbol{\kappa}}_{,m}$, and denote $\widetilde{\boldsymbol{\eta}} \equiv \boldsymbol{\eta}_1 - \boldsymbol{\eta}_2$. Assume that either $\widetilde{\mathbf{K}} \neq \mathbf{0}_{m\times m}$ or $\widetilde{\boldsymbol{\eta}} \neq \mathbf{0}_m$, otherwise there is nothing to prove. By Equation 4.1, writing $x_m \equiv 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$ and $\boldsymbol{x} = (\boldsymbol{x}_{-m}; x_m)$ and taking the gradient of the log of both sides of the equation with respect to x_j, $j = 1, \ldots, m-1$, we have

$$\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{,m}\right)^\top \boldsymbol{x}^a = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1} \tag{D.1}$$

for all $\boldsymbol{x}_{-m} \in \Delta_{-m} \equiv \left\{\boldsymbol{x}_{-m} \in \mathbb{R}_+^{m-1} \middle| \ \boldsymbol{x}_{-m} \succ \mathbf{0}, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1\right\}$. In the following when $a = 0$ by x^a we mean $\log(x)$, and by x^{a-1} we mean $1/x$ and we do not treat this case differently as the same expressions still hold.

(I) Suppose $\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{,m}\right)_{-j,-m} = \mathbf{0}_{m-2}$ for all $\boldsymbol{x}_{-m} \in \Delta_{-m}$ and $x_m = 1 - \mathbf{1}_{-m}^\top \boldsymbol{x}_{-m}$ or $\tilde{\eta}_j$.

(i) Suppose $a = 1$, then Equality D.1 becomes $\left(\widetilde{\kappa}_{j,j} - \widetilde{\kappa}_{j,m}\right) x_j + \left(\widetilde{\kappa}_{m,j} - \widetilde{\kappa}_{m,m}\right) x_m = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1}$, and we must have $b = 2$ or $b = 1$ or $\widetilde{\eta}_j = \widetilde{\eta}_m = 0$, i.e. $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$.

(ii) Suppose $a \neq 1$. The assumption implies $\left(\widetilde{\boldsymbol{\kappa}}_j\right)_{-j,-m} = \left(\widetilde{\boldsymbol{\kappa}}_m\right)_{-j,-m} = \mathbf{0}_{m-2}$. Then Equality D.1 becomes $\left(x_j^{a-1}\widetilde{\kappa}_{j,j} - x_m^{a-1}\widetilde{\kappa}_{j,m}\right) x_j^a +$

$\left(x_j^{a-1}\widetilde{\kappa}_{m,j} - x_m^{a-1}\widetilde{\kappa}_{m,m}\right) x_m^a = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1}$. Since $x_j > 0$ and $x_m > 0$ are arbitrary (as $\mathbf{1}_{-m}^\top \boldsymbol{x}_{-m}$ can vary) as long as $x_j + x_m < 1$, the cross terms must not exist, and so $\widetilde{\kappa}_{j,m} = \widetilde{\kappa}_{m,j} = 0$. It thus follows that $\tilde{\boldsymbol{\kappa}}_{-j,j} = \tilde{\boldsymbol{\kappa}}_{-m,m} = \mathbf{0}_{m-1}$ and hence $\widetilde{\mathbf{K}}$ is diagonal, and the original equality becomes $-\frac{1}{2}\mathrm{diag}(\widetilde{\mathbf{K}})^\top (\boldsymbol{x}^a)^2 + \widetilde{\boldsymbol{\eta}}^\top \boldsymbol{x}^b = 0$, in which by $\boldsymbol{x}^0$ we mean $\log(\boldsymbol{x})$. Thus we must have $2a = b \neq 0$ and $\mathbf{K}_1 - \mathbf{K}_2 = 2\boldsymbol{\eta}_1 - 2\boldsymbol{\eta}_2$.

(II) Now fix x_j and x_m such that $\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{,m}\right)_{-j,-m} \neq \mathbf{0}_{m-2}$. Note that $\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} = 1 - x_j - x_m$ is also fixed. Now the right-hand side and the first vector on the left-hand side of Equality D.1 are both constant, while $\boldsymbol{x}_{-j,-m}$ is allowed to vary freely as long as their sum is fixed. A necessary condition of Equality D.1 is thus

$$\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{,m}\right)_{-j,-m}^\top \boldsymbol{x}_{-j,-m}^a = \text{const depending on } x_j \text{ and } x_m \text{ only} \tag{D.2}$$

for all $\boldsymbol{x}_{-j,-m}^a \in \mathcal{U}_{x_j,x_m} \equiv \{\boldsymbol{y}^a : \boldsymbol{y} \succ \mathbf{0}_{m-2}, \mathbf{1}_{m-2}\boldsymbol{y} = \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} = 1 - x_j - x_m\}$.

Suppose by contradiction that $a \neq 1$. Then $\mathcal{U}_{x_j,x_m}$ is not entirely on a hyperplane, and by assumption $\left(x_j^{a-1}\widetilde{\boldsymbol{\kappa}}_{,j} - x_m^{a-1}\widetilde{\boldsymbol{\kappa}}_{,m}\right)_{-j,-m}$ is not a zero vector, so the equality cannot hold. We thus have $a = 1$, so that $\mathcal{U}_{x_j,x_m}$ lies on the hyperplane $\mathcal{H}_{x_j,x_m} \equiv \{\boldsymbol{y} : \mathbf{1}_{m-2}^\top \boldsymbol{y} = 1 - x_j - x_m\}$. Since $\mathcal{H}_{x_j,x_m} \equiv \{c\boldsymbol{x}_{-j,-m} : c \in \mathbb{R}, \boldsymbol{x} \in \mathcal{U}_{x_j,x_m}\}$, Equality D.2 must hold for all $\boldsymbol{x}_{-j,-m}$ in the hyperplane $\mathcal{H}_{x_j,x_m}$, and by the assumption that $\left(\widetilde{\boldsymbol{\kappa}}_{,j} - \widetilde{\boldsymbol{\kappa}}_{,m}\right)_{-j,-m}$ is nonzero it must be a constant multiple of $\mathbf{1}_m$, and the right-hand side of Equality D.2 is hence $c_0(1 - x_j - x_m)$ for some absolute constant $c_0 \neq 0$ assuming $\left(\widetilde{\boldsymbol{\kappa}}_{,j} - \widetilde{\boldsymbol{\kappa}}_{,m}\right)_{-j,-m} = c_0\mathbf{1}_m$. Plugging this back in Equality D.1 we get

$$c_0(1 - x_j - x_m) + \left(\widetilde{\kappa}_{j,j} - \widetilde{\kappa}_{j,m}\right) x_j + \left(\widetilde{\kappa}_{m,j} - \widetilde{\kappa}_{m,m}\right) x_m = \widetilde{\eta}_j x_j^{b-1} - \widetilde{\eta}_m x_m^{b-1},$$

and hence as in (I) (i) we have $b = 2$ or $b = 1$ or $\boldsymbol{\eta}_1 = \boldsymbol{\eta}_2$.

□

Proof of Theorem 3. Fix $j = 1, \ldots, m-1$ and $\boldsymbol{x}_{-j,-m} \in \mathcal{S}_{-j,\Delta_{-m}}$, i.e. $\boldsymbol{x}_{-j,-m} \in \mathbb{R}_+^{m-2}$ such that $\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} < 1$. In our discussion, for ease of notation, given x_j and $\boldsymbol{x}_{-j,-m}$, we may still write $x_m \equiv 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$, a function in x_j and $\boldsymbol{x}_{-j,-m}$, and for simplicity we may drop its dependence on x_j and $\boldsymbol{x}_{-j,-m}$. Note that $\mathcal{C}_j(\boldsymbol{x}_{-j,-m}) = \left(0, 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}\right)$.

(I) *Case $a > 0$ and $b \geq 0$:* For (A.1), we need $p_0(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \to 0$ as $x_j \searrow 0^+$ and $x_j \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$. As x_j goes to any finite constant, by Equation 3.1 $p_0(\boldsymbol{x}_{-m})$ converges to a non-zero constant when $b > 0$, or a finite constant times the limit of $x_j^{\eta_j} x_m(x_j)^{\eta_m}$ when $b = 0$. Note that

$$\begin{aligned}\partial_j \log p(\boldsymbol{x}_{-m}) = &-\left(\boldsymbol{\kappa}_{-m,j}^\top \boldsymbol{x}_{-m}^a\right) x_j^{a-1} + \left(\boldsymbol{\kappa}_{-m,m}^\top \boldsymbol{x}_{-m}^a\right) x_m^{a-1}(x_j) \\ &- x_m^a(x_j)\kappa_{jm}x_j^{a-1} + x_m^{2a-1}(x_j)\kappa_{mm} + \eta_j x_j^{b-1} - \eta_m x_m^{b-1}(x_j).\end{aligned}$$

i) If $b > 0$, by arguments above we only consider $(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m})$.

   a) As $x_j \searrow 0^+$, $x_m \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} > 0$, so $\partial_j \log p(\boldsymbol{x}_{-m}) = \mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}(1)$. Thus we need $\alpha_j > \max\{0, 1-a, 1-b\}$ so that $(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m}) \to 0$.
   b) The case where $x_j \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$ and $x_m \searrow 0^+$ is an analog of a) by noting that φ_j is symmetric in x_j about the midpoint of its domain $(1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m})/2$.

ii) If $b = 0$, we need $x_j^{\eta_j} x_m^{\eta_m}(x_j)(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})\partial_j \log p(\boldsymbol{x}_{-m}) \to 0$. Note that this quantity has the same form as in a) just with η_j or η_m added to the a and b ($= 0$) in the exponents, we thus require $\alpha_j > \max\{0, 1 - a - \eta_j, 1 - a - \eta_m, 1 - \eta_j, 1 - \eta_m\} = \max\{0, 1 - \eta_j, 1 - \eta_m\}$.

In conclusion, (A.1) requires $\alpha_j \geq \max\{0, 1-a, 1-b\}$ for $b > 0$ or $\alpha_j > \max\{0, 1-\eta_j\}$. For (A.2), we only prove the first integrability condition, since the second integrability condition is similar. For the first, we need to show that

$$\int_{\boldsymbol{x}_{-m} \succ \mathbf{0},\, \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1} p_0(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \left(\partial_j \log p(\boldsymbol{x}_{-m})\right)^2 \mathrm{d}\boldsymbol{x}_{-m} < +\infty.$$

Using the fact that $\mathbf{0} \prec \boldsymbol{x}_{-m} \prec \mathbf{1}$ and $0 < x_j^a < 1$, $0 < x_m^a < 1$ with the triangle inequality multiple times, we have

$$\begin{aligned}
&|\partial_j \log p(\boldsymbol{x}_{-m})| \\
&\leq \sum_{i=1}^{m-1} \left(|\kappa_{ij}| x_j^{a-1} + |\kappa_{im}| x_m^{a-1}\right) + |\kappa_{jm}| x_j^{a-1} + x_m^{a-1}|\kappa_{mm}| + |\eta_j| x_j^{b-1} + |\eta_m| x_m^{b-1} \\
&\leq \|\!|\mathbf{K}|\!\|_1 x_j^{a-1} + \|\!|\mathbf{K}|\!\|_1 x_m^{a-1} + |\eta_j| x_j^{b-1} + |\eta_m| x_m^{b-1},
\end{aligned}$$

where $\|\!|\mathbf{K}|\!\|_1 \equiv \max_{j=1,\ldots,m} \sum_{i=1}^m |\kappa_{ij}|$. We again consider the following two cases.

i) If $b > 0$, $p_0(\boldsymbol{x}_{-m})$ is bounded by an absolute constant, which we therefore ignore. We first fix $\boldsymbol{x}_{-j,-m}$ and denote $y_j(\boldsymbol{x}_{-j,-m}) \equiv 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} = x_j + x_m$. Then, writing $h_{\varphi,j,\boldsymbol{x}} \equiv (h_j \circ \varphi_j)(\boldsymbol{x}_{-m})$,

$$\begin{aligned}
&\int_0^{y_j} h_{\varphi,j,\boldsymbol{x}} (\partial_j \log p(\boldsymbol{x}_{-m}))^2 \mathrm{d}x_j \\
&\leq \int_0^{y_j} h_{\varphi,j,\boldsymbol{x}} \left[\|\!|\mathbf{K}|\!\|_1 \left(x_j^{a-1} + x_m^{a-1}(x_j)\right) + |\eta_j| x_j^{b-1} + |\eta_m| x_m^{b-1}(x_j)\right]^2 \mathrm{d}x_j \\
&\leq \int_0^{y_j/2} h_{\varphi,j,\boldsymbol{x}} \left(\|\!|\mathbf{K}|\!\|_1 x_j^{a-1} + |\eta_j| x_j^{b-1} + c_{1,m}(\boldsymbol{x}_{-j,-m})\right)^2 \mathrm{d}x_j \\
&\qquad + \int_{y_j/2}^{y_j} h_{\varphi,j,\boldsymbol{x}} \left(\|\!|\mathbf{K}|\!\|_1 x_m^{a-1}(x_j) + |\eta_m| x_m^{b-1}(x_j) + c_{1,j}(\boldsymbol{x}_{-j,-m})\right)^2 \mathrm{d}x_j \\
&= \int_0^{y_j/2} h_j(x_j) \left(\|\!|\mathbf{K}|\!\|_1 x_j^{a-1} + |\eta_j| x_j^{b-1} + c_{1,m}(\boldsymbol{x}_{-j,-m})\right)^2 \mathrm{d}x_j \\
&\qquad + \int_0^{y_j/2} h_j(x_j) \left(\|\!|\mathbf{K}|\!\|_1 x_j^{a-1} + |\eta_m| x_j^{b-1} + c_{1,j}(\boldsymbol{x}_{-j,-m})\right)^2 \mathrm{d}x_j
\end{aligned}$$

where in the last step we used change of variable $x_j \leftarrow x_m(x_j) = y_j - x_j$ for the second term, and where

$$0 < c_{1,j} \equiv \max_{y_j/2 \leq x_j \leq y_j} \left(\|\!|\mathbf{K}|\!\|_1 x_j^{a-1} + |\eta_j| x_j^{b-1}\right) = \mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1) < +\infty,$$

with $\mathcal{O}$ depending on $\mathbf{K}$ and $\boldsymbol{\eta}$. We thus have (dropping the dependency $y_j \equiv y_j(\boldsymbol{x}_{-j,-m})$ to save space)

$$\begin{aligned}
&\int_{\substack{\boldsymbol{x}_{-m} \succ \mathbf{0}, \\ \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1}} h_{\varphi,j,\boldsymbol{x}} \left(\partial_j \log p(\boldsymbol{x}_{-m})\right)^2 \mathrm{d}\boldsymbol{x}_{-m} \\
&= \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \int_0^{y_j} h_{\varphi,j,\boldsymbol{x}} (\partial_j \log p(\boldsymbol{x}_{-m}))^2 \mathrm{d}x_j \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\leq \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \int_0^{y_j/2} h_j(x_j) \times \\
&\qquad \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1)\right)^2 \mathrm{d}x_j \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\leq \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \int_0^{y_j/2} h_j(x_j) \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}(1)\right)^2 \mathrm{d}x_j \mathrm{d}\boldsymbol{x}_{-j,-m}
\end{aligned}$$

$$
\begin{aligned}
&+ \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \int_0^{y_j/2} h_j(x_j) \left(\mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \mathrm{d}\boldsymbol{x}_{-j,-m} \\
\leq & \int_{\mathbf{1} \succ \boldsymbol{x}_{-j,-m} \succ \mathbf{0}} \int_0^1 x_j^{\alpha_j} \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\quad + \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \frac{(y_j/2)^{\alpha_j+1}}{\alpha_j + 1} \left(\mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{b-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}\boldsymbol{x}_{-j,-m} \\
\leq & \int_0^1 \mathcal{O}\left(x_j^{2a-2+\alpha_j}\right) + \mathcal{O}\left(x_j^{2b-2+\alpha_j}\right) + \mathcal{O}\left(x_j^{\alpha_j}\right) \mathrm{d}x_j \\
&\quad + \sum_{\substack{p \in \{2a-1+\alpha_j, \\ 2b-1+\alpha_j,\, \alpha_j+1\}}} \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} < 1}} \mathcal{O}\left(\left(1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}\right)^p \right) \mathrm{d}\boldsymbol{x}_{-j,-m} \\
= & \int_0^1 o\left(x_j^{a-1}\right) + o\left(x_j^{b-1}\right) + o\left(x^0\right) \mathrm{d}x_j + \sum_{p \in \{a,b,1\}} \mathcal{O}\left(\Gamma(p+1)/\Gamma(p+m-1)\right) < +\infty
\end{aligned}
$$

for $\alpha_j > \max\{0, 1-a, 1-b\}$, where the second term of the last quantity follows from the normalizing constant of the Dirichlet distribution with parameters $(\mathbf{1}_{m-2}, p+1)$.

ii) If $b = 0$, then $p_0(\boldsymbol{x}_{-m})$ is bounded by $C_2 \prod_{j=1}^m x_j^{\eta_{0j}}$, where C_2 is the product of the inverse normalizing constant of $p_0(\boldsymbol{x}_{-m})$ and the supremum $\sup_{\boldsymbol{x} \succ \mathbf{0}, \mathbf{1}^\top \boldsymbol{x} = 1} \exp\left(-{\boldsymbol{x}^a}^\top \mathbf{K}_0 \boldsymbol{x}^a / (2a)\right)$, a positive and finite constant. Then by the same reasoning as in i), with $y_j(\boldsymbol{x}_{-j,-m}) \equiv 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j.-m}$ and noting that $\boldsymbol{\eta} \succ -\mathbf{1}_m$,

$$
\begin{aligned}
&\int_{\substack{\boldsymbol{x}_{-m} \succ \mathbf{0}, \\ \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m} < 1}} p_0(\boldsymbol{x}_{-m}) (h_j \circ \varphi_j)(\boldsymbol{x}_{-m}) \left(\partial_j \log p(\boldsymbol{x}_{-m})\right)^2 \mathrm{d}\boldsymbol{x}_{-m} \\
\leq & \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \int_0^{y_j/2} C_2 \prod_{k=1}^m x_k^{\eta_{0k}} h_j(x_j) \left(\mathcal{O}\left(x_j^{a-1}\right) + \mathcal{O}\left(x_j^{-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \mathrm{d}\boldsymbol{x}_{-j,-m} \\
&\quad + \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \int_0^{y_j/2} C_2 \prod_{k=1}^m x_k^{\eta_{0k}} h_j(x_j) \left(\mathcal{O}\left(y_j^{a-1}\right) + \mathcal{O}\left(y_j^{-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \mathrm{d}\boldsymbol{x}_{-j,-m} \\
\leq & C_2 \prod_{k \neq j,m} \int_0^1 x_k^{\eta_{0k}} \mathrm{d}x_k \int_0^1 x_j^{\eta_{0k}+\alpha_j} \left(\mathcal{O}\left(x_j^{-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}x_j \\
&\quad + C_2 \int_{\boldsymbol{x}_{-j,-m} \succ \mathbf{0},\, y_j > 0} \frac{(y_j/2)^{\alpha_j+1+\eta_{0j}}}{\alpha_j + 1 + \eta_{0j}} \times \prod_{k \neq j,m} x_k^{\eta_{0k}} \left(\mathcal{O}\left(y_j^{-1}\right) + \mathcal{O}(1) \right)^2 \mathrm{d}\boldsymbol{x}_{-j,-m} \\
\leq & C_2 \prod_{k \neq j,m} \frac{1}{\eta_{0k}+1} \int_0^1 \mathcal{O}\left(x_j^{-1}\right) + \mathcal{O}\left(x_j\right) \mathrm{d}x_j \\
&\quad + \sum_{p \in \{0,2\}} C_2 \int_{\substack{\boldsymbol{x}_{-j,-m} \succ \mathbf{0}, \\ y_j > 0}} \prod_{k \neq j,m} x_k^{\eta_{0k}} \mathcal{O}\left(y_j^p\right) \mathrm{d}\boldsymbol{x}_{-j,-m} \\
< & +\infty
\end{aligned}
$$

because the integral in the second term is the inverse normalizing constant of the Dirichlet distribution with parameters $(\boldsymbol{\eta}_{0,-j,-m} + \mathbf{1}_{m-2}, p+1)$, i.e. $\frac{\Gamma(p+1) \prod_{k \neq j,m} \Gamma(\eta_{0k}+1)}{\Gamma\left(\mathbf{1}_{m-2}^\top \boldsymbol{\eta}_{0,-j,-m} + p + m - 1\right)} < +\infty$.

This ends the proof for the first integrability condition for (A.1) for $a > 0$. For the second half, the integrand we consider is

$$
p_0(\boldsymbol{x}_{-m}) \left| \partial_j \left(\partial_j \log p(\boldsymbol{x}_{-m}) (h_j \circ \varphi)(\boldsymbol{x}_{-m}) \right) \right|.
$$

The arguments are similar to those for the first condition, where we first bound $\partial_{jj} \log p(\boldsymbol{x})$ using sums of products of powers of $\boldsymbol{x}$. Then for each fixed $\boldsymbol{x}_{-j,-m}$ we split the domain of x_j into two

halves and deal with the potential singularity at $x_j \searrow 0^+$, where one can show that the requirement on α_j is just enough for the integrand to be $o\left(x_j^{-1}\right)$ and thus the integral is finite. The detailed proof is tedious and is omitted.

(II) *Case $a = 0$ and $b \geq 0$:* First consider $b = 0$. We again write $y_j \equiv 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$. Fixing $\boldsymbol{x}_{-j,-m} \in \mathcal{S}_{-j,\Delta_{-m}}$,

$$
\begin{aligned}
& p_0(\boldsymbol{x}_{-m}) \left|\partial_j \log p(\boldsymbol{x}_{-m})\right| \\
\propto & \exp\left[-\frac{1}{2}\log(\boldsymbol{x}_{-m})^\top \mathbf{K}_{0,-m,-m} \log(\boldsymbol{x}_{-m}) - \log(\boldsymbol{x}_{-m})^\top \boldsymbol{\kappa}_{0,-m,m} \log x_m \right. \\
& \qquad \left. -\frac{1}{2}\kappa_{0,m,m} \left(\log x_m\right)^2 + \boldsymbol{\eta}_{0,-m}^\top \log(\boldsymbol{x}_{-m}) + \eta_{0,m} \log x_m\right] \times \\
& \left|-\boldsymbol{\kappa}_{-m,j}^\top \log(\boldsymbol{x}_{-m})/x_j + \boldsymbol{\kappa}_{-m,m}^\top \log(\boldsymbol{x}_{-m})/x_m - \kappa_{jm} \log x_m/x_j \right. \\
& \qquad \left. +\kappa_{mm} \log x_m/x_m + \eta_j/x_j - \eta_m/x_m\right| . \\
\leq & \prod_{k \neq j,m} \exp\left[-\frac{N_{\mathbf{K}_{0,-m,-m}}}{2}(\log x_k)^2 + \eta_{0,k} \log x_k\right] \exp\left[-\frac{N_{\mathbf{K}_{0,-m,-m}}}{2}(\log x_j)^2 + \right. \\
& \qquad \left. \eta_{0,j} \log x_j - \log(\boldsymbol{x}_{-m})^\top \boldsymbol{\kappa}_{0,-m,m} \log x_m - \frac{\kappa_{0,m,m} \left(\log x_m\right)^2}{2} + \eta_{0,m} \log x_m\right] \\
& \qquad \times \left|-\boldsymbol{\kappa}_{-m,j}^\top \log(\boldsymbol{x}_{-m})/x_j + \boldsymbol{\kappa}_{-m,m}^\top \log(\boldsymbol{x}_{-m})/x_m - \kappa_{jm} \log x_m/x_j \right. \\
& \qquad \left. +\kappa_{mm} \log x_m/x_m + \eta_j/x_j - \eta_m/x_m\right| .
\end{aligned}
$$

which is $\mathcal{O}\left(\exp\left(\mathcal{O}\left((\log x_j)^2\right) + \mathcal{O}\left(\log x_j\right) + \mathcal{O}\left(\log\log x_j\right)\right)\right)$ as $x_j \searrow 0^+$. Since the coefficient on the leading term is negative the entire term goes to 0. By symmetry the quantity goes to zero also when $x_j \nearrow y_j^-$. Thus, (A.1) holds for any $\alpha_j \geq 0$.

Similarly, $p_0(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})(\partial_j \log p(\boldsymbol{x}_{-m}))^2$ and $p_0(\boldsymbol{x}_{-m}) \left|\partial\left(\partial_j \log p(\boldsymbol{x}_{-m})(h_j \circ \varphi)(\boldsymbol{x}_{-m})\right)\right|$ are $\prod_{j=1}^m \mathcal{O}\left(\exp\left(-(\log x_j)^2\right)\right)$ times a polynomial, and are thus bounded and go to 0 at the boundaries of Δ_{-m}. Thus extending the integrands to 0 at the boundaries, they are continuous and bounded in the compact $\overline{\Delta_{-m}}$, so integrals $\int_{\Delta_{-m}} p_0(\boldsymbol{x}_{-m})(h_j \circ \varphi_j)(\boldsymbol{x}_{-m})(\partial_j \log p(\boldsymbol{x}_{-m}))^2 \mathrm{d}\boldsymbol{x}_{-m}$ and $\int_{\Delta_{-m}} p_0(\boldsymbol{x}_{-m}) \left|\partial\left(\partial_j \log p(\boldsymbol{x}_{-m})(h_j \circ \varphi)(\boldsymbol{x}_{-m})\right)\right| \mathrm{d}\boldsymbol{x}_{-m}$ are finite, thus proving (A.2).

For $a = 0$ and $b > 0$, $\boldsymbol{\eta} \preceq \mathbf{0}$, and the proof is similar and is omitted. In particular, $p_0(\boldsymbol{x}_{-m})$ is bounded by that with $a = 0$, $b = 0$, $\boldsymbol{\eta} \equiv \mathbf{0}_m$, and thus its product with any polynomial is bounded and goes to 0 at the boundary of $\overline{\Delta_{-m}}$. □

Proof of Theorem 4. It suffices to bound $\boldsymbol{\Gamma}$ and $\boldsymbol{g}$ using their forms in Section 4 and apply Theorem 1 in Lin et al. [22]. We first bound $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_m^{p_m}$ with $h_j(x) = x^{\alpha_j}$, $\alpha_j \geq \max\{0, -p_j, -p_m, -p_j - p_m\}$, $p_j, p_m \in \mathbb{R}$, and $0 < x_j + x_m < 1$. By the definition of $\boldsymbol{\varphi}$ on simplices, $\varphi_j(\boldsymbol{x}) = \min\{C_j, x_j, x_m\}$, so $(h_j \circ \varphi_j)(\boldsymbol{x}) x_j^{p_j} x_m^{p_m} = \min\{C_j, x_j, x_m\}^{\alpha_j} x_j^{p_j} x_m^{p_m}$ and is tightly lower bounded by 0. Noting that $\min\{x_j, x_m\} < 1/2$, we consider the following cases.

(I) If $C_j < \min\{x_j, x_m\}$, then $C_j < 1/2$ and the quantity is $C_j^{\alpha_j} x_j^{p_j} x_m^{p_m} \leq C_j^{\alpha_j + (p_j)_- + (p_m)_-} < 2^{-\alpha_j - (p_j)_- - (p_m)_-}$ where $(y)_- = y$ if $y < 0$ and 0 otherwise.

(II) Otherwise suppose $x_j \leq x_m$ and $x_j \leq C_j$, then the quantity is equal to $x_j^{\alpha_j + p_j} x_m^{p_m}$, which is upper bounded by $x_j^{\alpha_j + p_j + p_m} < 2^{-\alpha_j - p_j - p_m} < 1$ if $p_m \leq 0$; if $p_m > 0$ it is upper bounded by $((\alpha_j + p_j)/(\alpha_j + p_j + p_m))^{\alpha_j + p_j} (p_m/(\alpha_j + p_j + p_m))^{p_m}$ if $(\alpha_j + p_j)/(\alpha_j + p_j + p_m) \leq 1/2$ or by $2^{-\alpha_j - p_j - p_m}$ otherwise. Note that the statement for $p_m > 0$ covers the one for $p_m \leq 0$. The conclusion for $x_m \leq x_j$ and $x_m \leq C_j$ follows by symmetry, and note that at most one of $(\alpha_j + p_j)/(\alpha_j + p_j + p_m) \leq 1/2$ and $(\alpha_j + p_m)/(\alpha_j + p_j + p_m) \leq 1/2$ can hold.

In conclusion, defining

$$\zeta_2(\alpha_j,p_j,p_m)=\begin{cases}\left(\frac{\alpha_j+p_j}{\alpha_j+p_j+p_m}\right)^{\alpha_j+p_j}\left(\frac{p_m}{\alpha_j+p_j+p_m}\right)^{p_m}, & \text{if } p_m\geq\alpha_j+p_j,\\ \left(\frac{\alpha_j+p_m}{\alpha_j+p_j+p_m}\right)^{\alpha_j+p_m}\left(\frac{p_j}{\alpha_j+p_j+p_m}\right)^{p_j}, & \text{if } p_j\geq\alpha_j+p_m,\\ 2^{-\alpha_j-p_j-p_m}, & \text{otherwise,}\end{cases}$$

we have $(h_j\circ\varphi_j)(\boldsymbol{x})x_j^{p_j}x_m^{p_m}\leq\zeta_2(\alpha_j,p_j,p_m)<1$. Similarly, $\partial_j(h_j\circ\varphi_j)(\boldsymbol{x})x_j^{p_j}x_m^{p_m}\leq\alpha_j\zeta_2(\alpha_j-1,p_j,p_m)<\alpha_j$, if $\alpha_j-1\geq\max\{0,-p_j,-p_m,-p_j-p_m\}$.

Then for all j,k,ℓ, as long as $\alpha_j\geq\max\{1,2-a,2-b\}$, we have $0\leq\gamma_{j,k,\ell}<1$, and similarly $0\leq g_{j,k}<\max_{j=1,\ldots,m}\alpha_j+\max\{|a-1|+2a,|b-1|\}$. The rest follows from the same proof as Theorem 5.3 of Yu et al. [15].

Note that using the form of $\boldsymbol{\Gamma}$ in Section 4, a tighter bound for $\gamma_{j,k,\ell}$ is

$$\begin{aligned}\max_{j,k=1,\ldots,m}\max\{&\zeta_2(\alpha_j,2a-2,0),\zeta_2(\alpha_j,4a-2,0),\zeta_2(\alpha_j,2a-2,2a),\\&\zeta_2(\alpha_j,2b-2,0),\zeta_2(\alpha_j,0,2a-2),\zeta_2(\alpha_j,2a,2a-2),\\&\zeta_2(\alpha_j,0,4a-2),\zeta_2(\alpha_j,0,2b-2),\zeta_2(\alpha_j,a-1,a-1),\\&\zeta_2(\alpha_j,2a-1,2a-1),\zeta_2(\alpha_j,a-1,3a-1),\zeta_2(\alpha_j,3a-1,a-1),\\&\zeta_2(\alpha_j,b-1,b-1)\},\end{aligned}$$

and the one for $g_{j,k}$ can be similarly written in terms of $\zeta_2(\alpha_j,\cdot,\cdot)$ and $\alpha_j\zeta_2(\alpha_j-1,\cdot,\cdot)$. □

Proof of Theorem 6. Write $\iota_{+m}(\boldsymbol{x}_{-m})=\left(\boldsymbol{x}_{-m},1-\mathbf{1}_{m-1}^\top\boldsymbol{x}_{-m}\right)$ for any $\boldsymbol{x}_{-m}\in\Delta_{-m}$. We first prove the finiteness of the normalizing constant. If $\mathbf{K}$ is positive definite, the inverse normalizing constant is

$$\begin{aligned}&\int_{\Delta_{-m}}\exp\left(-\frac{1}{2}\log\iota_{+m}(\boldsymbol{x}_{-m})^\top\mathbf{K}\log\iota_{+m}(\boldsymbol{x}_{-m})+\boldsymbol{\eta}^\top\log\iota_{+m}(\boldsymbol{x}_{-m})\right)\mathrm{d}\boldsymbol{x}_{-m}\\ \leq&\int_{\Delta_{-m}}\exp\left(\sum_{j=1}^m\left(-\lambda_{\min}(\mathbf{K})\left(\log\iota_{+m}\left(\boldsymbol{x}_{-m}\right)\right)_j^2+\eta_j\left(\log\iota_{+m}\left(\boldsymbol{x}_{-m}\right)\right)_j\right)\right)\mathrm{d}\boldsymbol{x}_{-m}\\ \leq&\int_{\Delta_{-m}}\exp\left(\sum_{j=1}^m\frac{\eta_j^2}{4\lambda_{\min}(\mathbf{K})}\right)\mathrm{d}\boldsymbol{x}_{-m}<+\infty,\end{aligned}$$

proving (I). Now assume $\mathbf{K}$ is no longer positive definite. If $\mathbf{K}\mathbf{1}_m=\mathbf{0}_m$, for any $\boldsymbol{a}\in\mathbb{R}^m$,

$$\begin{aligned}\boldsymbol{a}^\top\mathbf{K}\boldsymbol{a}&=(\boldsymbol{a}_{-j},a_j)^\top\begin{bmatrix}\mathbf{K}_{-j,-j}&\boldsymbol{\kappa}_{-j,j}\\\boldsymbol{\kappa}_{-j,j}^\top&\kappa_{jj}\end{bmatrix}(\boldsymbol{a}_{-j},a_j)\\&=(\boldsymbol{a}_{-j},a_j)^\top\begin{bmatrix}\mathbf{K}_{-j,-j}&-\mathbf{K}_{-j,-j}\mathbf{1}_{m-1}\\-\mathbf{1}_{m-1}^\top\mathbf{K}_{-j,-j}&\mathbf{1}_{m-1}^\top\mathbf{K}_{-j,-j}\mathbf{1}_{m-1}\end{bmatrix}(\boldsymbol{a}_{-j},a_j)\\&=\left(\boldsymbol{a}_{-j}-a_j\mathbf{1}_m\right)^\top\mathbf{K}_{-j,-j}\left(\boldsymbol{a}_{-j}-a_j\mathbf{1}_{m-1}\right),\end{aligned}$$

which is zero if and only if $\boldsymbol{a}_{-j}=a_j\mathbf{1}_{m-1}$, i.e. $a_1=\cdots=a_m$, and is positive otherwise. Thus, if $\mathbf{K}\mathbf{1}_m=\mathbf{0}_m$, the condition that $\mathbf{K}_{-j,-j}$ is positive definite for some $j=1,\ldots,m$ is equivalent to that $\mathbf{K}_{-j,-j}$ is positive definite for all $j=1,\ldots,m$, and implies that $\mathbf{K}$ is positive semi-definite.

If $\mathbf{K}$ is positive semi-definite and $\boldsymbol{\eta}\succ-\mathbf{1}_m$, the inverse normalizing constant is

$$\begin{aligned}&\int_{\Delta_{-m}}\exp\left(-\frac{1}{2}\log\iota_{+m}(\boldsymbol{x}_{-m})^\top\mathbf{K}\log\iota_{+m}(\boldsymbol{x}_{-m})+\boldsymbol{\eta}^\top\log\iota_{+m}(\boldsymbol{x}_{-m})\right)\mathrm{d}\boldsymbol{x}_{-m}\\ \leq&\int_{\Delta_{-m}}\exp\left(\boldsymbol{\eta}^\top\log\iota_{+m}(\boldsymbol{x}_{-m})\right)\mathrm{d}\boldsymbol{x}_{-m}=\frac{\prod_{j=1}^m\Gamma(\eta_j+1)}{\Gamma\left(\mathbf{1}_m^\top\boldsymbol{\eta}+m\right)}<+\infty\end{aligned}$$

since the last quantity is the inverse normalizing constant of the Dirichlet distribution with parameters $(\boldsymbol{\eta}+\mathbf{1}_m)$, proving (III).

On the other hand, suppose $\mathbf{K}\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K}_{-j,-j}$ is positive definite for some/all $j = 1, \ldots, m$. Again letting $\boldsymbol{x}_{-m}$ be the free variables and letting $x_m = 1 - \mathbf{1}_{m-1}^\top \boldsymbol{x}_{-m}$, define the *additive log-ratio transformation* applied to $\boldsymbol{x}$: $\boldsymbol{y}_{-m} \equiv \log \boldsymbol{x}_{-m} - (\log x_m)\mathbf{1}_{m-1}$, a random vector supported on $\mathbb{R}^{m-1}$. Append an extra $y_m = 0$ for ease of notation. The transformation is thus bijective and the inverse transformation, the *additive logistic transformation* $\boldsymbol{x} = \exp(\boldsymbol{y})/\mathbf{1}_m^\top \exp(\boldsymbol{y})$. Since

$$\partial x_k/\partial y_j = -x_k x_j, \quad \partial x_j/\partial y_j = x_j(1 - x_j)$$

for $j \neq k$, $j, k = l, \ldots, m-1$, we have

$$\begin{aligned}\left|\frac{\partial \boldsymbol{x}_{-m}}{\partial \boldsymbol{y}_{-m}}\right| &= \begin{vmatrix} x_1(1-x_1) & -x_1x_2 & \cdots & -x_1x_{m-1} \\ -x_1x_2 & x_2(1-x_2) & \cdots & -x_2x_{m-1} \\ \vdots & \vdots & \ddots & \vdots \\ -x_1x_{m-1} & -x_2x_{m-1} & \cdots & x_{m-1}(1-x_{m-1}) \end{vmatrix} \\ &= \prod_{j=1}^m x_j = \exp\left(\mathbf{1}_m^\top \log \boldsymbol{x}\right).\end{aligned}$$

Then by $\mathbf{1}_m^\top \mathbf{K} = \mathbf{K}\mathbf{1}_m = \mathbf{0}_m$, $\boldsymbol{y}_{-m}$ has density proportional to

$$\begin{aligned}&p(\boldsymbol{x}_{-m})\,|\partial \boldsymbol{x}_{-m}/\partial \boldsymbol{y}_{-m}| \\ &\propto \exp\Bigg(-\frac{1}{2}\left(\log \boldsymbol{x} - (\log x_m)\mathbf{1}_m\right)^\top \mathbf{K}\left(\log \boldsymbol{x} - (\log x_m)\mathbf{1}_m\right) \\ &\qquad + (\boldsymbol{\eta} + \mathbf{1}_m)^\top \left(\log \boldsymbol{x} - (\log x_m)\mathbf{1}_m\right) + (\log x_m)\mathbf{1}_m^\top(\boldsymbol{\eta} + \mathbf{1}_m)\Bigg) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{y}^\top \mathbf{K}\boldsymbol{y} + (\boldsymbol{\eta} + \mathbf{1}_m)^\top \boldsymbol{y} + \mathbf{1}_m^\top(\boldsymbol{\eta} + \mathbf{1}_m)\log x_m\right) \\ &= \exp\Bigg(-\frac{1}{2}\boldsymbol{y}_{-m}^\top \mathbf{K}_{-m,-m}\boldsymbol{y}_{-m} + (\boldsymbol{\eta}_{-m} + \mathbf{1}_{m-1})^\top \boldsymbol{y}_{-m} \\ &\qquad - \left(\mathbf{1}_m^\top \boldsymbol{\eta} + m\right)\log\left(1 + \mathbf{1}_{m-1}^\top \exp(\boldsymbol{y}_{-m})\right)\Bigg).\end{aligned}$$

Note that $\log x_m = -\log\left(1 + \mathbf{1}_{m-1}^\top \exp(\boldsymbol{y}_{-m})\right) < 0$, so for $\mathbf{1}_m^\top \boldsymbol{\eta} + m \geq 0$ the last display is always upper-bounded by a constant times a normal density with a positive definite inverse covariance matrix $\mathbf{K}_{-m,-m}$, and thus the normalizing constant is finite, thus proving (II).

As for (A.1), fix $j = 1, \ldots, m-1$ and any $\ell \in \{1, \ldots, m-1\}\backslash\{j\}$, and write $\boldsymbol{z} \equiv \log \boldsymbol{x} - (\log x_\ell)\mathbf{1}_m$. Fix any $\boldsymbol{x}_{-j,-m} \in \mathbb{R}^{m-2}$ with $\boldsymbol{x}_{-j,-m} \succ \mathbf{0}$, $\mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m} < 1$. Then if (I) $\mathbf{K}_0$ is positive definite or (II) $\mathbf{K}_0\mathbf{1}_m = \mathbf{0}_m$ and $\mathbf{K}_{0,-\ell,-\ell}$ is positive definite, by the proof above, $p_0(\boldsymbol{x}_{-m})x_i^t$ is upper bounded by a finite constant depending on $\mathbf{K}_0$ and $\boldsymbol{\eta}_0$ for any $t \in \mathbb{R}$ and $i = j, m$, since it is a constant times the density with parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0 + t\boldsymbol{e}_i$, and since we did not impose any restriction on the $\boldsymbol{\eta}$ parameter. On the other hand, for (III) $\mathbf{K}_0\mathbf{1}_m = \mathbf{0}_m$, $\mathbf{K}_{0,-\ell,-\ell}$ is positive semi-definite and $\boldsymbol{\eta}_0 \succ -\mathbf{1}_m$,

$$\begin{aligned}p_0(\boldsymbol{x}_{-m}) &\propto \exp\left(-\frac{1}{2}\log \boldsymbol{x}^\top \mathbf{K}_0 \log \boldsymbol{x} + \boldsymbol{\eta}_0^\top \log \boldsymbol{x}\right) \\ &= \exp\left(-\frac{1}{2}\boldsymbol{z}_{-\ell}^\top \mathbf{K}_{0,-\ell,-\ell}\boldsymbol{z}_{-\ell} + \boldsymbol{\eta}_{0,-\ell}^\top \boldsymbol{z}_{-\ell} + \left(\mathbf{1}_m^\top \boldsymbol{\eta}_0\right)\log x_\ell\right) \\ &\leq \exp\left(\boldsymbol{\eta}_{0,-\ell}^\top \boldsymbol{z}_{-\ell} + \left(\mathbf{1}_m^\top \boldsymbol{\eta}_0\right)\log x_\ell\right) \\ &\propto \exp\left(\eta_{0,j} z_j + \eta_{0,m} z_m\right).\end{aligned}$$

On the other hand,

$$\begin{aligned}&|\partial_j \log p(\boldsymbol{x}_{-m})|\,\min\{x_j, x_m\}^{\alpha_j} \\ &= \left|-\boldsymbol{\kappa}_{,j}^\top \log \boldsymbol{x}/x_j + \boldsymbol{\kappa}_{,m}^\top \log \boldsymbol{x}/x_m + \eta_j/x_j - \eta_m/x_m\right| \min\{x_j, x_m\}^{\alpha_j} \\ &= \left(\left|\boldsymbol{\kappa}_{,j}^\top \log \boldsymbol{x}/x_j\right| + \left|\boldsymbol{\kappa}_{,m}^\top \log \boldsymbol{x}/x_m\right| + |\eta_j/x_j| + |\eta_m/x_m|\right)\min\{x_j, x_m\}^{\alpha_j} \\ &\leq \left(\left|\boldsymbol{\kappa}_{,j}^\top \log \boldsymbol{x}\right| + \left|\boldsymbol{\kappa}_{,m}^\top \log \boldsymbol{x}\right| + |\eta_j| + |\eta_m|\right)\min\{x_j, x_m\}^{\alpha_j - 1}.\end{aligned}$$

Thus, as $x_j \searrow 0^+$ or $x_m \searrow 0^+$ (i.e. $x_j \nearrow 1 - \mathbf{1}_{m-2}^\top \boldsymbol{x}_{-j,-m}$), by multiplying the two bounds we have $p_0(\boldsymbol{x}_{-m}) \left|\partial_j \log p(\boldsymbol{x}_{-m})\right| (h_j \circ \varphi_j)(\boldsymbol{x}) \searrow 0^+$ for any α_j for (I) and (II) (by letting t to e.g. $\alpha_j - 2$ in the discussion above), or for (III) by a constant times

$$\left(\left|\boldsymbol{\kappa}_{,j}^\top \log \boldsymbol{x}\right| + \left|\boldsymbol{\kappa}_{,m}^\top \log \boldsymbol{x}\right| + |\eta_j| + |\eta_m|\right) \min\{x_j, x_m\}^{\alpha_j - 1} x_j^{\eta_{0,j}} x_m^{\eta_{0,m}} \searrow 0^+$$

if $\alpha_j > \max\{1 - \eta_{0,j}, 1 - \eta_{0,m}\}$.

As for (A.2), the results follow by a similar discussion for the Gamma model (a-b model with $b = 0$) on the standard simplex in Section 3. □

Proof of Lemma 5. For $a > 0$ or $b > 0$, the proof of Lemma 5.1 of Yu et al. [15] works even for the simplex domain. We thus only consider the case where $a = b = 0$, for which we show that the moment-generating function of $\log X_j$ is finite and invoking Theorem 2.13 in Wainwright [27]. According to Theorem 6, assume

(I) $\mathbf{K}_0$ is positive definite, or

(II) $\mathbf{K}_0 \mathbf{1}_m = \mathbf{0}$, $\mathbf{K}_{0,-k,-k}$ is positive definite for some $k = 1, \ldots, m$, and $\mathbf{1}_m^\top \boldsymbol{\eta} + m \geq 0$, or

(III) $\mathbf{K}_0 \mathbf{1}_m = \mathbf{0}$, $\mathbf{K}_0$ is positive semi-definite, and $\boldsymbol{\eta} \succ -\mathbf{1}_m$.

For any suitable t, $\mathbb{E}_0 \exp(t \log X_j)$ is the inverse normalizing constant for the model with parameters $\mathbf{K}_0$ and $\boldsymbol{\eta}_0 + t\boldsymbol{e}_j$, and is thus finite for (I) with $t \in \mathbb{R}$ and (III) with $t \in (-1 - \eta_{0,j}, +\infty) \ni 0$. For (II), recall that in the proof of Theorem 6 we have shown that for any $k = 1, \ldots, m$, the density of $\log \boldsymbol{X}_{-k} - (\log X_k)\, \mathbf{1}_{m-1}$ is bounded by a constant times a Gaussian density, and thus $\mathbb{E}_0 \left[X_j^t / X_k^t\right] = \mathbb{E}_0 \exp\left(t(\log X_j - \log X_k)\right) < +\infty$ for any $k = 1, \ldots, m$ and $t \in \mathbb{R}$. So for any $t < 0$,

$$\begin{aligned}
&\mathbb{E}_0 X_j^t \\
\leq\, &\mathbb{E}_0 \left[X_j^t | X_j \geq 1/m\right] \mathbb{P}\left(X_j \geq 1/m\right) + \sum_{k \neq j} \mathbb{E}_0 \left[X_j^t \middle| X_k \geq 1/m\right] \mathbb{P}\left(X_k \geq 1/m\right) \\
\leq\, &m^{-t} \mathbb{P}\left(X_j \geq 1/m\right) + \sum_{k \neq j} m^{-t} \mathbb{E}_0 \left[X_j^t / X_k^t \middle| X_k \geq 1/m\right] \mathbb{P}\left(X_k \geq 1/m\right) \\
\leq\, &m^{-t} + \sum_{k \neq j} m^{-t} \mathbb{E}_0 \left[X_j^t / X_k^t\right] < +\infty.
\end{aligned}$$

On the other hand, $\mathbb{E}_0 X_j^t \leq 1$ for $t \geq 0$. Thus, $\mathbb{E}_0 \exp(t \log X_j) < +\infty$ for any $t \in \mathbb{R}$ for (II). Hence, for all of (I)–(III) we have $\mathbb{E}_0 \exp(t \log X_j) < +\infty$ for t in a neighborhood around 0. □

Proof of Corollary 8. Let the sub-exponential norm of $\log X_j$ be

$$\|\log X_j\|_{\psi_1} \equiv \sup_{q \geq 1} (\mathbb{E}_0 |\log X_j|^q)^{1/q} / q.$$

Then by Lemma 21.6 of Yu et al. [20] or Corollary 5.17 of Vershynin [28],

$$\mathbb{P}\left(-\log X_j + \mathbb{E}_0 \log X_j \geq \epsilon_3\right) \leq \exp\left(-\min\left(\frac{\epsilon_3^2}{8e^2 \|\log X_j\|_{\psi_1}^2}, \frac{\epsilon_3}{4e \|\log X_j\|_{\psi_1}}\right)\right).$$

Let

$$\epsilon_3 \equiv \max\Big\{2\sqrt{2}e \max_j \|\log X_j\|_{\psi_1} \sqrt{\log 3 + \log n + (\tau+1)\log m}, \\ 4e \max_j \|\log X_j\|_{\psi_1} (\log 3 + \log n + (\tau+1)\log m)\Big\}.$$

Then $0 \leq -\log X_j^{(i)} \leq \max_k \mathbb{E}_0 \log X_k + \epsilon_3$ for all $j = 1, \ldots, m$ and $i = 1, \ldots, n$ with probability at least $1 - 1/(3m^\tau)$. The rest follows as in the proof of Theorem 5.3 of Yu et al. [15] and Theorem 4. □