Rational partition models under iterative proportional scaling

Jane Ivy Coons[1,2], Carlotta Langer[3], and Michael Ruddy[4]

[1]St John's College, University of Oxford, United Kingdom
[2]Mathematical Institute, University of Oxford, United Kingdom
jane.coons@maths.ox.ac.uk
[3]Hamburg University of Technology, Hamburg, Germany
carlotta.langer@tuhh.de
[4]University of San Francisco, San Francisco, United States
mruddy@usfca.edu

April 29, 2022

Abstract

In this work we investigate *partition models*, the subset of log-linear models for which one can perform the iterative proportional scaling (IPS) algorithm to numerically compute the maximum likelihood estimate (MLE). Partition models include families of models such as hierarchical models and balanced, stratified staged trees. We define sufficient conditions, the Generalized Running Intersection Property (GRIP), on the matrix representation of a general partition model for the IPS algorithm to always exactly produce the MLE in one cycle. Additionally we connect the GRIP to the toric fiber product and to previous results for hierarchical models and balanced, stratified staged trees to the GRIP. This leads to a characterization of balanced, stratified staged trees in terms of the GRIP.

1 Introduction

The iterative proportional scaling (IPS) algorithm is a simple and efficient numerical algorithm for computing the maximum likelihood estimate (MLE) for certain families of log-linear statistical models, which we call *partition models*. The IPS algorithm is widely used in survey statistics, recent examples include [9, 26, 40], and has been researched, in its more restricted form as Sinkhorn algorithm, in connection to optimal transport problems [6, 33]. Partition models include hierarchical models, which have been heavily studied in connection with the IPS algorithm [21, 25, 39]. Maximum likelihood estimation for log-linear models and its connection to algebraic and combinatorial objects is also of interest in Algebraic Statistics, including in [2], where the authors connect invariant theory and algorithms to compute the MLE including a form of the

IPS algorithm, and in recent works exploring the number of complex critical points of the log-likelihood function, also known as the ML-degree (see [1, 8, 11, 19, 28, 29, 35]).

From an information geometric perspective, calculating the MLE for a partition model can be described as projecting to linear families defined by the partitions of matrix representing the model. Given a data vector and an estimate on the model, the IPS algorithm updates this estimate each step by projecting onto a different linear family, converging towards the MLE. We say that the IPS algorithm has completed *one cycle* after it has iterated through each linear family exactly once.

In particular at each step of the IPS algorithm, the estimate is a rational function of the data vector, implying that in this case the MLE is a rational function of the data vector. Models for which the MLE can always be described as a rational function of the data vector are called *rational* and are a subject of recent interest [12, 19, 30]. This is in contrast to what is sometimes called the *generalized*[1] IPS algorithm, another numerical algorithm for computing the MLE for log-linear models, which does not produce a rational function at each step, yet can be applied to any log-linear model.

In this work we are interested in the question, "When does the IPS algorithm exactly produce the MLE after one cycle?" We first note that the matrix representation of a partition model heavily influences the outcome of the IPS algorithm (see Example 2.11). In [25] the author defines the Running Intersection Property (RIP) for hierarchical models, which gives sufficient conditions on the matrix representation of a hierarchical model so that the IPS algorithm always produces the MLE exactly in one cycle. The author shows that for decomposable hierarchical models there always exists such a representation. Drawing inspiration from the RIP, we define the Generalized Running Intersection Property (GRIP) on the matrix representations of general partition models and show that it gives sufficient conditions for the IPS algorithm to always produce the MLE exactly in one cycle.

In the case of hierarchical models we show that the RIP is a special case of the GRIP (see Prop. 4.1 and Rem. 4.2). We also connect previous work on the rationality of graphical and hierarchical models with the GRIP. In particular we point out that the existence of a matrix representation satisfying the GRIP, the model being rational, and the model being decomposable are all equivalent.

We also investigate another family of partition models: balanced and stratified staged trees. Staged tree models, also known as chain event graphs, are probability tree models that also encode conditional independence relationships among events. They have been the subject of much study from an algebraic perspective in recent years [4, 19, 24, 34]. In this work we connect the GRIP to balanced and stratified staged trees.

One can associate a tree graph to any matrix representation of a partition model. If we restrict to matrices without repeated columns then this tree graph is unique. We show

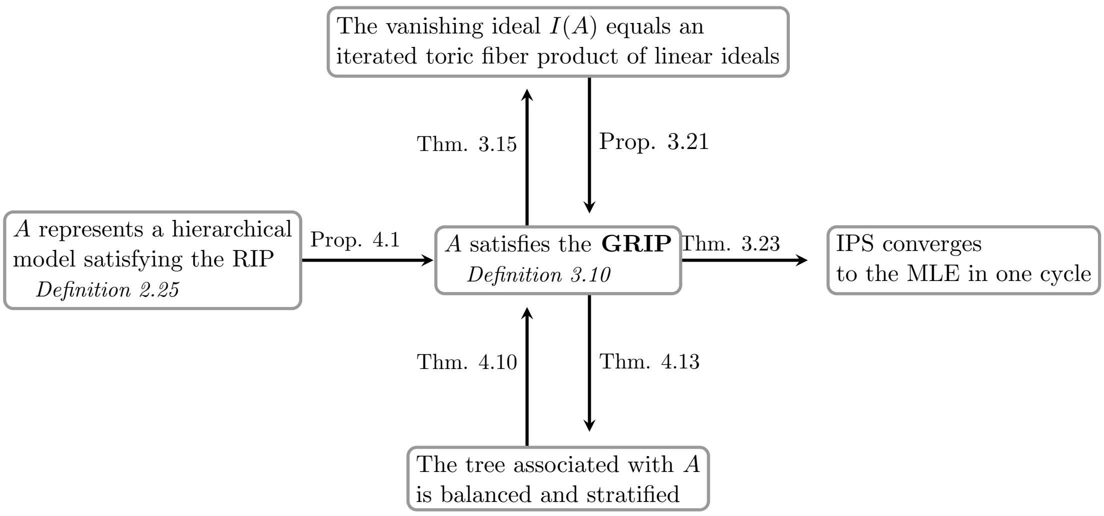
[1] Many works will refer to the generalized IPS algorithm as simply "the IPS algorithm" while referring to its predecessor, the algorithm we consider in this work, as the "classical IPS algorithm."

both that if a matrix representation satisfies the GRIP then the associated tree must be a balanced and stratified staged tree and that the unique matrix without repeated columns associated to a balanced and stratified staged tree satisfies the GRIP. This shows that, in some sense, the GRIP is equivalent to the notion of a balanced and stratified staged tree. In [20], the authors show that decomposable hierarchical models are also balanced staged tree models and claim that balanced staged trees are a natural generalization of such models. Our work also implies this result and supports this claim.

We make use of the toric geometry of a partition model in order to determine whether IPS achieves one-cycle convergence on a given parametrization of the model. In particular, we show that when a parametrization of the model satisfies the GRIP, the model can be written as a *toric fiber product* of smaller rational partition models [36]. This in turn allows us to write the MLE of the larger partition model as a normalized product of the MLEs of the smaller models [3], which facilitates our proof of one-cycle convergence.

Our work draws novel connections between the geometry of a rational partition model and the performance of the iterative proportional scaling algorithm applied to it. Moreover, it gives sufficient conditions for a partition model to be rational. If the partition model has a parametrization which satisfies the GRIP, then it is rational and one can read the MLE directly from the matrix of this parametrization (Corollary 3.19). These results highlight the significant role that model representation plays in the performance of numerical computations of MLEs.

We include a diagram of our main results below.



The work is structured as follows. In Section 2 we introduce some basic notions and previous results surrounding log-linear models, the MLE, the IPS algorithm, and some important partition model families. In particular, we define partition matrices and partition models in 2.1, describe describe the IPS algorithm that we are concerned with in 2.2, and discuss staged tree models and hierarchical models in 2.4 and 2.5 respectively. In Section 3 we introduce the GRIP and show this implies that the IPS algorithm produces the MLE after one cycle (Theorem 3.23). In Section 4.2 we investigate the connection between the GRIP and staged tree models, showing that the GRIP characterizes a subset of such models. In Section 4.1 we show that the GRIP is indeed a generalization of the RIP. We end the work with a discussion of open problems stemming from our results in Section 5.

2 Preliminaries

2.1 Log-Linear Models

In this section we introduce the class of models that the IPS algorithm can be performed on to find or estimate the maximum likelihood estimate. We start with some background on log-linear models before defining what we call *partition models*.

Consider an $n \times m$ matrix $A = (a_{ij})$ where $a_{ij} \in \mathbb{Z}_+$ and each column sum $\sum_{i=1}^{d} a_{ij}$ is equal. Then the matrix A defines a homogeneous polynomial map $\phi_A : \mathbb{R}^n \to \mathbb{R}^m$ where

$$\phi_A(t_1, \ldots, t_n) = \left(\prod_{i=1}^{n} t_i^{a_{i1}}, \prod_{i=1}^{n} t_i^{a_{i2}}, \ldots, \prod_{i=1}^{n} t_i^{a_{im}} \right). \tag{1}$$

Let $\Delta_{m-1} \subset \mathbb{R}^m$ denote the open $(m-1)$-dimensional probability simplex. The *support* of a row α of A, denoted $\mathrm{supp}(\alpha)$ is the set of column indices in which α has a nonzero entry.

Definition 2.1. The **log-linear** model associated to the integer matrix A described above, denoted $\mathcal{M}_A$, is the intersection of the Zariski closure of the image of ϕ_A with the open probability simplex; that is,

$$\mathcal{M}_A = \overline{\mathrm{Im}(\phi_A)} \cap \Delta_{m-1}.$$

$$\mathcal{M}_A = \{p \in \Delta_{m-1} | \log p \in rowspan(A)\}.$$

It is important to note the difference between a particular integer matrix A (with equal column sum) and the associated log-linear model $\mathcal{M}_A$. Many such matrices A may lead to the same log-linear model. In this work we are concerned with the relationship between the IPS algorithm and a matrix A which can be thought of as a particular representation of the log-linear model $\mathcal{M}_A$.

Note that the map ϕ_A is a monomial map. Hence the closure of $\mathcal{M}_A$ is a *toric variety*; for this reason the model $\mathcal{M}_A$ is also referred to as a *toric model* in the algebraic statistics literature. Let $I(A)$ denote the ideal of polynomials that vanishes on $\mathcal{M}_A$. The following proposition adapted from [Prop 6.2.4 in Sullivant] shows the relationship between the matrix A and the ideal of polynomials vanishing on $\mathcal{M}_A$.

Proposition 2.2. *Let $A \in \mathbb{Z}^{n\times m}$. Then the vanishing ideal of $\mathcal{M}_A$,*

$$I(A) = \langle p^u - p^v \mid u, v \in \mathbb{Z}_+^m \text{ and } Au = Av\rangle \tag{2}$$

*is the **toric ideal** of A and $\mathcal{M}_A$ is the intersection of the variety $V(I(A))$ with the open simplex Δ_{m-1}.*

Proposition 2.3. *Let $A \in \mathbb{Z}^{n\times m}$ and let $A' \in \mathbb{Z}^{n'\times m}$ be integer matrices such that* $\mathrm{rowspan}(A) = \mathrm{rowspan}(A')$. *Then $I(A) = I(A')$ and $\mathcal{M}_A = \mathcal{M}_{A'}$.*

Proof. Since $\mathrm{rowspan}(A) = \mathrm{rowspan}(A')$, the kernels of A and A' are equal. Therefore, by Equation 2, we have $I(A) = I(A')$. Furthermore, this implies that $V(I(A)) = V(I(A'))$. So the intersections of these with the open probability simplex are equal. Therefore $\mathcal{M}_A = \mathcal{M}_{A'}$ as well. □

The models we are interested in have a matrix A that can be structured in the following way. Suppose each column sum of A is equal to k, meaning $\sum_{i=1}^n a_{ij} = k$ for all $1 \le j \le n$. Then one can group the rows of A into matrices A^ℓ with column equal to one where $1 \le \ell \le k$. In each matrix A^ℓ, exactly one row of A^ℓ has a non-zero entry for each column of A and hence encodes a partition of the state space $\mathcal{X}$. For this reason we refer to matrices A^ℓ, $1 \le \ell \le k$ as the *partitions* of A.

Definition 2.4. A matrix A gives rise to a **partition model** $\mathcal{M}_A$ if the map in (1) is a homogeneous, multi-linear monomial map.

Let A^1A^2 denote the matrix obtained by stacking A^1 above A^2; that is,

$$A^1A^2 := \begin{bmatrix} A^1 \\ \hline A^2 \end{bmatrix}. \tag{3}$$

From the partitions of A we can build a new matrix, by stacking the partitions as defined above

$$A^{1,\dots,k} = A^1A^2 \dots A^k \tag{4}$$

where α_i^ℓ denotes the i-th row of the ℓ-th partition, as illustrated in Example 2.6. Although there are many ways to arrange the rows of A to build $A^{1,\dots,k}$, since they all have the collection of rows as A they clearly have the same row-span, and hence define the same toric model $\mathcal{M}_A$ as A. However, as we show in the next subsection, a different representation of the same model may affect the convergence of the iterative proportional scaling algorithm. Without loss of generality we assume A is of the above form.

Definition 2.5. The **index set** of α_i^ℓ, denoted I_i^ℓ, is the set of indices $j \in \{1, \dots, m\}$ such that the j-th entry of α_i^ℓ is one.

For a fixed partition A^ℓ and index $j \in \{1, \dots, m\}$ there is exactly one row α_i^ℓ such that j lies in its index set. We can define a function $\mathcal{S}(\ell, j) \in \{1, \dots, n_l\}$ where $\mathcal{S}(\ell, j)$ is the index such that $j \in I^\ell_{\mathcal{S}(\ell,j)}$. Then $\alpha^\ell_{\mathcal{S}(\ell,j)}$ is the row of A^ℓ where j lies in its index set.

Example 2.6. Let us now consider the following example, $m = 14, n_1, n_2 = 2$ and $n_3 = 4$.

$$
\begin{array}{cc}
 & j \in \{\ 1,\ 2, \cdots \qquad\qquad \cdots\ 13,\ 14\ \} \\
A = & \left(\begin{array}{cccccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\hdashline
1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \\
\hdashline
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1
\end{array}\right)
\begin{array}{l}
\left.\begin{array}{l} \alpha_1^1 \\ \alpha_2^1 \end{array}\right\} A^1 \\
\left.\begin{array}{l} \alpha_1^2 \\ \alpha_2^2 \end{array}\right\} A^2 \\
\left.\begin{array}{l} \alpha_1^3 \\ \alpha_2^3 \\ \alpha_3^3 \\ \alpha_4^3 \\ \alpha_5^3 \end{array}\right\} A^3
\end{array}
\end{array}
$$

Then $I_2^2 = \{4, 5, 6, 7, 11, 12, 13, 14\}$ and $\mathcal{S}(1, 10) = 2$.

We are able to apply the iterative proportional scaling algorithm as described in the next section to a partition model $\mathcal{M}_A$ in order to estimate the maximum likelihood of a data vector in $\mathbb{R}^m_+$. These are in fact the only integer matrices defining log-linear models for which we can apply the IPS algorithm as described here.

In this work we allow for repeated columns in A. We define the number of repetitions for a particular as the column's *weight*.

Definition 2.7. For a matrix A of the form (4), the **column weight** of the j-th column is number of times the column is repeated and is denoted $c_j \in \mathbb{Z}_+$.

Remark 2.8. For a matrix A defining a partition model, let $\bar{A}$ denote the $n \times \bar{m}$ matrix obtained by removing all repeated columns. This is then a **k-way quasi-independence model**. In fact what we call partition models can also be thought of as k-way quasi-independence models with repeated columns. We refer the reader to [12] for more information on the MLE of 2-way quasi-independence models.

2.2 Iterative Proportional Scaling

The iterative proportional scaling (IPS) algorithm is a method to calculate the maximum likelihood estimation of a normalized data vector d with respect to the model $\mathcal{M}_A$. Before we discuss the algorithm in more detail, we first define the maximum likelihood method.

Maximum likelihood estimation is a way to find an element in a statistical model, $\mathcal{M}_A$ that fits some observed data best. Let u be the vector of counts of the observed data.

Assuming that the observations are i.i.d., maximizing the likelihood of observing the data w.r.t. $\mathcal{M}_A$ leads to

$$\max_{p \in \mathcal{M}_A} \prod_j p_j^{u_j}.$$

This maximization is equivalent to maximizing the log-likelihood

$$\max_{p \in \mathcal{M}_A} log \left(\prod_j p_j^{u_j} \right) = \max_{p \in \mathcal{M}_A} u_j \, log \left(\prod_j p_j \right)$$

because of the monotonicity of the logarithm.

Let d be the empirical distribution of the data, then the maximum likelihood estimate is given by

$$p^* = \arg\max_{p \in \mathcal{M}_A} \sum_{j=1}^{m} d_j \, log \, p_j. \tag{5}$$

More details can be found in for example [37] and [17]. If $p^\star$ is rational for every d, then we say that $\mathcal{M}_A$ has rational MLE.

The MLE can be characterized by the following result.

Proposition 2.9 (Corollary 7.3.9 in [37])**.** *Let A be a matrix corresponding to a log-linear model and d be the empirical distribution of the data. If the maximum likelihood estimate in the model $\mathcal{M}_A$ exists, then it is the unique solution to the equations*

$$Ad = Ap, \qquad p \in \mathcal{M}_A.$$

This was referred to as Birch's theorem in [17] and [32].

The MLE can be calculated by applying an iterative algorithm, called IPS, iterative proportional fitting or iterative scaling algorithm. This is a well-known algorithm, first defined in the statistics literature by Deming and Stephan in 1940 in [16] and analyzed further by for example Csiszár in [13] and Brown in [7].

There exist various types of iterative scaling alorithms. In [17] and [37] the authors describe a variant of the original algorithm, proposed as generalized iterative scaling by Darroch and Ratcliff in [15]. We will focus on the earlier version defined below, because in this case each step is guaranteed to produce a rational function.

In order to guarantee that the initial distribution lies in the partition model, we will fix the uniform distribution $p^0 = (\frac{1}{n}, \ldots, \frac{1}{n})$ as the starting point. The ℓth-step of the algorithm is then defined as

$$p^\ell = p^{\ell-1} * \frac{A^i d}{A^i p^{\ell-1}} \tag{6}$$

for $i = \ell \bmod k$.

If p^ℓ is the MLE, then $Ad = Ap$ (Propositon 2.9) and every factor becomes 1. Every step is an information projection to the linear family L^i defined by the ith partition A^i.

$$L^i = \{p \in \Delta_{m-1} |\, A^i p = A^i d\} \tag{7}$$

The information projection of p to a non-empty, closed, convex set $M \subset \Delta_{m-1}$ is defined as the element $p' \in M$ that minimizes the KL-divergence between M and p

$$p' = \arg\min_{q \in M} D(q \parallel p) = \sum_{j=1}^{m} q_j \, log \, \frac{q_j}{p_j}.$$

The information projection of p to a linear family L^i is well-known and given by

$$p' = p * \frac{A^i d}{A^i p}.$$

see Lemma 4.1 in [14]. Hence the algorithm defined in (6) performs an information projection to a linear family in each step as sketched in Figure 1.

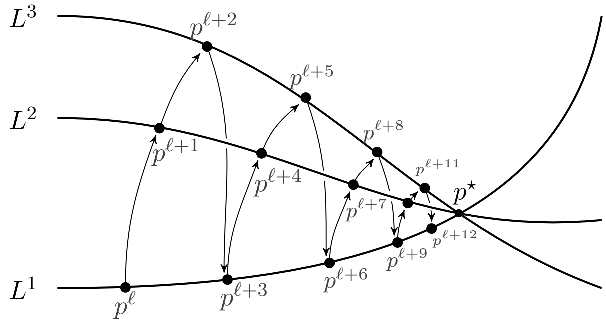


Figure 1: Sketch of the iterative proportional scaling algorithm in the case of three linear families.

Let $\mathcal{L}_A$ be the intersection of the linear families, $\mathcal{L}_A = \bigcap_i L_i$. This algorithm minimizes the KL-Divergence with respect to the first argument

$$p^\star = \arg\inf_{p \in \mathcal{L}_A} D(p \parallel q) \text{ for all } q \in \overline{\mathcal{M}}_A.$$

A proof of the convergence can be found in [14] Theorem 5.1. This is equivalent to minimizing the KL-Divergence with respect to the second argument as shown in Theorem 2.8 in [5]

$$p^\star = \arg\inf_{q \in \overline{\mathcal{M}}_A} D(p \parallel q) \text{ for all } p \in \mathcal{L}_A.$$

Since the normalized data vector d is in $\mathcal{L}_A$, this leads to the calculation of the MLE.

$$\begin{aligned} p^\star &= \arg\inf_{q \in \overline{\mathcal{M}}_A} D(d \parallel q) \\ &= \sum_{j=1}^{m} d_j\, log\, d_j + \arg\max_{q \in \overline{\mathcal{M}}_A} \sum_{j=1}^{m} d_j\, log\, q_j \end{aligned}$$

The last term is equivalent to our definition of the MLE as defined in (5).

Example 2.10 (The Independence model)**.** The normalized data is given by $d = (d_1, d_2, d_3, d_4)$ and the matrix for the 2x2 independence model is given below on the right.

Performing the first step of the algorithm results in:

$$\begin{aligned} p_1^0 = p_2^0 &= \frac{1}{4} \cdot \frac{d_1 + d_2}{\frac{2}{4}} = \frac{1}{2}(d_1 + d_2), \\ p_3^0 = p_4^0 &= \frac{1}{4} \cdot \frac{d_3 + d_4}{\frac{2}{4}} = \frac{1}{2}(d_3 + d_4) \end{aligned}$$

$$A = \begin{pmatrix} 1 & 1 & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 \\ \hline 1 & \cdot & 1 & \cdot \\ \cdot & 1 & \cdot & 1 \end{pmatrix}$$

In this case the algorithm converges to the MLE with the second step:

$$\begin{aligned} p_1^1 &= \frac{1}{2}(d_1 + d_2) \cdot \frac{d_1 + d_3}{\frac{1}{2}(d_1 + d_2) + \frac{1}{2}(d_3 + d_4)} = (d_1 + d_2)(d_1 + d_3) \\ p_2^1 &= \frac{1}{2}(d_1 + d_2) \cdot \frac{d_2 + d_4}{\frac{1}{2}(d_1 + d_2) + \frac{1}{2}(d_3 + d_4)} = (d_1 + d_2)(d_2 + d_4) \\ p_3^1 &= \frac{1}{2}(d_3 + d_4) \cdot \frac{d_1 + d_3}{\frac{1}{2}(d_1 + d_2) + \frac{1}{2}(d_3 + d_4)} = (d_3 + d_4)(d_1 + d_3) \\ p_4^1 &= \frac{1}{2}(d_3 + d_4) \cdot \frac{d_2 + d_4}{\frac{1}{2}(d_1 + d_2) + \frac{1}{2}(d_3 + d_4)} = (d_3 + d_4)(d_2 + d_4) \end{aligned}$$

Note that at each step of the algorithm the approximation of the MLE is a rational function of the data d. Thus if after finitely many steps the algorithm results in the MLE of d with respect to $\mathcal{M}_A$, the model has a rational maximum likelihood estimator. The question naturally arises, does iterative proportional scaling always result in the MLE after finitely many steps for every model with rational MLE?

Example 2.11. Here we consider the two matrices A and $\tilde{A}$ depicted on the right below.

Both matrices have full rowspan and with Proposition 2.3 also $\mathcal{M}_A = \mathcal{M}_{\tilde{A}}$ holds. Although they represent the same model, the convergence of the IPS algorithm is heavily influenced by the chosen representation. Using matrix $\tilde{A}$ the IPS algorithm converges in one step to the mle $p^\star = d$.

$$A = \left(\begin{array}{ccc} 1 & 1 & \cdot \\ \cdot & \cdot & 1 \\ \hdashline 1 & \cdot & \cdot \\ \cdot & 1 & 1 \end{array}\right) \qquad \tilde{A} = \begin{pmatrix} 1 & \cdot & \cdot \\ \cdot & 1 & \cdot \\ \cdot & \cdot & 1 \end{pmatrix}$$

The IPS algorithm on A with normalized data vector $d = (d_1, d_2, d_3)$ results in

$$p^0 = \left(\frac{1}{2}(d_1 + d_2),\, \frac{1}{2}(d_1 + d_2), d_3\right)$$

$$p^1 = \left(d_1,\, \frac{1}{2}(d_1 + d_2)\frac{(d_2 + d_3)}{\frac{1}{2}(d_1 + d_2) + d_3}, d_3\frac{(d_2 + d_3)}{\frac{1}{2}(d_1 + d_2) + d_3}\right)$$

$$p^2 = \left(d_1\frac{(d_1 + d_2)}{d_1 + \frac{\frac{1}{2}(d_1 + d_2)(d_2 + d_3)}{\frac{1}{2}(d_1 + d_2) + d_3}},\, \frac{1}{2}(d_1 + d_2)\frac{(d_2 + d_3)}{\frac{1}{2}(d_1 + d_2) + d_3}\frac{(d_1 + d_2)}{d_1 + \frac{\frac{1}{2}(d_1 + d_2)(d_2 + d_3)}{\frac{1}{2}(d_1 + d_2) + d_3}}, d_3\right)$$

In general the different projections have the following form:

$$p^k = \left(d_1, \qquad p_2^{k-1}\frac{(d_2 + d_3)}{p_2^{k-1} + d_3}, d_3\frac{(d_2 + d_3)}{p_2^{k-1} + d_3}\right), \quad k \text{ odd} \tag{8}$$

$$p^k = \left(d_1\frac{(d_1 + d_2)}{d_1 + p_2^{k-1}}, p_2^{k-1}\frac{(d_1 + d_2)}{d_1 + p_2^{k-1}}, d_3\right), \quad k \text{ even} \tag{9}$$

Assuming that there exists an index k such that the second index of p^k is exactly d_2 in (8) as well as (9) leads to $d_2 = p_2^{k-1}$. Hence IPS can only result in the exact result, if $d_1 = d_2$.

In a practical evaluation with 20 000 random input distributions the arithmetic mean of the iteration steps taken to get a step size smaller than 10^{-8} was $113,4767$ with a minimum value of 8 and a maximum value of 287 478. Recall that in case of $\tilde{A}$ the necessary iteration steps are only 1.

2.3 Toric Fiber Products

In this section, we review an algebraic construction, known as the *toric fiber product*, which will facilitate the proofs in the following sections. Toric fiber products were first introduced by Sullivant, and this exposition follows the notation of [36]. For the purposes of this article, we may broadly think of the toric fiber product as describing a way to concatonate the parametrizations of two log-linear models that preserves some of their properties.

For any positive integer r, let $[r] := \{1, \dots, r\}$. Fix a positive integer r and positive integers s_i, t_i for each $i \in [r]$. Define three polynomial rings,

$$\begin{aligned}\mathbb{C}[x] &= \mathbb{C}[x_j^i \mid i \in [r], j \in [s_i]],\\ \mathbb{C}[y] &= \mathbb{C}[y_k^i \mid i \in [r], k \in [t_i]], \text{ and}\\ \mathbb{C}[z] &= \mathbb{C}[z_{jk}^i \mid i \in [r], j \in [s_i], k \in [t_j]].\end{aligned}$$

Fix multigradings on $\mathbb{C}[x]$ and $\mathbb{C}[y]$ given by

$$\deg(x_j^i) = \deg(y_j^i) = \mathbf{d}^i,$$

for integer vectors $\mathbf{d}^i$. Note that the multigrading of each indeterminate depends only on its superscript, i. Let D be the matrix with columns $\mathbf{d}^i$ for $i \in [r]$.

Let $I \subset \mathbb{C}[x]$ and $J \subset \mathbb{C}[y]$ be homogeneous ideals with respect to this multigrading. We define a ring homomorphism,

$$\begin{aligned}\psi_{I,J} : \qquad \mathbb{C}[z] &\to \qquad (\mathbb{C}[x]/I) \otimes_{\mathbb{C}} (\mathbb{C}[y]/J)\\ z_{jk}^i &\mapsto \qquad x_j^i \otimes_{\mathbb{C}} y_k^i\end{aligned}$$

Definition 2.12. The **toric fiber product** of I and J with multigrading D is the ideal in $\mathbb{C}[z]$,

$$I \times_D J = \ker(\psi_{I,J}).$$

We assume throughout that I and J are toric, as the ideals associated to rational partition models are always toric. In order to parametrize the toric fiber product of I and J, we may simply take the product of their parametrizations. More precisely, if ϕ_I and ϕ_J are monomial parametrizations of I and J, then we obtain a parametrization for their toric fiber product according to multigrading D via

$$z_{jk}^i \mapsto \phi_I(x_j^i)\phi_J(y_k^i).$$

Now assume further that D is linearly independent. In this case, we can explicitly describe a generating for $I \times_D J$ using the following families of polynomials. First, let $i \in [r]$. Then we define

$$\mathrm{Quad}^i := \{z_{jk}^i z_{j'k'}^i - z_{jk'}^i z_{j'k}^i \mid j, j' \in [s_i], k, k' \in [t_i]\}.$$

By replacing each indeterminate with its parametrization in terms of ϕ_I and ϕ_J, we may see that each element of Quad^i belongs to $I \times_D J$. We let

$$\mathrm{Quad} = \cup_{\alpha=i}^r \mathrm{Quad}^i$$

Let $f = \prod_{\alpha=1}^d x_{j_\alpha}^{i_\alpha} - \prod_{\alpha=1}^d x_{j'_\alpha}^{i_\alpha} \in I$. Note that all binomials in I can be written in this form as D is linearly independent and I is homogeneous with respect to the multigrading specified by D. We define the set of binomials,

$$\mathrm{Lift}(f) = \left\{ \prod_{\alpha=1}^d z_{j_\alpha k_\alpha}^{i_\alpha} - \prod_{\alpha=1}^d z_{j'_\alpha k_\alpha}^{i_\alpha} \mid k_\alpha \in [t_{i_\alpha}] \text{ for all } \alpha \in [d] \right\}.$$

Observe that each binomial in $\mathrm{Lift}(f)$ also lies in $I \times_D J$. For each binomial $g \in J$, we define $\mathrm{Lift}(g)$ analogously. Let F be a generating set for I. Then we define

$$\mathrm{Lift}(F) = \cup_{f \in F} \mathrm{Lift}(F),$$

and similarly for any generating set of J. We are now able to describe the generating set, and in fact, a Gröbner basis, for $I \times_D J$. The following is an adaptation of Theorem 2.9 of [36] for our purposes.

Theorem 2.13 ([36]). *Let I and J be toric ideals that are homogeneous with respect to the multigrading specified by D. Suppose further that the columns of D are linearly independent. Let F be a Gröbner basis for I and let G be a Gröbner basis for J. Then*

$$\mathrm{Quad} \cup \mathrm{Lift}(F) \cup \mathrm{Lift}(G)$$

is a Gröbner basis for $I \times_D J$ with respect to a certain weight order.

We conclude this introduction to the toric fiber product by describing the MLE of the model associated to the toric fiber product in terms of the MLE of its factors. Let B, C be matrices representing the log-linear models associated to I and J. Let $\hat{p}(B), \hat{p}(C)$ and $\hat{p}(D)$ denote the maximum likelihood estimators for the log-linear models given by B, C and D, respectively. The following is Theorem 5.5 of [3] and plays a critical role in our proofs of the results in Section 3.

Theorem 2.14 ([3]). *The (i, j, k)th coordinate function of the maximum likelihood estimator for the log-linear model associated to the toric fiber product $I \times_D J$ is*

$$\hat{p}^i_{jk} = \frac{\hat{p}^i_j(B)\hat{p}^i_k(C)}{\hat{p}^i(D)}.$$

2.4 Balanced Staged Tree Models

Here we introduce staged tree models, following the presentation in [4]. Staged tree models are a type of probability tree model, which are graphical models that encode conditional independence statements between events rather than between random variables as in hierarchical models. For a detailed introduction see [10]. As a class of models, staged tree models contain discrete Bayesian networks and decomposable graphical models [20]. In this section we note that *balanced* staged tree models correspond to partition models with a rational maximum likelihood estimator.

Let $\mathcal{T} = (V, E)$ be a connected, acyclic graph where V denotes its vertex set and E is the set of edges. Such a graph is known as a *tree*, and it is a *directed* tree if each edge is an ordered pair of vertices. Furthermore a directed tree is *rooted* if there exists a vertex v_0, which we called the *root*, with no incoming edges. Conversely, a vertex v is called a *leaf*, if there are no outgoing edges.

For a particular vertex $v \in V$, we can define the set of *children* of v as the set $\mathrm{ch}(v) = \{u \,|\, (v, u) \in E\}$. Then the outgoing edges from v can be denoted as the set $E(v) = \{(v, u) \,|\, u \in \mathrm{ch}(v)\}$, and v is a leaf if $E(V) = \emptyset$. A tree is *labelled* if there exists a label set S and a surjective mapping, called a labelling, $\theta : E \to S$. From now on we assume that $(\mathcal{T}, \theta)$ is a directed, rooted, and labelled tree graph.

Definition 2.15. For any vertex $v \in V$, the set of labels of its outgoing edges is the **floret** of v, denoted $\mathcal{F}_v := \{\theta(e) \,|\, e \in E(v)\}$. The set of florets for $(\mathcal{T}, \theta)$ is denoted $F_{\mathcal{T}}$.

Definition 2.16. The labelled tree graph $(\mathcal{T}, \theta)$ is a **staged tree** if the following are satisfied:

- For each $v \in V$, no two edges in $E(v)$ have the same labelling, i.e. $|\theta_v| = |E(v)|$.
- For any two vertices $v, w \in V$, the florets $\mathcal{F}_v$ and $\mathcal{F}_w$ are equal or disjoint. When $\mathcal{F}_v = \mathcal{F}_w$ we say that the vertices v and w are in the same *stage*.

A *staged tree model* can be associated to a staged tree $(\mathcal{T}, \theta)$ as follows. Let Λ denote the set of root-to-leaf paths in $\mathcal{T}$ and suppose $|\Lambda| = m$. For $\lambda \in \Lambda$ let $E(\lambda)$ be the set of edges in Λ and define $\theta(\lambda) := \{s_i \,|\, \theta(e) = s_i \text{ for } e \in E(\lambda)\}$. The paths λ parameterize the model $\mathcal{M}_{\mathcal{T}}$ defined below.

Definition 2.17. For a staged tree $(\mathcal{T}, \theta)$ with label set $S = \{s_1, \ldots, s_r\}$, the **staged tree model** $\mathcal{M}_{\mathcal{T}}$ is the image of the map $\phi_{\mathcal{T}} : \Theta \to \Delta_{m-1}$ defined by

$$\phi_{\mathcal{T}}(s_1, \ldots, s_r) = \left(\prod_{s_i \in \theta(\lambda)} s_i \right)_{\lambda \in \Lambda} \tag{10}$$

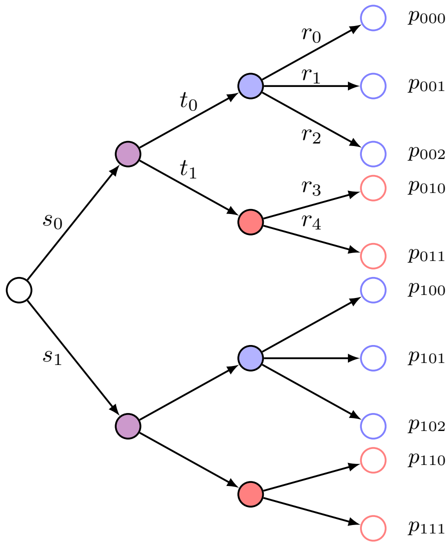
where the parameter space is

$$\Theta := \{(s_1, \ldots, s_r) \in (0, 1]^r \,|\, \sum_{s_i \in \mathcal{F}} s_i = 1 \text{ for all florets } \mathcal{F} \in F_{\mathcal{T}}\}.$$

Example 2.18. Consider the staged tree $(\mathcal{T}, \theta)$ on the right.

In this diagram, the vertices in the same stage have the same coloring. We can immediately see that each floret has equal or disjoint label set. We can refer to the vertices of the tree by the edge labels defining the path to the vertex, i.e. the blue vertices are $v_{s_0t_0}$ and $v_{s_1t_0}$. If v_0 is the root vertex then the florets of the tree are

$$\mathcal{F}_{v_0}, \quad \mathcal{F}_{v_{s_0}} = \mathcal{F}_{v_{s_1}}, \\ \mathcal{F}_{v_{s_0t_0}} = \mathcal{F}_{v_{s_1t_0}}, \\ \mathcal{F}_{v_{s_0t_1}} = \mathcal{F}_{v_{s_1t_1}}.$$



The associated staged tree model $\mathcal{M}_\mathcal{T}$ is image of the map $\phi_\mathcal{T} : \Theta \to \Delta_9$ where

$$\phi_\mathcal{T}(s_0, s_1, t_0, t_1, r_0, r_1, r_2) = (s_0t_0r_0, s_0t_0r_1, s_0t_0r_2, s_0t_1r_3, s_0t_1r_4, \\ s_1t_0r_0, s_1t_0r_1, s_1t_0r_2, s_1t_1r_3, s_1t_1r_4),$$

and

$$\Theta = \{(s_0, s_1, t_0, t_1, r_0, r_1, r_2) \in (0, 1]^7 \mid s_0 + s_1 = 1, t_0 + t_1 = 1, \\ r_0 + r_1 + r_2 = 1, r_3 + r_3 + r_4 = 1\}.$$

Remark 2.19. We use the slightly more general definitions of staged tree and staged tree models introduced in [4] that allow for staged trees with singleton florets. The authors show that the staged tree obtained by contracting these florets, which satisfies the definition used in previous literature (i.e. [10, 23]), results in the same staged tree model [4, Lemma 5.11].

The map $\phi_\mathcal{T}$ in Definition 2.17 also defines a toric model if one "forgets" the conditions imposed on the parameter space by the florets. We denote the map obtained by extending $\phi_\mathcal{T}$ to all of $\mathbb{R}^r$ as $\overline{\phi}_\mathcal{T} : \mathbb{R}^r \to \mathbb{R}^m$ and the corresponding toric model as $\overline{\mathcal{M}}_\mathcal{T}$. Immediately it follows that $\mathcal{M}_\mathcal{T} \subset \overline{\mathcal{M}}_\mathcal{T}$, but in general it is not true that $\mathcal{M}_\mathcal{T} = \overline{\mathcal{M}}_\mathcal{T}$. In [18] the authors found necessary and sufficient graphical conditions on $(\mathcal{T}, \theta)$ such that $\mathcal{M}_\mathcal{T} = \overline{\mathcal{M}}_\mathcal{T}$. We detail those conditions in Section 4.2.

Definition 2.20. For any vertex in a staged tree $(\mathcal{T}, \theta)$, we define the **level** $\ell(v)$ as the number of edges in the unique root-to-v path. A staged tree $(\mathcal{T}, \theta)$ is **stratified** if all its leaves have the same level and for any two vertices v, w such that $\mathcal{F}_v = \mathcal{F}_w$, $\ell(v) = \ell(w)$. We say that the **level** of a stratified staged tree is the level of its leaves.

There is a natural partition of the label set of a stratified staged tree by level. In other words for a stratified staged tree of level k, we can write $S = \{s^0_0, s^0_1, \dots, s^0_{n_0}, s^1_0, \dots, s^k_{n_k}\}$

where n_ℓ is the number of distinct edge labels for edges emanating from vertices with level ℓ. Thus the monomial map $\phi_\mathcal{T}$ for a stratified staged tree of level k is multilinear and homogeneous which implies that the associated matrix is a matrix of form (4) with k partitions. We denote the matrix associated with $\phi_\mathcal{T}$ as $A_\mathcal{T}$.

Similarly with any matrix A of the form (4) we can associate a directed, rooted tree with a labelling induced by the rows of A. We define the label set associated with A as $S_A = \{s_1^1, \dots, s_{n_1}^1, \dots, s_1^k, \dots, s_{n_k}^k\}$ associated to the row vectors $\alpha_1^1, \alpha_2^1, \dots, \alpha_{n_k}^k$.

Starting with a root vertex, we add a labelled edge s_i^1 for each associated row vector in A^1. To each edge s_i^1, we attach an edge labelled s_j^2 to the outgoing vertex for each s_j^2 is the label set S_A such that $I_i^1 \cap I_j^2 \neq \emptyset$. The resulting directed, rooted, and labelled tree graph is $\mathcal{T}_{A^{1,2}}$ where $A^{1,2}$ is the matrix with partitions A_1, A_2. The paths of $\mathcal{T}_{A^{1,2}}$ exactly correspond to the unique columns of A. In this way we can define $\mathcal{T}_A$ as the directed, rooted tree graph with labelling such that the paths in $\mathcal{T}_A^{(\ell)}$ correspond to the unique columns of $A^{1,\dots,\ell}$.

We will see that staged trees, as combinatorial objects, are useful in describing certain properties of partition matrices. In particular, when we introduce conditions on partition matrices to produce the MLE after one cycle of IPS, we require that $\mathcal{T}_A$ is a staged tree. We also show in Section 2.4 that staged tree models, when $\mathcal{M}_\mathcal{T} = \overline{\mathcal{M}_\mathcal{T}}$, also produce the MLE after one cycle of IPS.

2.5 Hierarchical Models

A hierarchical model is a of log-linear models for which the interaction structure between the columns can be described by a simplicial complex. We will introduce hierarchical models using the notation in [37]. They compose another interesting class of partition models.

Definition 2.21. Let 2^J be the powerset of $J = \{1, \dots, l\}$, meaning the set of all the subsets of J. A *simplicial complex* Γ with the ground set J is a subset of 2^J with the property that if $F \in \Gamma$ and $F' \subset F$, then $F' \in \Gamma$. The elements of a simplicial complex are the faces of Γ and all inclusion maximal faces are called facets.

A simplicial complex can be identified by its set of facets $facet(\Gamma)$, therefore we write $\Gamma = [12][13]$ for $\Gamma = \{\emptyset, \{1\}, \{2\}, \{3\}, \{1,2\}, \{1,3\}\}$.

Up to this point we did not assume any additional structure on Δ_{m-1}, but now we will consider a collection of random variables $X_1, \dots, X_l$ with the state spaces $R_1, \dots, R_l$. A simplicial complex with the ground set $\{1, \dots, l\}$ now describes the interactions among the different random variables on the joint state space $R = R_1 \times \cdots \times R_l$. The corresponding log-linear model is called hierarchical model and defined as follows.

Definition 2.22. Let Γ be a simplicial complex and for every facet $F = \{f_1, f_2, \dots\} \in \Gamma$ let $r_F = \prod_{f \in F} |R_f|$. For $i = (i_1, \dots, i_l) \in R$ let $i_F = (i_{f_1}, i_{f_2}, \dots)$. Additionally, we introduce a set of r_F positive parameters $\theta^F_{f_i}$ for every facet. Now the **hierarchical model** corresponding to Γ is

$$\mathcal{M}_\Gamma = \left\{ p \text{ p.d. on } R \,|\, p_i = \frac{1}{Z(\theta)} \prod_{F \in facet(\Gamma)} \theta^F_{i_F} \right\}$$

with the normalizing factor

$$Z(\theta) = \sum_{i \in R} \prod_{F \in facet(\Gamma)} \theta^F_{i_F}.$$

A hierarchical model with the simplicial complex Γ can be associated with a graph $G(\Gamma)$. The vertex set of this graph is the ground set J and two vertices are connected by an edge if they lie together in a face in Γ. A subgraph is called complete if every vertex is connected to every other vertex in this subgraph. If a complete subgraph is maximal with respect to inclusion, then it is called a clique. If the cliques of the graph $G(\Gamma)$ are exactly the facets of Γ, then Γ is called conformal [38] or graphical [31]. These concepts are demonstrated in the next example.

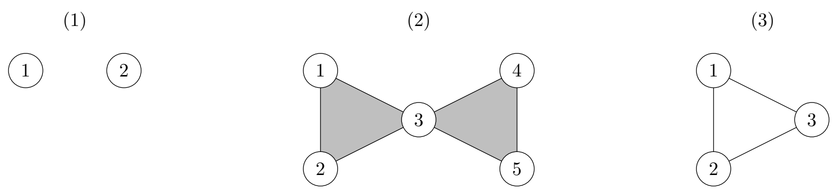


Figure 2: Three different examples of hierarchical models.

Example 2.23. (1) We already mentioned the 2x2 independence model in Example 2.10. Each partition in the matrix corresponds to a facet of the corresponding simplicial complex $\Gamma = [1][2]$. This corresponds to a graph with two vertices X_1, X_2 that are not connected, as depicted in Figure 2 (1).

(2) An example for a conformal simplicial complex is $\Gamma = [1,2,3][3,4,5]$. The associated graph is depicted in Figure 2 in the middle. This graph has two cliques $\{1,2,3\}$ and $\{3,4,5\}$, which are exactly the facets of Γ.

(3) The simplicial complex $\Gamma = [12][13][23]$ is not conformal, since the clique $\{1,2,3\}$ is not a face of Γ.

Definition 2.24. A simplicial complex Γ is called **decomposable** if it consists only of one facet or if $facet(\Gamma)$ is the union of two disjoint sets Γ_1, Γ_2 such that there exist $F_1 \in \Gamma_1$ and $F_2 \in \Gamma_2$ with

$$\bigcup_{F \in \Gamma_1} F \cap \bigcup_{\tilde{F} \in \Gamma_2} \tilde{F} = F_1 \cap F_2$$

In [31] in Theorem 2.25 Lauritzen shows that every decomposable simplicial complex is conformal. The simplicial complex in Example 2.23 (2) is decomposable with $F_1 = \{\{1,2,3\}\}$ and $F_2 = \{\{3,4,5\}\}$. The example in 2.23 (3) is not decomposable.

Now we will introduce a special ordering to the facets of a simplicial complex.

Definition 2.25. Let $E = \{F_1, \ldots, F_s\}, s \in \mathbb{N}$ be an ordering of the facets of a simplicial complex Γ. Then this ordering satisfies the **running intersection property (RIP)**, if for each $r \in \{1, \ldots, s\}$ there exists a k_r such that

$$\left(\bigcup_{k=1}^{r} F_k\right) \cap F_{r+1} = F_{k_r} \cap F_{r+1}$$

The connection between decomposable simplicial complexes and the running intersection property, as stated in the Lemma below, is proven for example in Lemma 5.10 in [25].

Lemma 2.26. *Let Γ be a decomposable simplicial complex. Then there exists an ordering of the facets such that this ordering satisfies the RIP.*

Theorem 2.27. *Suppose the simplicial complex Γ is decomposable and the initial distribution lies in the model $\mathcal{M}_\Gamma$, then the IPS algorithm converges in one cycle.*

Proof. This theorem is proven in Theorem 5.3 in [25]. □

3 Generalized Running Intersection Property

In this section, we define the generalized running intersection property, or GRIP. We show that a partition model satisfies the GRIP if and only if it can be written as an iterated toric fiber product of partition models. This allows us to give a formula for the maximum likelihood estimate of a model that satisifies the GRIP. We use this formula to show that iterative proportional scaling exhibits one-cycle convergence on models that satisfy the GRIP.

Definition 3.1. Let B and C be two partition matrices with the same number of columns and with rows α_i^B and $\alpha_{i'}^C$. Two rows α_i^B and $\alpha_{i'}^C$ are **connected** if their supports intersect nontrivially; that is, if $I_i^B \cap I_{i'}^C \neq \emptyset$.

For a matrix to satisfy the GRIP, we require that rows from one partition to the next are connected in way that satisfies certain conditions. The following definition is motivated by the way column weights behave under steps of the IPS algorithm.

Definition 3.2. For a matrix A of the form (4), define c_j^ℓ as the j-th column weight (see Definition 2.7) for the matrix obtained by only considering the first ℓ partitions of A. Then A is **well-connected** if for any row vector α_i^ℓ, where $\ell > 1$, we have that

$(c_j^\ell / c_j^{\ell-1}) = (c_{j'}^\ell / c_{j'}^{\ell-1})$ for all $j, j' \in I_i^\ell$. We call this quantity the **connection ratio** for α_i^ℓ and denote this quantity as C_i^ℓ with the convention that $C_i^1 = |I_i^1|$.

Remark 3.3. Suppose that A is a well-connected matrix. Then for any index $j \in \{1, \dots, m\}$, we see that the column weight can be expressed as the product of the connection ratios:

$$c_j = \prod_{\ell=1}^{k} C^\ell_{\mathcal{S}(\ell,j)}.$$

Example 3.4. Consider the matrix A below. We note that it is well-connected and display the connection ratios to the right of the matrix.

$$\begin{array}{cc}
A & C_i^\ell \\
\left(\begin{array}{cccccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\hdashline
1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \\
\hdashline
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1
\end{array}\right)
&
\begin{array}{c}
7 \\ 7 \\ \frac{3}{7} \\ \frac{4}{7} \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{1}{3} \\ \frac{2}{4} \\ \frac{2}{4}
\end{array}
\end{array}$$

Let $A^1, \dots, A^\ell$ be a set of partition matrices with m columns and $A^{1,\dots,\ell} = A^1 A^2 \dots A^\ell$. Let β be number of *distinct* columns of $A^{1,\dots,\ell}$. Define a labeling of the columns of $A^{1,\dots,\ell}$, $\lambda : [m] \to [\beta]$ such that $\lambda(j) = \lambda(j')$ if and only if the jth and j'th columns of $A^{1,\dots,\ell}$ are equal.

Definition 3.5. We define the partition matrix $B = \Cup_{n=1}^{\ell} A^n$ to be the $\beta \times m$ matrix with jth column $e_{\lambda(j)}$. Since the labeling λ of the columns of $A^{1,\dots,\ell}$ simply permutes the rows of $\Cup_{n=1}^{\ell} A^n$, we omit the specification of λ from this notation.

In order to define the GRIP, we require that the tree associated to the partition matrix be staged. The word choice of "floret" in the definition below is deliberately suggestive of the terminology for staged trees discussed in Section 2.4.

Definition 3.6. Let B and C be two partition matrices with the same number of columns and with rows α_u^B and α_v^C. The matrices B and C satisfy the **floret condition** if for every two rows of B, α_u^B and $\alpha_{u'}^B$, the sets of rows of C that are connected to α_u^B and $\alpha_{u'}^B$ are disjoint or equal. In this case, the set of rows of B connected to a row α_v^C is called a **floret** of B and the set of rows of C connected to a row α_u^B is a **floret** of C.

The florets of the matrix C correspond exactly to the second-level florets of the tree associated with the partition matrix BC (defined in (3)). In fact we can say that BC satisfies the floret condition if and only if $\mathcal{T}_{BC}$ is a staged tree.

Suppose that B and C are two partition matrices that satisfy the floret condition, and let $\mathcal{F}_1^C, \ldots, \mathcal{F}_f^C$ be the distinct florets of C with respect to the matrix B. For each $t \in [f]$, let $\mathcal{F}_t^B$ be the set of all rows of B that are connected to the rows of $\mathcal{F}_t^C$.

Remark 3.7. The floret condition in Definition 3.6 implies that $\mathcal{F}_t^B$ is well-defined for any $t \in [f]$ and that this encompasses all florets of B. Indeed the florets of B can be identified in a one-to-one correspondence with the florets of C in this way.

Let $t(u,v)$ be the index in $[f]$ such that $\alpha_u^B \in \mathcal{F}_{t(u,v)}^B$ and $\alpha_v^C \in \mathcal{F}_{t(u,v)}^C$. In fact, $t(u,v)$ can be determined by the row α_u^B *or* the row α_v^C; it is not necessary to provide both indices. Hence, we write $t(u,\bullet)$ to be the index such that $\alpha_u^B \in \mathcal{F}_{t(u,\bullet)}^B$. Similarly, $t(\bullet,v)$ is the index such that $\alpha_v^C \in \mathcal{F}_{t(\bullet,v)}^C$. Thus α_u^B and α_v^C are connected if and only if $t(u,v) = t(u,\bullet) = t(\bullet,v)$. Note that the function $t(u,v)$ implicitly depends on the matrices B and C.

Example 3.8. Consider the matrix in Example 3.4, and let $B = A^1 \uplus A^2$, which results in the following matrix:

$$B = A^1 \uplus A^2 = \begin{pmatrix} 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \end{pmatrix}.$$

Now let $C = A^3$. Note that B and C satisfy the floret condition and that the florets of C are given by $\mathcal{F}_1^C = \{\alpha_1^C, \alpha_2^C, \alpha_3^C\}$ and $\mathcal{F}_2^C = \{\alpha_4^C, \alpha_5^C\}$. Note that the florets of C are exactly the third-level florets of the staged tree $\mathcal{T}_A$.

From the florets of C, we can clearly see that $t(\bullet,1) = t(\bullet,2) = t(\bullet,3) = 1$ and $t(\bullet,4) = t(\bullet,5) = 2$. The corresponding florets of B are given by $\mathcal{F}_1^B = \{\alpha_1^B, \alpha_3^B\}$ and $\mathcal{F}_2^B = \{\alpha_2^B, \alpha_4^B\}$. This implies that $t(1,\bullet) = t(3,\bullet) = 1$ and $t(2,\bullet) = t(4,\bullet) = 2$.

At this point the function $t(u,v)$ is completely defined and it can be used to answer questions like, "Which floret of C is connected to the row α_3^B?" This is then given by $\mathcal{F}_{t(3,\bullet)}^C = \mathcal{F}_1^C$.

For a set of partition matrices $A^1, \ldots, A^\ell$ with m columns, let $B = \uplus_{n=1}^{\ell} A^n$. Let $C = A^{\ell+1}$ have γ rows and m columns. Then the matrix BC has m columns of the form $[e_u\ e_v]^T$ where $u \in [\beta]$ and $v \in [\gamma]$. Let ω_{uv} denote the number of columns of C of the form $[e_u\ e_v]^T$. Then the columns of BC can be indexed by triples of the form (u,v,s) for $s \in \{0, \ldots, \omega_{uv} - 1\}$ where column (u,v,s) is the $(s+1)$st column of BC of the form $[e_u\ e_v]^T$.

Definition 3.9. Suppose that B and C satisfy the floret condition with f florets, $\mathcal{F}_1^C, \dots, \mathcal{F}_f^C$. We define the matrix $B \Cap C$ to be the $f \times m$ matrix whose columns are indexed by the triples (u, v, s) such that the (u, v, s) entry of the tth row of $B \Cap C$ is equal to 1 if $t = t(u, v)$ and 0 otherwise. Note that any two columns with indices (u, v, s) and (u, v, s') are identical.

In other words, the columns of $B \Cap C$ are indicator vectors for the florets that each column's non-zero rows belong to. Note that $B \Cap C$ is only defined when B and C satisfy the floret condition.

Definition 3.10. Let $A^1, \dots, A^k$ be partition matrices. For each ℓ, let B_ℓ denote $\Cup_{n=1}^{\ell} A^n$. Then $A^{1,\dots,k}$ satisfies the **generalized running intersection property**, or **GRIP** if for each $1 \leq \ell \leq k-1$,

1. the matrix $B_\ell A^{\ell+1}$ is well-connected,
2. $B_\ell A^{\ell+1}$ satisfies the floret condition, and
3. the rows of $B_\ell \Cap A^{\ell+1}$ lie in the rowspan of $A^{1,\dots,\ell}$.

Remark 3.11. The first point in the above definition is equivalent to $A^{1,\dots,k}$ being well-connected. Additionally it is easy to see that the second point is satisfied if and only if $\mathcal{T}_{A^{1,\dots,k}}$ is a staged tree. We frame these conditions above in an iterative way to facilitate the proofs in this section.

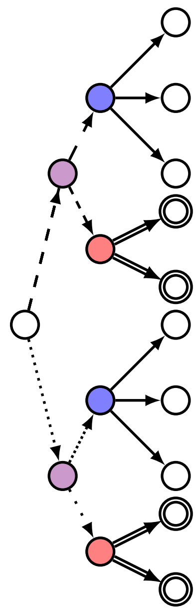
Example 3.12. We show that the matrix A from Example 3.4 satisfies the GRIP. Since A is well-connected and $\mathcal{T}_A$ is a staged tree, (see Example 2.18), A satisfies conditions (1) and (2) from Definition 3.10. We have included the matrix A and the associated tree below, as well as all the necessary partition matrix operations to check the third condition of the GRIP.

$$A = \left(\begin{array}{cccccccccccccc}
1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\
\hdashline
1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \\
\hdashline
1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot \\
\cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1
\end{array}\right)$$

$$A^1 \Cap A^2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix}$$

$$A^1 \Cup A^2 = \begin{pmatrix}
1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1
\end{pmatrix}$$

$$(A^1 \Cup A^2) \Cap A^3 = \begin{pmatrix}
1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\
\cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1
\end{pmatrix}$$



For $\ell = 1$, $B^1 = A^1$ and (3) is trivially satisfied. Thus all that is left is to check that the rows of $B^2 \Cap A^3$ lie in the rowspan of A. First consider $B^2 = A^1 \Cup A^2$. This matrix indexes all the unique columns of the two-partition matrix A^1A^2, and describes all the distinct paths in $\mathcal{T}_{A^1A^2}$.

Note the rows of the matrix $B^2 \Cap A^3 = (A^1 \Cup A^2) \Cap A^3$ indexes the level three florets of the staged tree $\mathcal{T}_A$. Since $B^2 \Cap A^3 = A^2$, we can easily see that (3) is satisfied.

We now describe the iterated toric fiber product structure on the partition model defined by a matrix $A^{1,\dots,k}$ that satisfies the GRIP. The following lemma is crucial to our construction of the toric fiber product.

Lemma 3.13. *Suppose that BC is well-connected. Then there exist positive integers $x_1, \dots, x_\beta, y_1, \dots, y_\gamma$ such that $\omega_{uv} = x_u y_v$ for all $u \in [\beta]$ and $v \in [\gamma]$.*

Proof. See Proposition 1 in [27]. $\square$

Let $A^{1,\dots,k}$ satisfy the GRIP and consider the submatrix $A^{1,\dots,\ell+1} = A^1A^2\dots A^{\ell+1}$, where $1 \le \ell \le n-1$. Fix $B = \Cup_{n=1}^{\ell} A^n$ and $C = A^{\ell+1}$ as in the paragraph preceding Definition 3.10. Define the polynomial ring,

$$R = \mathbb{C}[p_{u,v,s}^{t(u,v)} \mid u \in [\beta], v \in [\gamma], s \in \{0, \dots, \omega_{uv} - 1\}].$$

Note that the index $t(u,v)$ is determined by u and v and keeps track of which floret the support of a column belongs to.

Let $x_1, \dots, x_\beta, y_1, \dots, y_\gamma$ be positive integers such that $\omega_{uv} = x_u y_v$; these are guaranteed to exist by Lemma 3.13. Define two other polynomial rings,

$$R_B = \mathbb{C}[q_{u,s'}^{t(i,\bullet)} \mid u \in [\beta], s' \in \{0, \dots, x_u - 1\}], \text{ and}$$
$$R_C = \mathbb{C}[r_{v,s''}^{t(\bullet,v)} \mid v \in [\gamma], s'' \in \{0, \dots, y_v - 1\}].$$

Recall that λ maps the columns of $A^{1,\dots,\ell}$ onto $[\beta]$ by labeling the columns of $A^{1,\dots,\ell}$ so that $\lambda(u) = \lambda(v)$ if and only if the uth and vth columns of $A^{1,\dots,\ell}$ are equal. Let $\tilde{A}^{1,\dots,\ell}$ be the matrix defined as follows. Its columns are indexed by ordered pairs (u, s') such that $u \in [\beta]$ and $s' \in \{0, \dots, x_u - 1\}$. The (u, s') column of $\tilde{A}^{1,\dots,\ell}$ is equal to the column u' of $A^{1,\dots,\ell}$ with $\lambda(u') = u$. In other words, the matrix $\tilde{A}^{1,\dots,\ell}$ has the same underlying set of columns as $A^{1,\dots,\ell}$; however, the column with label u is repeated in $\tilde{A}^{1,\dots,\ell}$ only x_u times. In general, we say that a matrix $\tilde{A}$ is a *compression* of A if A and $\tilde{A}$ have the same underlying set of columns, but $\tilde{A}$ contains at most as many copies of each column as A.

Similarly, let $\tilde{A}^{\ell+1}$ have columns indexed by ordered pairs (v, s'') such that $v \in [\gamma]$ and $s'' \in \{0, \dots, y_v - 1\}$. The (v, s'') column of $\tilde{A}^{\ell+1}$ is equal to the vth standard basis vector. The matrix $\tilde{A}^{\ell+1}$ has the same underlying set of columns as $A^{\ell+1}$, and this underlying set is equal to $\{e_1, \dots, e_\gamma\}$; however, the vth standard basis vector is repeated in $\tilde{A}^{\ell+1}$ only y_v times.

Recall that for any integer matrix A, $I(A)$ is vanishing ideal of the log-linear statistical model $\mathcal{M}_A$ as described in Proposition 2.3 . We define a ring homomorphism by

$$\begin{aligned} \psi : R &\to R_B/I(\tilde{A}^{1,\dots,\ell}) \otimes_{\mathbb{C}} R_C/I(\tilde{A}^{\ell+1}) \\ p_{u,v,s}^{t(u,v)} &\mapsto q_{u,s'}^{t(u,\bullet)} \otimes_{\mathbb{C}} r_{v,s''}^{t(\bullet,v)}, \end{aligned} \tag{11}$$

where s', s'' are the nonnegative integers such that $s = s'y_v + s''$. Note that each $s \in \{0, \dots, \omega_{uv} - 1\}$ has a unique representation in this way. Define a multigrading on R_B by $\deg(q_{u,s'}^{t(u,\bullet)}) = e_{t(u,\bullet)}$ and a multigrading on R_C by $\deg(r_{v,s''}^{t(\bullet,v)}) = e_{t(\bullet,v)}$. We will show that $\ker(\phi)$ is the toric fiber product of $I(\tilde{A}^{1,\dots,\ell})$ and $I(\tilde{A}^{\ell})$ with respect to this multigrading and that $\ker(\psi) = I(A^{1,\dots,\ell+1})$.

Proposition 3.14. *The toric ideals $I(\tilde{A}^{1,\dots,\ell})$ and $I(\tilde{A}^{\ell+1})$ are multihomogeneous with respect to the multigradings $\deg(q_{u,s'}^{t(u,\bullet)}) = e_{t((u,\bullet)}$ and $\deg(r_{v,s''}^{t(\bullet,v)}) = e_{t(\bullet,v)}$.*

Proof. Let $D = B \Cap C$. The rows of D are in the rowspan of $A^{1,\dots,\ell}$ by the generalized running intersection property. Moreover, the row of D corresponding to a floret $\mathcal{F}$ in C is $\displaystyle\sum_{\alpha_v^C \in \mathcal{F}} \alpha_v^C$. Hence, the rows of D are in the rowspan of $A^{\ell+1}$ as well.

The matrix $\tilde{A}^{1,\dots,\ell}$ is obtained by deleting some columns of $A^{1,\dots,\ell}$. Let $\tilde{D}$ be obtained from $D = B \Cap C$ by deleting the same columns. Then the matrix $\tilde{D}$ defines the grading

$\deg(q_{u,s'}^{t(u,\bullet)}) = e_{t(u,\bullet)}$ on R_B. The rows of $\tilde{D}$ remain in the rowspan of $\tilde{A}^{1,\dots,\ell}$. Thus $I(\tilde{A}^{1,\dots,\ell})$ is multihomogeneous with respect to the grading given by $\tilde{D}$ since every vector in $\ker(\tilde{A}^{1,\dots,\ell})$ is also in $\ker(\tilde{D})$. The argument for the linear ideal $I(\tilde{A}^{\ell+1})$ is analogous. □

Theorem 3.15. *If a partition model $A^{1,\dots,\ell+1}$ satisfies the generalized running intersection property, then $I(A^{1,\dots,\ell+1})$, the vanishing ideal of $\mathcal{M}_A$, can be obtained via a toric fiber product. In particular this toric fiber product is given by* $\ker(\psi)$*, defined in* (11)*.*

Proof. By Proposition 3.14, $I(\tilde{A}^{1,\dots,\ell})$ and $I(\tilde{A}^{\ell+1})$ are both multihomogeneous with respect to the grading specified by D. For each $m \in [\ell+1]$, let ρ^k denote the number of rows of A^m. We can replace each ideal with its parametrization to rewrite the map ψ as

$$\begin{aligned} R \to \quad & \mathbb{C}[\eta_g^h \mid h \in [\ell], g \in [\rho^h]] \otimes_{\mathbb{C}} \mathbb{C}[\eta_g^{\ell+1} \mid g \in [\rho^{\ell+1}]] \\ p_{u,v,s}^{t(u,v)} \mapsto \quad & \prod_{h=1}^{\ell} \eta^h_{\mathcal{S}(h,(u,v,s))} \otimes_{\mathbb{C}} \eta^{\ell+1}_{\mathcal{S}(\ell+1,(u,v,s))} \\ & = \prod_{h=1}^{\ell+1} \eta^p_{\mathcal{S}(h,(u,v,s))}. \end{aligned}$$

This is exactly the map $\phi_{A^{1,\dots,\ell+1}}$ whose kernel is $I(A^{1,\dots,\ell+1})$, as needed. □

For each $u \in [\beta]$, $\mathcal{F}^C_{t(u,\bullet)}$ is the set of rows of C that are connected to α_u^B. Let x_u, y_v be the positive integers such that

$$\omega_{uv} = \begin{cases} x_u y_v & \text{if } t(u,\bullet) = t(\bullet, v) \\ 0 & \text{otherwise.} \end{cases}$$

Then the number of copies of column e_u in B is

$$\sum_{\alpha_v^C \in \mathcal{F}^C_{t(u,\bullet)}} \omega_{uv} = x_u \sum_{\alpha_v^C \in \mathcal{F}^C_{t(u,\bullet)}} y_v.$$

For each u, let

$$Y_u = \sum_{\alpha_v^C \in \mathcal{F}^C_{t(u,\bullet)}} y_v.$$

Then we can index columns of $A^{1,\dots,\ell}$ by (u,s) where $u \in [\beta]$ and $s \in \{0, \dots, x_u Y_u - 1\}$.

Let d be an m-dimensional normalized data vector. Let $p^*_{(u,s)}(d)$ denote the (u,s) component of the MLE for d in $\mathcal{M}(A^{1,\dots,\ell})$. Let $\tilde{d}$ be the data vector of $\tilde{A}^{1,\dots,\ell}$ with components

$$\tilde{d}_{(u,s')} = \sum_{\alpha_v^C \in \mathcal{F}^C_{t(u,\bullet)}} \sum_{n=0}^{y_v - 1} d_{(u,v,s' y_v + n)}$$

for each $u \in [\beta]$ and $s' \in \{0, \ldots, x_u - 1\}$.

Proposition 3.16. *Let d be a data vector for $\mathcal{M}(A^{1,\ldots,\ell})$ and let $\tilde{d}$ be as described above. Then the MLE for $\tilde{d}$ in $\mathcal{M}(\tilde{A}^{1,\ldots,\ell})$ has (u, s') component*

$$Y_u \cdot p^*_{(u,0)}(d).$$

Proof. Let q^* be the vector with (u, s') component equal to $Y_u \cdot p^*_{(u,0)}(d)$ for each $u \in [\beta]$, $s' \in \{0, \ldots, x_u - 1\}$. First, we show that $\tilde{A}^{1,\ldots,\ell}q^* = \tilde{A}^{1,\ldots,\ell}\tilde{d}$. Let $\tilde{\alpha}$ be a row of $\tilde{A}^{1,\ldots,\ell}$ and let α be the corresponding row in $A^{1,\ldots,\ell}$. Then we have

$$\tilde{\alpha} \cdot q^* = \sum_{(u,s')\in \mathrm{supp}(\tilde{\alpha})} Y_u \cdot p^*_{(u,0)}(d).$$

Note that if $(u, s') \in \mathrm{supp}(\tilde{\alpha})$ for some s', then $(u, h) \in \mathrm{supp}(\tilde{\alpha})$ for all $h \in \{0, \ldots, x_u - 1\}$. Moreover, $p^*_{(u,s)} = p^*_{(u,0)}$ for all $s \in \{0, \ldots, x_uY_u - 1\}$. Thus

$$\begin{aligned}
\tilde{\alpha} \cdot q^* &= \sum_{u:(u,0)\in \mathrm{supp}(\tilde{\alpha})} x_uY_u \cdot p^*_{(u,0)}(d) \\
&= \sum_{u:(u,0)\in \mathrm{supp}(\tilde{\alpha})} \sum_{n=0}^{x_uY_u-1} p^*_{(u,n)}(d) \\
&= \alpha \cdot p^*.
\end{aligned}$$

Similarly, we have that

$$\begin{aligned}
\tilde{\alpha} \cdot \tilde{d} &= \sum_{(u,s')\in \mathrm{supp}(\alpha)} \tilde{d}_{(u,s')} \\
&= \sum_{(u,s')\in \mathrm{supp}(\alpha)} \Big(\sum_{n=0}^{x_u-1} d_{s'Y_u+n}\Big) \\
&= \sum_{u:(u,0)\in \mathrm{supp}(\alpha)} \sum_{n=0}^{x_uY_u} d_{(u,n)} \\
&= \sum_{(u,n)\in \mathrm{supp}(\alpha)} d_{(u,n)} \\
&= \alpha \cdot d
\end{aligned}$$

By Birch's theorem, Proposition 2.9, $\alpha \cdot p^* = \alpha \cdot d$. Hence $\tilde{\alpha} \cdot q^* = \tilde{\alpha} \cdot \tilde{d}$, as needed.

Now we must argue that q^* lies in $I(\tilde{A}^{1,\ldots,\ell})$. Let $\bar{A}^{1,\ldots,\ell}$ denote the matrix obtained from $A^{1,\ldots,\ell}$ by removing all repeated columns. The columns of $\bar{A}^{1,\ldots,\ell}$ are indexed by $u \in [\beta]$. Let J denote the inclusion of $I(\bar{A}^{1,\ldots,\ell})$ into the polynomial ring,

$$\mathbb{C}[p_{(u,s)} \mid u \in [\beta], s \in \{0, \ldots, x_uY_u - 1\}]$$

obtained by mapping p_u to $p_{(u,0)}$. Then by definition of the toric ideal associated to a matrix, we have that

$$I(A^{1,\dots,\ell}) = \langle p_{(u,s)} - p_{(u,n)} \mid u \in [\beta], s, n \in \{0, \dots, x_u Y_u - 1\}\rangle + J.$$

Similarly, let $\tilde{J}$ denote the inclusion of $I(\bar{A}^{1,\dots,\ell})$ into R_B obtained by mapping p_u to $q_{(u,0)}$. Then we have that

$$I(\tilde{A}^{1,\dots,\ell}) = \langle q_{(u,s)} - q_{(u,n)} \mid u \in [\beta], s, n \in \{0, \dots, x_u - 1\}\rangle + \bar{J}.$$

Thus we must check that q^* lies in each of the summands of $I(\tilde{A}^{1,\dots,\ell})$.

First, $q^*_{(u,s)} = q^*_{(u,n)}$ for all $u \in [\beta]$ and $s, n \in \{0, \dots, x_u - 1\}$ by construction. Next, let

$$\tilde{f} = \prod_{n=1}^{N} q_{(u_n,0)} - \prod_{n=1}^{N} q_{(v_n,0)}$$

lie in $\tilde{J}$ and $f = \prod_{n=1}^{N} p_{(u_n,0)} - \prod_{n=1}^{N} p_{(v_n,0)}$ be the corresponding element of J where $N = \deg(\tilde{f}) = \deg(f)$. By 3.14, $\tilde{J}$ is multihomogeneous with respect to the multigrading specified by D. Thus, without loss of generality, we may assume that for each $n \in [N]$, $\alpha^B_{u_n}$ and $\alpha^B_{v_n}$ lie in the same floret of B with respect to $A^{\ell+1}$. Thus $Y_{u_n} = Y_{v_n}$ for all n. Evaluating $\tilde{f}$ at q^* yields

$$\begin{aligned}
\tilde{f}(q^*) &= \prod_{n=1}^{N} q^*_{(u_n,0)} - \prod_{n=1}^{N} q^*_{(v_n,0)} \\
&= \prod_{n=1}^{N} Y_{u_n} \cdot p^*_{(u_n,0)} - \prod_{n=1}^{N} Y_{v_n} \cdot p^*_{(v_n,0)} \\
&= \big(\prod_{n=1}^{N} Y_{u_n}\big)\big(\prod_{n=1}^{N} p^*_{(u_n,0)} - \prod_{n=1}^{N} p^*_{(v_n,0)}\big) \\
&= 0,
\end{aligned}$$

since $p^* \in V(J)$. Thus q^* lies in the closure of $\mathcal{M}(\tilde{A}^{1,\dots,\ell})$. So by Birch's theorem, Proposition 2.9 , q^* is the MLE for $\tilde{d}$ in $\mathcal{M}(\tilde{A}^{1,\dots,\ell})$. □

Definition 3.17. Let A be a partition matrix with k partitions that satisfies the GRIP. The ℓth level **florets** of the matrix A are given by the florets of $A^{\ell+1}$ (as defined in Definition 3.6) in the pair of matrices $B_\ell = \uplus_{n=1}^{\ell} A^n$ and $A^{\ell+1}$, with the convention that the 1st level florets are given by the rows of A^1. We denote the ℓth level floret containing the row α^ℓ_i as $\mathcal{F}^\ell_i$ which implies that the ℓth level floret corresponding to the jth column of A is $\mathcal{F}^\ell_{\mathcal{S}(\ell,j)}$.

Remark 3.18. There is a clear relationship between the notation in Definition 3.17 and the notation used for florets of rows used up until this point. In particular for the pair of matrices B_ℓ and $A^{\ell+1}$ we have that $\mathcal{F}^{\ell+1}_{\mathcal{S}(\ell+1,j)} = \mathcal{F}^{A^{\ell+1}}_{t(\bullet,\mathcal{S}(\ell+1,j))}$ for $1 \leq \ell \leq k-1$. We make this distinction since the function $t(\bullet, \mathcal{S}(\ell+1,j))$ depends on ℓ. As the rows of A correspond to the labels of the edges in the stratified, staged tree $\mathcal{T}_A$, the florets of A correspond exactly to the florets of $\mathcal{T}_A$ (see Remark 4.12).

Corollary 3.19. *Let A be a partition matrix with k partitions that satisfies the GRIP. Then the MLE p^* of d has as its jth coordinate function:*

$$p^*_j = \frac{1}{c_j}\left(\prod_{\ell=1}^{k} \frac{\alpha^\ell_{\mathcal{S}(\ell,j)}(d)}{\sum\limits_{\alpha^\ell_i \in \mathcal{F}^\ell_{\mathcal{S}(\ell,j)}} \alpha^\ell_i(d)}\right). \tag{12}$$

Proof. We induct on k. First, let $k=1$. Then

$$I(A^1) = \langle p_j - p_{j'} \mid \text{ the } j \text{ and } j' \text{ columns of } A^1 \text{ are equal}\rangle.$$

Thus, it is straightforward to check that

$$p^*_j = \frac{1}{c_j}\sum_{i \in I^1_{\mathcal{S}(1,j)}} d_i = \frac{1}{c_j}\alpha^1_{\mathcal{S}(1,j)} \cdot d,$$

as needed.

Now suppose that the result holds for all $k \leq K$ for some natural number $K \geq 1$. Consider a patrition matrix $A^{1,\ldots,K+1}$ that satisfies the GRIP; then by definition of the GRIP so does $A^{1,\ldots,K}$. Now let $B = \Cup_{\ell=1}^{K} A^\ell$, $C = A^{K+1}$, and $D = B \Cap C$. By the GRIP, BC satisfies the floret condition. Recall that by well-connectedness and Lemma 3.13, for each $u \in [\beta]$ and $v \in [\gamma]$, there exist $x_u, y_v \in \mathbb{Z}_+$ such that

$$\omega_{uv} = \begin{cases} x_u y_v & \text{if } t(u,\bullet) = t(\bullet, v) \\ 0 & \text{otherwise.} \end{cases}$$

Let $\tilde{A}^{K+1}$ have columns indexed by (v, s'') where $s'' \in \{0, \ldots, y_v - 1\}$. The (v, s'') column of $\tilde{A}^{K+1}$ is equal to the vth standard basis vector in $\mathbb{Z}^\gamma$. For each v, let

$$X_v = \sum_{u : \alpha^B_u \in \mathcal{F}^B_{t(\bullet,v)}} x_u$$

so that $\sum_{u:\alpha^B_u \in \mathcal{F}^B_{t(\bullet,v)}} \omega_{uv} = X_v y_v$. Let d^C be the vector of data with entries indexed by (v, s'') such that

$$d^C_{(v,s'')} = \sum_{u:\alpha^B_u \in \mathcal{F}^B_{t(\bullet,v)}} \sum_{n=0}^{x_u - 1} u_{(u,v,s''x_u+n)}.$$

Finally, $\tilde{D}$ be the $f \times f$ identity matrix where f is the number of distinct florets of BC. Let d^D be the vector of data indexed by the distinct florets of BC with t component

$$d_t^D = \sum_{u,v:t(u,v)=t} \sum_{s=0}^{\omega_{uv}-1} d_{(u,v,s)}.$$

Let $(p^*)^B$ denote the MLE for $\tilde{d}$ in $\mathcal{M}(\tilde{A}^{1,\dots,K})$. Let $(p^*)^C$ denote the MLE for d^C in $\mathcal{M}(\tilde{A}^{K+1})$. Let $(p^*)^D$ denote the MLE for d^D in $\mathcal{M}(\tilde{D})$. Since $I(A^{1,\dots,K+1}) = I(\tilde{A}^{1,\dots,K}) \times_D I(\tilde{A}^{K+1})$, by Amendola-Kosta-Kubjas theorem, we have that

$$p^*_{(u,v,s)}(d) = \frac{(p^*)^B_{(u,s')}\,(p^*)^C_{(v,s'')}}{(p^*)^D_{t(u,v)}},$$

where s' and s'' are such that $s = s'y_v + s''$. We compute $(p^*)^B, (p^*)^C$ and $(p^*)^D$ to prove the desired result. Fix $u \in [\beta]$ and $v \in [\gamma]$ By induction and Proposition 3.16, we have that

$$\begin{aligned}
(p^*)^B_{(u,s')} &= Y_u \cdot p^*_{(u,v,0)} \\
&= \frac{Y_u}{x_u Y_u}\left(\prod_{\ell=1}^{K} \frac{\alpha^\ell_{\mathcal{S}(\ell,(u,v,0))}(d)}{\sum_{\alpha^\ell_{u'} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,(u,v,0))}} \alpha^\ell_{u'}(d)}\right) \\
&= \frac{1}{x_u}\left(\prod_{\ell=1}^{K} \frac{\alpha^\ell_{\mathcal{S}(\ell,(u,v,s))}(d)}{\sum_{\alpha^\ell_{u'} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,(u,v,s))}} \alpha^\ell_{u'}(d)}\right),
\end{aligned}$$

since $\mathcal{S}(\ell,(u,v,s)) = \mathcal{S}(\ell,(u,v,0))$ for all s. Since $\bar{A}^{K+1}$ is an identity matrix with repeated columns, we have that

$$\begin{aligned}
(p^*)^C_{(v,s'')} &= \frac{1}{y_v}\sum_{n=0}^{y_v-1} d^C_{(v,n)} \\
&= \frac{1}{y_v}\sum_{u:\alpha^B_u \in \mathcal{F}^B_{t(\cdot,v)}} \sum_{n=0}^{\omega_{uv}} d_{(u,v,n)} \\
&= \frac{1}{y_v}(\alpha^{K+1}_v d) \\
&= \frac{1}{y_v}\left(\alpha^{K+1}_{\mathcal{S}(K+1,(u,v,s))} d\right).
\end{aligned}$$

Let $t = t(u, v)$. Finally, since $\bar{D}$ is an identity matrix, we have that

$$\begin{aligned}
(p^*)_t^D &= d_t^D \\
&= \sum_{u',v' : t(u',\bullet)=t(\bullet,v')=t} \sum_{s=0}^{\omega_{u'v'}-1} d_{(u',v',s)} \\
&= \sum_{\alpha_{v'}^{K+1} \in \mathcal{F}_{\mathcal{S}(\ell,(u,v,s))}^{K+1}} \alpha_{v'}^{K+1} d.
\end{aligned}$$

Thus by the theorem from [3], the (u, v, s) coordinate of the MLE for u in $\mathcal{M}(A^{1,\dots,K+1})$ is

$$\begin{aligned}
p^*_{(u,v,s)}(d) &= \frac{1}{x_u} \left(\prod_{\ell=1}^{K} \frac{\alpha^\ell_{\mathcal{S}(\ell,(u,v,s))}(d)}{\sum_{\alpha^\ell_{u'} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,(u,v,s))}} \alpha^\ell_{u'}(d)} \right) \cdot \frac{1}{y_v} (\alpha^{K+1}_{\mathcal{S}(K+1,(u,v,s))} d) \cdot \frac{1}{\sum_{\alpha_{v'}^{K+1} \in \mathcal{F}_{\mathcal{S}(\ell,(u,v,s))}^{K+1}} \alpha_{v'}^{K+1}(d)} \\
&= \frac{1}{x_u y_v} \left(\prod_{\ell=1}^{K+1} \frac{\alpha^\ell_{\mathcal{S}(\ell,(u,v,s))}(d)}{\sum_{\alpha^\ell_{u'} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,(u,v,s))}} \alpha^\ell_{u'}(d)} \right) \\
&= \frac{1}{\omega_{uv}} \left(\prod_{\ell=1}^{K+1} \frac{\alpha^\ell_{\mathcal{S}(\ell,(u,v,s))}(d)}{\sum_{\alpha^\ell_{u'} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,(u,v,s))}} \alpha^\ell_{u'}(d)} \right),
\end{aligned}$$

proving the result for $A^{1,\dots,K+1}$ and Corollary 3.19 by induction. □

We now define an iterated toric fiber product of linear ideals. These are defined so that if $I(A^{1,\dots,k})$ can be constructed as an iterated toric fiber product of models corresponding to partition matrices, then $A^{1,\dots,k}$ satisfies the GRIP.

Definition 3.20. Let $A^1, \dots, A^k$ be partition matrices such that $A^{1,\dots,\ell}$ and $A^{\ell+1}$ satisfy the floret condition for each $\ell \in [k-1]$. Thus, we can let $D_\ell = \left(\Cup_{n=1}^{\ell} A^n \right) \Cap A^{\ell+1}$ for each $\ell \in [k-1]$. Suppose there exist compressions $\tilde{A}^{1,\dots,\ell}$ and $\tilde{A}^{\ell+1}$ of $A^{1,\dots,\ell}$ and $A^{\ell+1}$, respectively, such that $I(\tilde{A}^{1,\dots,\ell}) \times_{D_\ell} I(\tilde{A}^{\ell+1})$ is defined and equal to $I(A^{1,\dots,\ell+1})$ for each $\ell \in [k-1]$. Then the ideal I is an **iterated toric fiber product of linear ideals**.

Proposition 3.21. *Let $A^{1,\dots,k}$ be a partition matrix such that $I(A^{1\dots,k})$ is an iterated toric fiber product of linear ideals. Then $A^{1,\dots,k}$ is well-connected and satisfies the GRIP.*

Proof. Fix $\ell \in [k-1]$. Let $B = \Cup_{n=1}^{\ell} A^n$ and let $C = A^{\ell+1}$. By definition of the iterated toric fiber product, BC satisfies the floret condition. Thus, $D_\ell = B \Cap C$ is well-defined. Let $\tilde{A}^{1,\dots,\ell}$ and $\tilde{A}^{\ell+1}$ be such that $I(\tilde{A}^{1,\dots,\ell}) \times_{D_\ell} I(\tilde{A}^{\ell+1})$ is defined and equal to

$I(A^{1,\dots,\ell+1})$. Then since the toric fiber product is defined, both $I(\tilde{A}^{1,\dots,\ell})$ and $I(\tilde{A}^{\ell+1})$ are multihomogeneous with respect to the multigrading given by D_ℓ. Thus each row of D_ℓ is contained in the rowspans of $A^{1,\dots,\ell}$ and $A^{\ell+1}$.

Let $\tilde{A}^{1,\dots,\ell}$ have β distinct columns with labels $1,\dots,\beta$. Let x_u be the number of copies of column u in $\tilde{A}^{1,\dots,\ell}$. Let y_v be the number of copies of the vth standard basis vector in $\tilde{A}^{\ell+1}$. Then by the construction of the toric fiber product, there are $\omega_{uv} = x_u y_v$ columns of $A^{1,\dots,\ell+1}$ that consist of column u of $\tilde{A}^{1,\dots,\ell}$ concatenated with e_v.

Let $u, u' \in [\beta]$ be such that α_u^B and $\alpha_{u'}^B$ belong to the same floret of B. Let α_v^C belong to the corresponding floret of C and let $\mathcal{F}$ denote the set of all rows of C in the same floret as α_v^C. Then the number of copies of column u in $A^{1,\dots,\ell}$ is $x_u \sum_{v':\alpha_{v'}^C \in \mathcal{F}} y_{v'}$. Similarly, the number of copies of column u' in $A^{1,\dots,\ell}$ is $x_{u'} \sum_{v':\alpha_{v'}^C \in \mathcal{F}} y_{v'}$. Thus we have

$$\omega_{uv}/\big(x_u \sum_{v':\alpha_{v'}^C \in \mathcal{F}} y_{v'}\big) = y_v/\big(\sum_{v':\alpha_{v'}^C \in \mathcal{F}} y_{v'}\big) = \omega_{u'v}/\big(x_{u'} \sum_{v':\alpha_{v'}^C \in \mathcal{F}} y_{v'}\big).$$

Thus BC is well-connected. The above argument holds for all $\ell \in [k-1]$. Thus $A^{1,\dots,k}$ satisfies the GRIP. $\square$

We now show that for partition matrices that satisfy the GRIP, the IPS algorithm produces the MLE in exactly one cycle.

Lemma 3.22. *If A is a partition matrix with k partitions that satisfies the GRIP given in Definition 3.10 then the following equation holds for any row vector α_i^k of A^k:*

$$\sum_{j \in I_i^k} \frac{1}{c_j^{k-1}} \left(\prod_{\ell=1}^{k-1} \frac{\alpha^\ell_{\mathcal{S}(\ell,j)}(d)}{\sum\limits_{\alpha^\ell_{i'} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,j)}} \alpha^\ell_{i'}(d)} \right) = C_i^k \sum_{\alpha^k_{i'} \in \mathcal{F}_i^k} \alpha^k_{i'}(d).$$

Proof. By Corollary 3.19, the MLE of $\mathcal{M}_A$ is given by $p = (p_1, \dots, p_m)$ where

$$p_j = \frac{1}{c_j} \left(\prod_{\ell=0}^{k} \frac{\alpha^\ell_{\mathcal{S}(\ell,j)}(d)}{\sum\limits_{\alpha^\ell_{i} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,j)}} \alpha^\ell_{i}(d)} \right),$$

for $j \in \{1, \dots, m\}$. Since p is the MLE it also satisfies $Ap = Ad$ (see Proposition 2.9). In particular for any row vector α_i^k we have that

$$\alpha_i^k(p) = \sum_{j \in I_i^k} \frac{1}{c_j} \left(\prod_{\ell=1}^{k} \frac{\alpha^\ell_{\mathcal{S}(\ell,j)}(d)}{\sum\limits_{\alpha^\ell_{i'} \in \mathcal{F}^\ell_{\mathcal{S}(\ell,i)}} \alpha^\ell_{i'}(d)} \right) = \alpha_i^k(d).$$

Note that for any $j \in I_i^k$, we have $(1/c_j) = (c_j^{k-1}/c_j^{k-1}c_j) = (1/C_i^k c_j^{k-1})$ since A is well-connected. Thus we can pull out a factor of $(1/C_i^k)$ out of each term. Additionally for any $j \in I_i^k$, we have that $\mathcal{S}(k,j) = i$ by definition, and hence the middle term of the above equation can be rewritten as

$$\frac{\alpha_i^k(d)}{C_i^k \sum_{\alpha_{i'}^k \in \mathcal{F}_j^k} \alpha_{i'}^k(d)} \left(\sum_{j \in I_i^k} \frac{1}{c_j^{k-1}} \left(\prod_{\ell=1}^{k-1} \frac{\alpha_{\mathcal{S}(\ell,j)}^\ell(d)}{\sum_{\alpha_{i'}^\ell \in \mathcal{F}_{\mathcal{S}(\ell,i)}^\ell} \alpha_{i'}^\ell(d)} \right) \right).$$

The result immediately follows from this . $\square$

Theorem 3.23. *If A is a partition matrix with k partitions that satisfies the GRIP given in Definition 3.10 then the IPS algorithm results in the MLE after one cycle.*

Proof. We proceed by induction on k. For $k = 1$, the IPS produces the MLE of $\mathcal{M}_A$ after just one step. Indeed the information projection onto the linear family defined by A in this case is simply the MLE of $\mathcal{M}_A$ (see (7) and Proposition 2.9). Now assume that for any partition matrix A satisfying the GRIP where $k \leq K$ for some natural number $K \geq 1$, the IPS algorithm results in the MLE. By Corollary 3.19 p^k is of the form in (12) in this case.

Now let $A^{1,\ldots,K+1}$ be a partition matrix with $K+1$ partitions satisfying the GRIP. Note that based on the definition of the GRIP, $A^{1,\ldots,K}$ also satisfies the GRIP.

Performing the first K steps of the IPS algorithm on A is equivalent to one cycle on $A^{1,\ldots,K}$. Thus we have, by the induction hypothesis and Corollary 3.19, that p^K is the MLE of $\mathcal{M}_{A^{1,\ldots,K}}$, i.e.

$$p_j^K = \frac{1}{c_j^K} \left(\prod_{\ell=1}^{K} \frac{\alpha_{\mathcal{S}(\ell,j)}^\ell(d)}{\sum_{\alpha_i^\ell \in \mathcal{F}_{\mathcal{S}(\ell,j)}^\ell} \alpha_i^\ell(d)} \right).$$

Then the $K+1$-th step of IPS scales each p_j^K by

$$\frac{\alpha_{\mathcal{S}(K+1,j)}^{K+1}(d)}{\sum_{j' \in I_{\mathcal{S}(K+1,j)}^{K+1}} \frac{1}{c_{j'}^K} \left(\prod_{\ell=1}^{K} \frac{\alpha_{\mathcal{S}(\ell,j')}^\ell(d)}{\sum_{\alpha_i^\ell \in \mathcal{F}_{\mathcal{S}(\ell,j')}^\ell} \alpha_i^\ell(d)} \right)} = \frac{\alpha_{\mathcal{S}(K+1,j)}^{K+1}(d)}{C_{\mathcal{S}(K+1,j)}^k \sum_{\alpha_i^{K+1} \in \mathcal{F}_{\mathcal{S}(K+1,j)}^{K+1}} \alpha_i^{K+1}(d)},$$

where equality follows from Lemma 3.22. Since $C_{\mathcal{S}(K+1,j)}^{K+1} = \frac{c_j}{c_j^K}$ this implies that p^{K+1} is exactly the MLE of $\mathcal{M}_{A^{1,\ldots,K+1}}$ by Proposition 3.19. By induction this proves the result. $\square$

Remark 3.24. In order to prove the one-cycle convergence for GRIP in Theorem 3.23 we took three important steps, which are similar to the structure of the proof of one-cycle convergence in the case of the RIP in [25] Theorem 5.3. There the author first proves in Lemma 5.8 that the MLE can be written as a normalized product of the MLE for two submodels. We use a generalization of this result from [3] in Lemma 3.22 to show the structure of the MLE. This corresponds to Theorem 5.1 in [25]. Finally, in Theorem 3.23 and Theorem 5.3 in [25], respectively, the now known structure of the MLE is used in the induction to show one-cycle convergence.

Not only does the proof of Theorem 3.23 show that the IPS algorithm constructs the MLE in one cycle, but also that at the ℓ-th step the vector p^ℓ is exactly the MLE of the $A^{1,\dots,\ell}$. It is natural to ask if the property of producing the MLE in one cycle also characterizes partition matrices satisfying the GRIP.

However, we can immediately see that this is not the case in general. Consider a $n \times m$ partition matrix A with k partitions such that A^k is the identity matrix I_m. Then $\mathcal{M}_A$ is the entire simplex Δ_{m-1} and IPS automatically produces vector d (the MLE of $\mathcal{M}_A$ in this case) regardless of the first $k-1$ partitions.

Remark 3.25. For a partition matrix A with $k = 2$, it is possible to show that if the IPS algorithm produces the MLE in one cycle, then A satisfies the GRIP. This relies on being able to explicitly write the coordinate functions of the MLE as a product of linear forms that must satisfy the Horn uniformization [19, 30]. While the above counter-example implies that this is not the case for $k = 3$, it would be interesting to investigate whether requiring IPS to produce the MLE of $A^{1,\dots,\ell}$ at the ℓ-th step is sufficient to guarantee that A satisfies the GRIP.

Example 3.26. Here we apply the IPS to the matrix A from Example 3.12.

$$A = \begin{array}{c} \left(\begin{array}{cccccccccccccc} 1 & 1 & 1 & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \\ \hline 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & 1 \end{array}\right) \end{array} \quad \begin{array}{c} \alpha_1^1 \\ \alpha_2^1 \\ \hline \alpha_1^2 \\ \alpha_2^2 \\ \hline \alpha_1^3 \\ \alpha_2^3 \\ \alpha_3^3 \\ \alpha_4^3 \\ \alpha_5^3 \end{array}$$

Let $d = (d_1, \dots, d_{14})$ be the normalized data vector. Projecting to the first partition of A leads to:

$$p_1^0 = \dots = p_7^0 = \frac{1}{7}\,\alpha_1^1(d), \qquad p_8^0 = \dots = p_{14}^0 = \frac{1}{7}\,\alpha_2^1(d)$$

The second step of the algorithm results in four different types of indices. These are given by the different rows in the matrix $A^1 \uplus A^2$, indicated by the different dashed and dotted lines below the matrix.

$$\begin{array}{ll} A^1 \Cap A^2 = \begin{pmatrix} 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 & 1 \end{pmatrix} & \alpha_1^{1\Cap 2} \end{array}$$

$$\begin{array}{ll} A^1 \Cup A^2 = \begin{pmatrix} 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \end{pmatrix} & \begin{array}{l} \alpha_1^{1\Cup 2} \\ \alpha_2^{1\Cup 2} \\ \alpha_3^{1\Cup 2} \\ \alpha_4^{1\Cup 2} \end{array} \end{array}$$

Note that since $A^1 \Cup A^2$ consists only of the ones vector and d is normalized, we have $\alpha_1^{1\Cap 2}(d) = 1$. Therefore the second projection simplifies in the following way:

$$\begin{aligned}
p_1^1 = p_2^1 = p_3^1 &= \frac{1}{C_1^1}\alpha_1^1(d)\frac{\alpha_1^2(d)}{C_1^2\,\alpha_1^{1\Cap 2}(d)} = \frac{1}{3}\alpha_1^1(d)\alpha_1^2(d) \\
p_4^1 = p_5^1 = p_6^1 = p_7^1 &= \frac{1}{C_1^1}\alpha_1^1(d)\frac{\alpha_2^2(d)}{C_2^2\,\alpha_1^{1\Cap 2}(d)} = \frac{1}{4}\alpha_1^1(d)\alpha_2^2(d) \\
p_8^1 = p_9^1 = p_{10}^1 &= \frac{1}{C_2^1}\alpha_2^1(d)\frac{\alpha_1^2(d)}{C_1^2\,\alpha_1^{1\Cap 2}(d)} = \frac{1}{3}\alpha_2^1(d)\alpha_1^2(d) \\
p_{11}^1 = p_{12}^1 = p_{13}^1 = p_{14}^1 &= \frac{1}{C_2^1}\alpha_2^1(d)\frac{\alpha_2^2(d)}{C_2^2\,\alpha_1^{1\Cap 2}(d)} = \frac{1}{4}\alpha_2^1(d)\alpha_2^2(d)
\end{aligned}$$

For the last projection we only demonstrate four indices as an example.

$$\begin{array}{ll} (A^1 \Cup A^2) \Cap A^3 = \begin{pmatrix} 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 \end{pmatrix} & \begin{array}{l} \alpha_1^{(1\Cup 2)\Cap 3} \\ \alpha_2^{(1\Cup 2)\Cap 3} \end{array} \end{array}$$

In this case $A^2 = (A^1 \Cup A^2) \Cap A^3$, hence $\alpha_i^2(d) = \alpha_i^{(1\Cup 2)\Cap 3}(d)$. In order to visualize the structure created by the IPS, we also display the general formula of p_j^2 without simplifications

$$\begin{aligned}
p_1^2 &= \frac{1}{C_1^1}\alpha_1^1(d)\frac{\alpha_1^2(d)}{C_1^2\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_1^3(d)}{C_1^3\alpha_1^{(1\Cup 2)\Cap 3}(d)} = \frac{1}{7}\alpha_1^1(d)\frac{\alpha_1^2(d)}{\frac{3}{7}\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_1^3(d)}{\frac{1}{3}\alpha_1^{(1\Cup 2)\Cap 3}(d)} = \alpha_1^1(d)\alpha_1^3(d) \\
p_4^2 &= \frac{1}{C_1^1}\alpha_1^1(d)\frac{\alpha_2^2(d)}{C_2^2\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_4^3(d)}{C_4^3\alpha_2^{(1\Cup 2)\Cap 3}(d)} = \frac{1}{7}\alpha_1^1(d)\frac{\alpha_2^2(d)}{\frac{4}{7}\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_4^3}{\frac{2}{4}\alpha_2^{(1\Cup 2)\Cap 3}(d)} = \frac{1}{2}\alpha_1^1(d)\alpha_4^3(d) \\
p_8^2 &= \frac{1}{C_2^1}\alpha_2^1(d)\frac{\alpha_1^2(d)}{C_1^2\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_1^3(d)}{C_1^3\alpha_1^{(1\Cup 2)\Cap 3}(d)} = \frac{1}{7}\alpha_2^1(d)\frac{\alpha_1^2(d)}{\frac{3}{7}\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_1^3(d)}{\frac{1}{3}\alpha_1^{(1\Cup 2)\Cap 3}(d)} = \alpha_2^1(d)\alpha_1^3(d) \\
p_{11}^2 &= \frac{1}{C_2^1}\alpha_2^1(d)\frac{\alpha_2^2(d)}{C_2^2\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_4^3(d)}{C_4^3\alpha_2^{(1\Cup 2)\Cap 3}(d)} = \frac{1}{7}\alpha_2^1(d)\frac{\alpha_2^2(d)}{\frac{4}{7}\,\alpha_1^{1\Cap 2}(d)}\frac{\alpha_4^3}{\frac{2}{4}\alpha_2^{(1\Cup 2)\Cap 3}(d)} = \frac{1}{2}\alpha_1^1(d)\alpha_4^3(d)
\end{aligned}$$

For the indices 4 and 11 there exist two identical columns, respectively. Hence the connection ratios do not add up to one, but to $\frac{1}{2}$.

We end this section by noting that the formula for the MLE of a matrix satisfying the GRIP in (12) factors through the associated monomial map ϕ_A. In particular we have that

$$p^\star(d) = \phi_A\left(\frac{s_1^1(d)}{C_1^1}, \ldots, \frac{s_{n_1}^1(d)}{C_{n_1}^1}, \frac{s_1^2(d)}{C_1^2}, \ldots, \frac{s_{n_k}^k(d)}{C_{n_k}^k}\right)$$

where $s_i^\ell(d) = \frac{\alpha_i^\ell(d)}{\sum_{\alpha_{i'}^\ell \in \mathcal{F}_i^\ell} a_i^\ell(d)}$. This factorization implies a nice correspondence between the MLE of the matrix $\bar{A}$ obtained by removing all repeated columns of A and the MLE of A. Suppose that $\bar{A}$ is of size $n \times m$ and let $\bar{d} = (\bar{d}_1, \ldots, \bar{d}_m)$ by a normalized data vector of counts. Re-ordering if necessary, we can denote d in the ambient space of $\mathcal{M}_A$ as

$$d = (d_1^1, \ldots, d_1^{c_1}, d_2^1, \ldots, d_2^{c_2}, \ldots, d_m^{c_m})$$

where c_j is the column weight, or number of repetitions of the column, of the jth column of $\bar{A}$ (see Definition 2.7).

Proposition 3.27. *If A is a partition matrix with k partitions that satisfies the GRIP given in Definition 3.10 then the MLE of $\mathcal{M}_{\bar{A}}$ is given by*

$$p_{\bar{A}}^*(d) = \phi_{\bar{A}}\left(s_1^1(\bar{d}), \ldots, s_{n_1}^1(\bar{d}), s_1^2(\bar{d}), \ldots, s_{n_k}^k(\bar{d})\right)$$

where $s_i^\ell(\bar{d}) = \frac{\bar{\alpha}_i^\ell(\bar{d})}{\sum_{\bar{\alpha}_{i'}^\ell \in \mathcal{F}_i^\ell} \bar{\alpha}_{i'}^\ell(\bar{d})}$ for row vectors $\bar{\alpha}_i^\ell$ of $\bar{A}$.

Proof. Clearly $p_{\bar{A}}^*(d)$ lies on the model $\mathcal{M}_{\bar{A}}$; it remains to be shown that $A(p_{\bar{A}}^*(d)) = A(d)$ which implies that $p_{\bar{A}}^*(d)$ is the MLE of $\mathcal{M}_{\bar{A}}$ by Proposition 2.9. Since A satisfies the GRIP the MLE of $\mathcal{M}_A$, denoted $p_A^*(d)$, is given in (12). We can write the coordinate functions of $p_A^*(d)$ in terms of the coordinate functions of $p_{\bar{A}}^*(d)$:

$$p_A^*(d)_j^i = \frac{\prod_{\ell=1}^k s_{\mathcal{S}(\ell,j)}^\ell(d)}{c_j} = \frac{\prod_{\ell=1}^k s_{\mathcal{S}(\ell,j)}^\ell(\bar{d})|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}}}{c_j} = \frac{p_{\bar{A}}(\bar{d})_j|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}}}{c_j}.$$

where $p_A^*(d)_j^i$, $1 \le i \le c_j$, is any coordinate function of $p_A^*(d)$ corresponding to the j-th column of $\bar{A}$. Thus we have that

$$\sum_{i=1}^{c_j} p_A^*(d)_j^i = p_{\bar{A}}(\bar{d})_j|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}},$$

which implies that

$$\alpha_i^\ell(p_A^*(d)) = \sum_{j=1}^m \sum_{i=1}^{c_j} p_A^*(d)_j^i = \sum_{j=1}^m p_{\bar{A}}^*(\bar{d})_j|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}} = \bar{\alpha}_i^\ell(p_{\bar{A}}^*(\bar{d}))|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}}.$$

In addition, for any row of $\bar{\alpha}_i^\ell$ of $\bar{A}$, it is easy to see that

$$\alpha_i^\ell(d) = \bar{\alpha}_i^\ell(\bar{d})|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}},$$

Putting the above two equalities together, along with the fact that $p_A^*(d)$ is the MLE of $\mathcal{M}_A$, we have that

$$\bar{\alpha}_i^\ell(\bar{d})|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}} = \alpha_i^\ell(d) = \alpha_i^\ell(p_A(d)) = \bar{\alpha}_i^\ell(p_{\bar{A}}(\bar{d}))|_{\bar{d}_j = d_j^1 + d_j^2 + \ldots + d_j^{c_j}},$$

which implies that $\bar{\alpha}_i^\ell(\bar{d}) = \bar{\alpha}_i^\ell(p_{\bar{A}}(\bar{d}))$. □

Example 3.28. In Proposition 3.27 we see that the partition matrix obtained by removing all repeated columns from a partition matrix that satisfies the GRIP also satisfies the GRIP. The reverse is, however, not true. Consider the independence model from Example 2.10. We see in that example that the IPS algorithm produces the MLE in one cycle. One can easily check that this model satisfies the GRIP as well, so this is not surprising.

However, if we simply repeat the last column to obtain a new partition matrix, shown here to the right, then the IPS algorithm does not produce the MLE after one cycle. In fact, one can check (by computing the solution set to the polynomial equations defining the ML-degree) that the resulting partition model has ML-degree greater than one, and hence non-rational MLE. This implies that the IPS algorithm *cannot* produce the MLE exactly.

$$\begin{pmatrix} 1 & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & 1 & 1 \\ \hdashline 1 & \cdot & 1 & \cdot & \cdot \\ \cdot & 1 & \cdot & 1 & 1 \end{pmatrix}$$

4 Model families satisfying the GRIP

4.1 Hierarchical models

In this section, we justify our use of the name "generalized running intersection property" in the case of binary hierarchical models. Recall that decomposable implies exists ordering satisfying RIP (see Lemma 2.26).

As discussed in Section 2.5, we can associate a hierarchical model with a simplicial complex Γ. Here we define the matrix A_Γ corresponding to a simplicial complex Γ. The columns of the matrix A_Γ are indexed by subsets of $[n]$. Note that this is the same as being indexed by 0/1 strings of length n by taking each $S \subset [n]$ to be the the set of positions in the string equal to 1.

The rows of A_Γ are divided into blocks $A^1, \ldots, A^k$ each of size $2^{|F|}$. The rows of block A^i are of the form a_S^i where $S \subset F_i$ with

$$a_S^i(T) = \begin{cases} 1, & \text{if } T \cap F_i = S \\ 0, & \text{otherwise,} \end{cases}$$

for each $T \subset [n]$. The *hierarchical model* $\mathcal{M}_\Gamma$ can also be defined as the closure of the image of the monomial map defined by A_Γ intersected with the probability simplex, Δ_{2^n}.

Proposition 4.1. *Let* Γ *be a simplicial complex and let* A *be a partition matrix representing the binary hierarchical model on* Γ *that satisfies the running intersection property. Then* A *satisfies the generalized running intersection property as well.*

Remark 4.2. This proposition holds for non-binary hierarchical models as well. For the sake of notational simplicity, we provide the argument for binary hierarchical models below; however, the analogous argument proves the result for non-binary random variables as well.

Proof. Let Γ be a simplicial complex on $[p]$ with facets $F_1, \dots, F_k$ such that the resulting partition matrix $A^1 A^2 \dots A^k$ satisfies the running intersection property. Let $1 \le \ell \le k$, $B = \Cup_{n=1}^{\ell} A^n$ and let $C = A^{\ell+1}$. Without loss of generality, let $F_1, \dots, F_\ell$ form a simplicial complex on $[q]$ for $q \le p$. Since $A^1 A^2 \dots A^k$ satisfies the RIP, there exists an $s \le \ell$ such that $[q] \cap F_{\ell+1} = F_s \cap F_{\ell+1}$.

First, we wish to show that the matrix BC satisfies the floret condition. The matrix B has rows indexed by subsets of $[q]$. For each $S \subset [p]$ and $T \subset [q]$, the S entry of α_T^B is equal to 1 if $T = S \cap [q]$ and 0 otherwise. Let $U \subset F_{\ell+1}$. We claim that α_U^C is connected to α_T^B if and only if $U \cap [q] = T \cap F_{\ell+1}$.

Indeed, suppose that $U \cap [q] = T \cap F_{\ell+1}$. Let $S = U \cap T$. Then

$$S \cap [q] = (U \cap [q]) \cup (T \cap [q]) = (T \cap F_{\ell+1}) \cup T = T.$$

Similarly, $S \cap F_{\ell+1} = U$. Hence the S coordinates of α_T^B and α_U^C are both equal to one, and these rows are connected.

Conversely, suppose that α_T^B and α_U^C are connected. Then there exists an $S \subset [p]$ such that $U = S \cap F_{\ell+1}$ and $T = S \cap [q]$. Intersecting the first equation with $[q]$ and the second with $F_{\ell+1}$ yields that $U \cap [q] = T \cap F_{\ell+1}$, as needed.

Thus, for each $T, T' \subset [q]$, if $T \cap F_{\ell+1} = T' \cap F_{\ell+1}$, then we have that the sets of rows of C connected to α_T^B and $\alpha_{T'}^B$ are equal. Otherwise these sets are disjoint. Hence BC satisfies the floret condition.

For each $S \subset [p]$, the number of columns of B identical to the column associated to S is 2^{p-q}. Indeed, the columns of B associated to S and S' are equal if and only if $S \cap [q] = S' \cap [q]$, and there are 2^{p-q} such subsets of $[p]$. Similarly, the number of columns of BC identical to the column associated to S is $2^{p-q'}$, where $q' = \#([q] \cup F_{\ell+1})$. Thus, since all column weights within B are equal to one another and all column weights within BC are equal to one another, BC is well-connected.

Finally, we must show that $B \Cap C$ lies in the rowspan of $A^{1,\dots,\ell}$. Recall that there exists an $s \le \ell$ such that $[q] \cap F_{\ell+1} = F_s \cap F_{\ell+1}$. We showed above that the rows of $D = B \Cap C$

are indexed by subsets of $[q] \cap F_{\ell+1}$. For each $T \subset [q] \cap F_{\ell+1}$ and each $S \subset [p]$, the row α_T^D has S entry equal to 1 if $S \cap ([q] \cap F_{\ell+1}) = T$ and 0 otherwise.

We claim that for each $T \subset [q] \cap F_{\ell+1}$, α_T^D is the sum of all rows of A^s, α_U^s, such that $U \cap F_{\ell+1} = T$, i.e. α_T^D is equal to

$$\sum_{\substack{U \subset F_s: \\ U \cap F_{\ell+1} = T}} \alpha_U^s. \tag{13}$$

Indeed, let $S \subset [p]$ satisfy $S \cap F_{\ell+1} \cap [q] = T$. Then we have that $(S \cap F_s) \cap F_{\ell+1} = T$. In particular, the term of Equation 13 corresponding to $U = S \cap F_s$ has S entry equal to 1, as needed.

Similarly, if $S \cap F_{\ell+1} \cap [q] = T' \neq T$, then we have $T' = (S \cap F_s) \cap F_{\ell+1}$. In particular, the row $\alpha_{S \cap F_s}^s$ is the row of A^s supported on S. But it is not a term of the sum in Equation 13. Hence the sum in 13 has S entries equal to 0, as needed.

Since each row of D lies in the rowspan of A^s, it lies in the rowspan of $A^1 A^2 \ldots A^\ell$, as needed. Hence $A^1 A^2 \ldots A^k$ satisfies the generalized running intersection properety. $\square$

4.2 Staged tree models

In this section, we explore the implications of the results in Section 3 for staged tree models. We show that a staged tree model is *balanced* and stratified if and only if the associated matrix satisfies the GRIP, leading to a new characterization of balanced, stratified staged trees. In particular for the associated partition matrix, the IPS algorithm results in the MLE after one cycle.

To define balanced first we must introduce the interpolating polynomial of a staged tree $\mathcal{T}$, first introduced in [23]. For any vertex $v \in V$, we can consider the sub-tree graph $\mathcal{T}_v$ rooted at v. We denote the set of root-to-leaf paths of $\mathcal{T}_v$ (or equivalently the v-to-leaf paths in $\mathcal{T}$) as Λ_v.

Definition 4.3. For a staged tree $(\mathcal{T}, \theta)$ and a vertex $v \in V$, define

$$t(v) := \sum_{\lambda \in \Lambda_v} \prod_{s_i \in \theta(\lambda)} s_i.$$

If v_0 is the root of $\mathcal{T}$, then $t(v_0)$ is the **interpolating polynomial** of $\mathcal{T}$.

Definition 4.4. A staged tree $(\mathcal{T}, \theta)$ is **balanced** if for any two vertices $v, w \in V$ such that $\mathcal{F}_v = \mathcal{F}_w$ the following equation

$$t(v')t(w'') = t(w')t(v'') \tag{14}$$

is satisfied for all distinct pairs of vertices $v', v'' \in \mathrm{ch}(v)$ and $w', w'' \in \mathrm{ch}(w)$ where $\theta(v, v') = \theta(w, w')$ and $\theta(v, v'') = \theta(w, w'')$.

Example 4.5. Consider the staged tree $(\mathcal{T}, \theta)$ in Example 2.18. To determine if the tree is balanced we must compare all pairs of vertices in the same stage. The pairs $v_{s_0 t_0}, v_{s_1 t_0}$ and $v_{s_0 t_1}, v_{s_1 t_1}$ trivially satisfy the condition in (14). Thus to determine whether $\mathcal{T}$ is balanced it is enough to check whether (14) for v_{s_0} and v_{s_1}. We have that

$$\mathrm{ch}(v_{s_0}) = \{v_{s_0 t_0}, v_{s_0 t_1}\}, \quad \mathrm{ch}(v_{s_1}) = \{v_{s_1 t_0}, v_{s_1 t_1}\},$$

and hence we have to check that $t(v_{s_0 t_0})t(v_{s_1 t_1}) = t(v_{s_0 t_1})t(v_{s_1 t_0})$. Writing out the interpolating polynomials,

$$t(v_{s_0 t_0}) = t(v_{s_1 t_0}) = r_0 + r_1 + r_2, \quad t(v_{s_0 t_1}) = t(v_{s_1 t_1}) = r_3 + r_4,$$

we see that this is satisfied. The staged tree $(\mathcal{T}, \theta)$ additionally satisfies the property that for any two vertices v, w in the same stage, $t(v) = t(w)$. In [4, Lemma 2.13] it was shown that this property implies balanced. To "unbalance" the staged tree, one could re-color as blue one of the red vertices, changing the resulting floret. For another example of an unbalanced staged tree see [4, Example 2.14].

One can determine statistical and algebraic properties of a staged tree model from its interpolating polynomial [22, 23] and when it is balanced [4, 18]. In particular the following proposition follows from [18, Thm. 10] and applies to our definition of staged tree models due to [4, Lem. 5.13]

Proposition 4.6. *A staged tree model $\mathcal{M}_\mathcal{T}$ is equal to the toric model $\overline{\mathcal{M}_\mathcal{T}}$ if and only if $(\mathcal{T}, \theta)$ is balanced.*

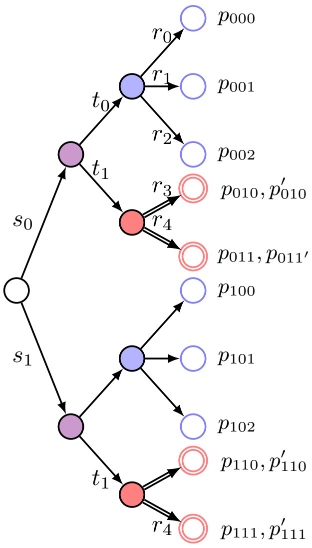
Example 4.7. For the staged tree $(\mathcal{T}, \theta)$ introduced in Example 2.18, we can define the toric model $\overline{\mathcal{M}_\mathcal{T}}$ by the monomial map $\phi_\mathcal{T}$. This is equivalent to monomial map $\phi_{A_\mathcal{T}}$ arising from the partition matrix $A_\mathcal{T}$ on the right.
Thus $\overline{\mathcal{M}_\mathcal{T}}$ is the log-linear model $\mathcal{M}_{A_\mathcal{T}}$. In Example 4.5 we showed that $(\mathcal{T}, \theta)$ is a balanced staged tree, and hence Proposition 4.6 implies that $\mathcal{M}_\mathcal{T} = \overline{\mathcal{M}_\mathcal{T}} = \mathcal{M}_{A_\mathcal{T}}$.

$$A_\mathcal{T} = \begin{array}{cl} \left(\begin{array}{cccccccccc} 1 & 1 & 1 & 1 & 1 & \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot & 1 & 1 & 1 & 1 & 1 \\ \hline 1 & 1 & 1 & \cdot & \cdot & 1 & 1 & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & 1 & \cdot & \cdot & \cdot & 1 & 1 \\ \hline 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot \\ \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot \\ \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot \\ \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 & \cdot \\ \cdot & \cdot & \cdot & \cdot & 1 & \cdot & \cdot & \cdot & \cdot & 1 \end{array}\right) & \begin{array}{l} s_0 \\ s_1 \\ t_0 \\ t_1 \\ r_0 \\ r_1 \\ r_2 \\ r_3 \\ r_4 \end{array} \\ \begin{array}{cccccccccc} p_{000} & p_{001} & p_{002} & p_{010} & p_{011} & p_{100} & p_{101} & p_{102} & p_{110} & p_{111} \end{array} & \end{array}$$

In the case that $(\mathcal{T}, \theta)$ is stratified and balanced, the staged tree model $\mathcal{M}_\mathcal{T}$ is also a partition model. Since the tree is stratified the associated matrix $A_\mathcal{T}$ with the monomial map $\phi_\mathcal{T}$ (defined in (10)) is a partition matrix, and being balanced implies that $\mathcal{M}_\mathcal{T} = \overline{\mathcal{M}_\mathcal{T}} = \mathcal{M}_{A_\mathcal{T}}$.

Note that the set of models defined by matrices whose associated tree $A_\mathcal{T}$ is a balanced, stratified staged tree is slightly larger than the set of staged tree models arising from balanced, stratified staged trees. When the matrix A has repeated columns, this information gets lost when constructing the tree graph $\mathcal{T}_A$. Thus the staged tree model arising from $(\mathcal{T}_A, \theta_A)$ is the model obtained by removing repeated columns from a staged matrix A.

Example 4.8. Consider the matrix A from Example 3.4, which is also used throughout Section 3. On the left the associated tree $(\mathcal{T}_A, \theta_A)$ is drawn and on the right we color portions of the matrix corresponding to the blue and red florets. We draw double circles around four of the leaves to indicate that the associated column is repeated in the matrix A. However, the underlying labelled tree graph is exactly the same as in Examples 2.18 and 4.5, and hence is a stratified and balanced staged tree.



$$A = \begin{array}{c} \\ \left(\begin{array}{cccccccccccccc} 1&1&1&1&1&1&1&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&1&1&1&1&1&1&1 \\ \hline 1&1&1&\cdot&\cdot&\cdot&\cdot&1&1&1&\cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&\cdot&1&1&1&1&\cdot&\cdot&\cdot&1&1&1&1 \\ \hline 1&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&1&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot \\ \cdot&1&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&1&\cdot&\cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&1&\cdot&\cdot&\cdot&\cdot&\cdot&\cdot&1&\cdot&\cdot&\cdot&\cdot \\ \cdot&\cdot&\cdot&1&\cdot&1&\cdot&\cdot&\cdot&\cdot&1&\cdot&1&\cdot \\ \cdot&\cdot&\cdot&\cdot&1&\cdot&1&\cdot&\cdot&\cdot&\cdot&1&\cdot&1 \end{array}\right) \end{array} \begin{array}{c} E \\ s_0 \\ s_1 \\ t_0 \\ t_1 \\ r_0 \\ r_1 \\ r_2 \\ r_3 \\ r_4 \end{array}$$

$$\mathcal{F}_{v_{s_0 t_0}} = \mathcal{F}_{v_{s_1 t_0}} = \{r_0, r_1, r_2\}$$

$$\mathcal{F}_{v_{s_0 t_1}} = \mathcal{F}_{v_{s_1 t_1}} = \{r_3, r_4\}$$

We show that the stratified and balanced conditions on a staged tree are related to the generalized running intersection property for partition matrices.

Lemma 4.9. *The associated partition matrix for a stratified and balanced staged tree is well-connected.*

Proof. For a balanced and stratified staged tree we have that $\mathcal{M}_A = \mathcal{M}_{\mathcal{T}_A}$ and each column of A represents a distinct root-to-leaf path λ_j in $\mathcal{T}_A$. Consider an edge (v, v') of $\mathcal{T}_A$ with label s_i^ℓ, $1 \le \ell \le m$. Then there exists a subset of $I \subset I_i^\ell \subset \{1, \dots, m\}$ representing the set of root-to-leaf paths containing (v, v'), i.e.

$$\{\lambda \in \Lambda \,|\, (v, v') \in E(\lambda)\} = \{\lambda_j \,|\, j \in I \subset I_i^\ell\}.$$

The number of these paths, $|I|$ is exactly c_j^ℓ for any $j \in I$ since c_j^ℓ describes the number of repeated columns of the first ℓ partitions each of which is a distinct column of A. Thus the number of terms in the summation

$$t(v') = \sum_{\lambda \in \Lambda_v} \prod_{s \in \theta(\lambda)} s = \sum_{\lambda_j \,|\, j \in I} \prod_{s \in \theta(\lambda_j)} s$$

is c_j^ℓ. In other words if we denote $\sigma(t(v'))$ as the sum of the coefficients of $t(v')$, then $\sigma(t(v')) = c_j^\ell$. Note that, for any $v, w \in V$, since $t(v)$ is a polynomial with positive integers coefficients, $\sigma(t(v)t(w)) = \sigma(t(v))\sigma(t(w))$.

Now fix a row α_i^ℓ in the ℓ-th partition of A and an index $j \in I_i^\ell$. Let (v, v') be the edge in $\mathcal{T}_a$ such that $\theta(v, v') = s_i^\ell$ and $(v, v') \in \theta(\lambda_j)$. Similarly let $\tilde{j} \in I_i^\ell$ be any other index in I_i^ℓ and (w, w') be the edge such that $\theta(w, w') = s_i^\ell$ and $(w, w') \in \theta(\lambda_{\tilde{j}})$. Since $\mathcal{T}_A$ is balanced, for any edges $(v, v'') \in E(v)$ and $(w, w'') \in E(w)$ we have

$$t(v')t(w'') = t(w')t(v'').$$

In particular we can sum over all the vertices in $\mathrm{ch}(v)$ and $\mathrm{ch}(w)$:

$$\begin{aligned} t(v') \sum_{w'' \in \mathrm{ch}(w)} t(w'') &= t(w') \sum_{v'' \in \mathrm{ch}(v)} t(v'') \\ \Rightarrow \quad t(v')t(w) &= t(w')t(v). \end{aligned}$$

Similarly as before, the number of paths containing an edge (v, v'') or (w, w'') for $v'' \in \mathrm{ch}(v), w'' \in \mathrm{ch}(w)$ is $c_j^{\ell-1}$ or $c_{\tilde{j}}^{\ell-1}$ respectively. Thus $\sigma(t(v)) = c_j^{\ell-1}$ and $\sigma(t(w)) = c_{\tilde{j}}^{\ell-1}$. Then the above equality implies that

$$\sigma(t(v')t(w)) = \sigma(t(w')t(v)) \ \Rightarrow \ c_j^\ell c_{\tilde{j}}^{\ell-1} = c_{\tilde{j}}^\ell c_j^{\ell-1}$$

proving the result. □

Theorem 4.10. *The associated partition matrix for a stratified and balanced staged tree satisfies the GRIP.*

Proof. By Lemma 4.9 we know that the associated partition matrix A is well-connected, and by assumption satisfies property (2) of Definition 3.10. Thus it remains to be shown that for each $1 \leq \ell < k$ we have that the rows of $B_\ell \Cap A^{\ell+1}$ lie in the rowspan of $A^{1,\ldots,\ell}$ where $B^\ell = \Cup_{m=1}^{\ell} A^\ell$.

In the proof [4][Theorem 2.5], the authors compute the generators of the toric ideal $I(A^{1,\ldots,\ell})$. In order to accomplish this, they show in [4][Proposition 4.5] that $I(A^{1,\ldots,\ell})$ is multihomogeneous with respect to the grading given by D. In particular, this implies that for all b in the *integer* kernel of $A^{1,\ldots,\ell}$ and all rows α^D of D, $\alpha^D \cdot b = 0$. Since $A^{1,\ldots,\ell}$ is an integer matrix, there exists an integer basis for its kernel. Thus for each row α^D of D, we have $\alpha^D \in (\ker(A^{1,\ldots,\ell}))^\perp = \mathrm{rowspan}(A^{1,\ldots,\ell})$, as needed. □

Thus stratified and balanced staged tree models lie in the family of partition models whose matrix satisfies the GRIP which implies the following corollary.

Corollary 4.11. *For the partition matrix associated with a stratified and balanced staged tree, the IPS algorithm results in the MLE after one cycle.*

Proof. Follows from Theorems 3.23 and 4.10. □

Remark 4.12. Consider the model $\mathcal{M}_\mathcal{T}$ obtained from the stratified staged tree $\mathcal{T}$ with k levels. Denote the florets of $\mathcal{T}$ by $\mathcal{F}_i^\ell$ where $1 \leq \ell \leq k$ indexes the levels of $\mathcal{T}$ and $1 \leq i \leq n_\ell$ indexes the edge labels of the ℓth level of $\mathcal{T}$; thus $\mathcal{F}_i^\ell$ is the floret associated with edge label s_i^ℓ. As discussed in Remark 3.18 there is a one to one correspondence between the florets of $\mathcal{T}$ and of $A_\mathcal{T}$. Let $A_\mathcal{T}$ be the associated partition matrix with $\mathcal{T}$, and d be a data vector. Then by [19, Proposition 11] the MLE of $\mathcal{M}_\mathcal{T}$ is given by

$$p^*(d) = \phi_{A_\mathcal{T}}\left(s_1^1(d), \ldots, s_{n_1}^1(d), s_1^2(d), \ldots, s_{n_k}^k(d)\right) \tag{15}$$

where $s_i^\ell(d) = \dfrac{\alpha_i^\ell(d)}{\displaystyle\sum_{\alpha_{i'}^\ell \in \mathcal{F}_i^\ell} a_{i'}^\ell(d)}$.

We now show that the GRIP is, in some sense, a characterization of stratified and balanced staged trees. All partition matrices that satisfy the GRIP, in fact, have an associated tree that is a stratified and balanced staged tree.

Theorem 4.13. *For any partition matrix A satisfying the GRIP, the associated tree $\mathcal{T}_A$ is a stratified and balanced staged tree.*

Proof. Consider the matrix $\bar{A}$ obtained by removing the repeated columns of A. The associated tree $\mathcal{T}_{\bar{A}}$ is equivalent to $\mathcal{T}_A$, and hence is a stratified staged tree. Since there are no repeated columns, we have that the toric model for the staged tree $\mathcal{T}_A$ is given by the partition model of $\bar{A}$, i.e. $\overline{\mathcal{M}}_{\mathcal{T}_A} = \mathcal{M}_{\bar{A}}$.

By Proposition 4.6, the staged tree model $\mathcal{M}_{\mathcal{T}_A}$ is equal to $\mathcal{M}_{\bar{A}}$ if and only if $\mathcal{T}_A$ is a balanced staged tree. By Proposition 3.27 the MLE of $\mathcal{M}_{\bar{A}}$ is equal to the MLE of the staged tree model $\mathcal{T}_A$ (Remark 4.12). Since the MLE map also parameterizes the model, this implies that

$$\overline{\mathcal{M}}_{\mathcal{T}_A} = \mathcal{M}_{\bar{A}} = \mathcal{M}_{\mathcal{T}_A},$$

and hence that $\mathcal{T}_A$ is balanced. □

Putting the previous results together, we can say that a partition matrix without repeated columns satisfies the GRIP if and only if the associated tree is a stratified and balanced staged tree. We also note that this implies that the IPS algorithm provides another method for determining whether a stratified staged tree is balanced. If the IPS algorithm produces the MLE exactly in one cycle for an arbitrary data vector and, in each case, the MLE is given by (15), then the stratified staged tree is balanced. Otherwise the stratified staged tree is not balanced via Corollary 4.11.

5 Discussion

We conclude with a discussion of some natural questions that arise from this work.

Question 1. For a partition matrix A with k partitions, does requiring that the IPS algorithm produces the MLE for $A^{1,\dots,\ell}$ at each step imply that A must satisfy the GRIP?

In Remark 3.25, we claim that IPS producing the MLE in 2 steps for a partition matrix with 2 partitions is enough to guarantee that the GRIP is satisfied. While the GRIP is sufficient for the algorithm to produce the MLE for $A^{1,\dots,\ell}$ at each step, it is possible that the reverse logical direction holds.

Question 2. Can one always find a representation of a log-linear partition model with rational MLE such that it satisfies the GRIP?

It is clear that the matrix representation of a particular model affects the IPS algorithm. In [12], the authors show sufficient conditions for 2-way quasi-independence models to have rational MLE. Since k-way quasi-independence models are just partition models without repeated columns, it would be interesting to see if these results can be used to show that every such matrix has a representation satisfying the GRIP. Indeed our preliminary results indicate that this is true, but we do not include a formal proof.

Finally there are many results and questions related to the convergence of the generalized IPS algorithm on log-linear models and its connections to tools from algebraic statistics (see [2, Sec. 5] and [17, Sec. 7.3]). We feel that our work falls adjacent to this line of inquiry and that it would be interesting to investigate whether tools that have recently produced results in this area can connect to our work.

Acknowledgments

We would like to thank Bernd Sturmfels, Nihat Ay, Jiří Vomlel, Eliana Duerte, Orlando Marigliano and Carlos Améndola for helpful discussions and suggestions regarding this project. Jane Coons was partially supported by the US National Science Foundation (DGE 1746939). This work was also supported in part by the Max Planck Institute for Mathematics in the Sciences.

References

[1] Améndola, C., Bliss, N., Burke, I., Gibbons, C. R., Helmer, M., Hoşten, S., Nash, E. D., Rodriguez, J. I., and Smolkin, D. The maximum likelihood degree of toric varieties. *J. Symbolic Comput. 92* (2019), 222–242.

[2] Améndola, C., Kohn, K., Reichenbach, P., and Seigal, A. Toric invariant theory for maximum likelihood estimation in log-linear models. *to appear in Algebraic Statistics* (10 2021).

[3] Améndola, C., Kosta, D., and Kubjas, K. Maximum likelihood estimation of toric Fano varieties. *Algebr. Stat. 11*, 1 (2020), 5–30.

[4] Ananiadi, L., and Duarte, E. Gröbner bases for staged trees. *Algebraic Statistics 12*, 1 (Apr. 2021), 1–20.

[5] Ay, N., Jost, J., Lê, H. V., and Schwachhöfer, L. *Information Geometry*, vol. 64. Springer International Publishing, 2017.

[6] Berman, R. J. The sinkhorn algorithm, parabolic optimal transport and geometric monge–ampère equations. *Numerische Mathematik 145* (2020), 771–836.

[7] Brown, J. B., Chase, P. J., and Pittenger, A. O. Order independence and factor convergence in iterative scaling. *Linear Algebra and its Applications 190* (sep 1993), 1–38.

[8] Catanese, F. F., Hosten, S., Khetan, A., and Sturmfels, B. The maximum likelihood degree. *American Journal of Mathematics 128*, 3 (2006), 671–697.

[9] Chang, S., Pierson, E., Koh, P., Geradin, J., Redbird, B., Grusky, D., and Leskovec, J. Mobility network models of covid-19 explain inequities and inform reopening. *Nature 589* (2021), 82–87.

[10] Collazo, R. A., Görgen, C., and Smith, J. Q. *Chain Event Graphs*. CRC Press, Jan. 2018.

[11] Coons, J. I., Marigliano, O., and Ruddy, M. Maximum likelihood degree of the two-dimensional linear gaussian covariance model. *Algebraic Statistics 11*, 2 (Dec. 2020), 107–123.

[12] Coons, J. I., and Sullivant, S. Quasi-independence models with rational maximum likelihood estimator. *Journal of Symbolic Computation 104* (May 2021), 917–941.

[13] Csiszár, I. I-divergence geometry of probability distributions and minimization problems. *The Annals of Probability 3* (1975).

[14] Csiszár, I., and Shields, P. C. *Information Theory and Statistics: A Tutorial*. Foundations and Trends in Communications and Information Theory. Now Publishers Inc, 2004.

[15] Darroch, J., and Ratcliff, D. Generalized iterative scaling for log-linear models. *The Annals of Mathematical Statistics* (1972).

[16] Deming, W. E., and Stephan, F. F. On a least squares adjustment of a sampled frequency table when the expected marginal totals are known. *The Annals of Mathematical Statistics* (1940).

[17] Drton, M., Sturmfels, B., and Sullivant, S. *Lectures on algebraic statistics*, vol. 39. Springer Science & Business Media, 2008.

[18] Duarte, E., and Görgen, C. Equations defining probability tree models. *Journal of Symbolic Computation 99* (July 2020), 127–146.

[19] Duarte, E., Marigliano, O., and Sturmfels, B. Discrete statistical models with rational maximum likelihood estimator. *Bernoulli 27*, 1 (Feb. 2021).

[20] Duarte, E., and Solus, L. A new characterization of discrete decomposable models.

[21] Endo, Y., and Takemura, A. Iterative proportional scaling via decomposable submodels for contingency tables. *Comput. Statist. Data Anal. 53*, 4 (2009), 966–978.

[22] Görgen, C., Bigatti, A., Riccomagno, E., and Smith, J. Q. Discovery of statistical equivalence classes using computer algebra. *International Journal of Approximate Reasoning 95* (Apr. 2018), 167–184.

[23] Görgen, C., and Smith, J. Q. Equivalence classes of staged trees. *Bernoulli 24*, 4A (Nov. 2018), 2676–2692.

[24] Görgen, C., Maraj, A., and Nicklasson, L. Staged tree models with toric structure, 2021.

[25] Haberman, S. J. *The Analysis of Frequency Data.* University of Chicago Press, 1974.

[26] Hauer, M. E., Hardy, R. D., Mishra, D. R., and Pippin, J. S. No landward movement: examining 80 years of population migration and shoreline change in louisiana. *Population and Environment 40* (2019), 369–387.

[27] (https://mathoverflow.net/users/36721/iosif pinelis), I. P. Independent decomposition of coordinate distribution. MathOverflow. URL:https://mathoverflow.net/q/298294 (version: 2018-04-19).

[28] Huh, J. The maximum likelihood degree of a very affine variety. *Compos. Math. 149*, 8 (2013), 1245–1266.

[29] Huh, J. Varieties with maximum likelihood degree one. *J. Algebr. Stat. 5*, 1 (2014), 1–17.

[30] Huh, J., and Sturmfels, B. Likelihood geometry. In *Lecture Notes in Mathematics.* Springer International Publishing, 2014, pp. 63–117.

[31] Lauritzen, S. L. *Graphical Models.* Clarendon Press, 1996.

[32] Pachter, L., and Sturmfels, B., Eds. *Algebraic Statistics for Computational Biology.* Cambridge University Press, New York, 2005, ch. §1.2.

[33] Peyré, G., and Cuturi, M. Computational optimal transport. *Foundations and Trends in Machine Learning 11*, 5-6 (2019), 355–607.

[34] Smith, J. Q., and Anderson, P. E. Conditional independence and chain event graphs. *Artificial Intelligence 172*, 1 (2008), 42–68.

[35] Sturmfels, B., Timme, S., and Zwiernik, P. Estimating linear covariance models with numerical nonlinear algebra. *Algebraic Statistics 11*, 1 (Oct. 2020), 31–52.

[36] Sullivant, S. Toric fiber products. *J. Algebra 316*, 2 (2007), 560–577.

[37] Sullivant, S. *Algebraic Statistics*, vol. 194. American Mathematical Soc., 2018.

[38] Vomel, J. i. Methods of probabilistic knowledge integration, 1999.

[39] Xu, P.-F., Sun, J., and Shan, N. Local computations of the iterative proportional scaling procedure for hierarchical models. *Comput. Statist. Data Anal. 95* (2016), 17–23.

[40] Yiannoutsos, C. T., Halverson, P. K., and Menachemi, N. Bayesian estimation of sars-cov-2 prevalence in indiana by random testing. *Proceedings of the National Academy of Sciences 118*, 5 (2021).