

Matrix Completion for the Independence Model

Kaie Kubjas, Zvi Rosen

October 26, 2014

Abstract

We investigate the problem of completing partial matrices to rank-1 probability matrices. The motivation for studying this problem comes from statistics: A lack of desired completion can provide a falsification test for partial observations to come from the independence model. For each type of partial matrix, we give an inequality in the observed entries which is satisfied if and only if a desired completion exists. We explain how to construct such completions and, in case a partial matrix has more than one rank-1 probability completion, completions that minimize a distance function are studied.

1 Introduction

This paper addresses the following problem:

Problem 1.1. *Given some entries of a matrix, is it possible to add the missing entries so that the matrix has rank 1, is nonnegative, and its entries sum up to one?*

The answer to this question takes many different forms. For example, as we shall prove later, the partial probability matrix

$$\begin{pmatrix} 0.16 & & & \\ & 0.09 & & \\ & & 0.04 & \\ & & & 0.01 \end{pmatrix}$$

has a unique completion:

$$\begin{pmatrix} 0.16 & 0.12 & 0.08 & 0.04 \\ 0.12 & 0.09 & 0.06 & 0.03 \\ 0.08 & 0.06 & 0.04 & 0.02 \\ 0.04 & 0.03 & 0.02 & 0.01 \end{pmatrix}.$$

On the other hand, perturbing any entry of the original matrix by $\epsilon > 0$ makes the matrix have no eligible completions, and perturbing any entry by $\epsilon < 0$ introduces an infinite number of completions.

The motivation for studying Problem 1.1 comes from statistics. Let X and Y be two discrete random variables with m and n states respectively. Define the probability matrix:

$$P = (p_{ij})_{1 \leq i \leq m, 1 \leq j \leq n}, \quad \text{where } p_{ij} = \Pr(X = i, Y = j).$$

For any probability matrix P , we have $p_{ij} \geq 0$ for all i, j and $\sum p_{ij} = 1$. We say that random variables X and Y are independent, if

$$\Pr(X = i, Y = j) = \Pr(X = i) \cdot \Pr(Y = j)$$

for all i, j . This can be translated into the statement

$$P = \begin{pmatrix} \Pr(X = 1) \\ \Pr(X = 2) \\ \vdots \\ \Pr(X = m) \end{pmatrix} (\Pr(Y = 1) \quad \Pr(Y = 2) \quad \cdots \quad \Pr(Y = n)).$$

Hence, the probability matrix P of two independent random variables has rank 1, is nonnegative, and its entries sum to one.

Suppose that probabilities $\Pr(X = i, Y = j)$ are measurable only for certain pairs (i, j) . A situation in which this might arise in applications is a pair of compounds in a laboratory that only react when in certain states. A complete answer to Problem 1.1 will allow us to reject a hypothesis of independence of the events X and Y , based only on this collection of probabilities.

For each type of partial matrix, we derive an inequality which is satisfied if and only if the partial matrix can be completed to a rank-1 probability matrix. This main result is derived in Theorem 2.14. In Theorem 3.3, we generalize this characterization to diagonal partial tensors which can be completed to rank-1 probability tensors, i.e. rank-1 tensors whose entries are nonnegative and sum to one. Rank-1 probability tensors correspond to joint probabilities of independent random variables X_1, X_2, \dots, X_n . An entry $p_{i_1 i_2, \dots, i_n}$ of a probability tensor expresses the joint probability $\Pr(X_1 = i_1, X_2 = i_2, \dots, X_n = i_n)$.

We design Algorithm 4.2 for checking completability of partial matrices to rank-1 probability matrices. Moreover, we will explain how to construct desired completions. In case there is more than one desired completion, we will show how to use Lagrange multipliers to find a completion that minimizes a distance measure to a fixed probability distribution.

Problem 1.1 is a variation on the well-studied problem of low-rank matrix completion. Király *et al* [KTTU12] introduced algebraic matroid techniques for matrix completion problems. In Section 5, we will study the algebraic matroids that arise in the context of completions of probability matrices.

Problem 1.1 has an immediate generalization to probability matrices of higher rank: We can ask if a partial matrix can be completed to a probability matrix of (nonnegative) rank r . Nonnegative rank r probability matrices express joint probabilities of random variables X and Y that are independent given a hidden random variable Z with r states. However, to study nonnegative rank r probability completions, one has to consider rank- r probability completions first. We will explore an example for matrices of rank 2.

1.1 Outline

In Section 2, we derive, for each type of partial matrix, an inequality which is fulfilled if and only if the partial matrix is completable to a rank-1 probability distribution. Our discussion starts with diagonal probability masks in Section 2.1 and continues with general probability masks in Section 2.2. Theorem 2.2 and Theorem 2.14 characterize when a diagonal, respectively a general, partial matrix is completable.

Partial matrices which can be completed to rank-1 probability matrices form a semialgebraic set, see Proposition 3.1. In Section 3, we will study the semialgebraic description of completable partial matrices and tensors. In Theorem 3.3, we will derive a characterization of diagonal partial tensors which can be completed to rank-1 probability tensors, and in Proposition 3.4, we study the polynomial inequalities defining this semialgebraic set.

In Section 4.1, we use results in Section 2 to define an algorithm which checks completability. In Section 4.2, we present an algorithm to recover the ≤ 2 possible solutions when the solution set is finite. In Section 4.3, we show how to construct a completion if there are infinitely many completions. We will use Lagrange multipliers to construct a rank-1 probability completion which maximizes or minimizes a certain function, e.g. the distance from the uniform distribution or the probability of a particular state.

In Section 5, we examine the algebraic matroids arising from this problem, following the approach of Király *et al* [KTTU12] for low-rank matrix completion.

In Section 6.1, we study generalization of our results to higher rank probability matrices and probability tensors.

Implementations of algorithms can be found on

math.berkeley.edu/~zhrosen/probCompletion.html

The following notation will be used throughout the paper:

Notation	Definition
Δ^n	Standard n -simplex $\{\mathbf{x} \in \mathbb{R}^{n+1} : \sum x_i = 1 \text{ and } x_i \geq 0 \text{ for all } i\}$
Δ_0^n	n -simplex as a corner of the n -cube $\{\mathbf{x} \in \mathbb{R}^n : \sum x_i \leq 1 \text{ and } x_i \geq 0 \text{ for all } i\} = \text{conv}(\{0\} \cup \Delta^{n-1})$
$\Pi_{m \times n}$	Ideal of algebraic relations among entries of a rank-1 matrix with entries summing to 1.
$\mathcal{V}(\Pi_{m \times n})$	Variety of rank-1 matrices with entries summing to 1.
π_S	Probability mask. Projection of $\mathbb{R}^{m \times n}$ onto the coordinates indexed by S . Summarized by 0-1 matrix with 1's for coordinates in the image, and 0's for coordinates in the kernel.
$\pi_S(M)$	Partial matrix. The image of a matrix M under the probability mask π_S . Summarized by matrix with values in the coordinates indexed by S and blanks elsewhere.

The completion problem can be restated as a question about geometry.

Remark 1.2. *The set of rank-1 matrices in $\mathbb{R}^{m \times n}$ with entries summing to 1 is an algebraic variety $\mathcal{V}(\Pi_{m \times n})$. The projection map $\pi_S : \mathcal{V}(\Pi_{m \times n}) \rightarrow \mathbb{R}^{|S|}$ takes a rank-1 matrix with entries summing to 1 and returns some subset of its entries. To move backwards from that subset to the full rank-1 matrix summing to 1, i.e. matrix completion, amounts to computing the fiber of the projection, in particular $\pi_S^{-1}(M) \cap (\mathcal{V}(\Pi_{m \times n}) \cap \Delta^{mn-1})$. The intersection with the simplex restricts the entries to those that qualify for probability matrices.*

2 Completability of Partial Probability Matrix

2.1 Diagonal Partial Matrices

The simplest case to analyze is the probability mask with ones along the diagonal. For a 1×1 matrix, this is trivially completable, indeed completed, if and only if the observed entry is 1. For a 2×2 matrix, there is more to consider.

Example 2.1. *Let M be the partial probability matrix given by:*

$$M = \begin{pmatrix} a & \\ & b \end{pmatrix}.$$

In order for the matrix to be completed, both the rank 1 requirement and the summing to 1 must be addressed. First, for rank 1, the off-diagonal entries are set to x and ab/x , then the quantity $a + ab/x + x + b$ is set equal to 1. The equivalent quadratic equation is $x^2 + (a + b - 1)x + ab = 0$. In order for a real solution for x to exist, the discriminant must be ≥ 0 , i.e.

$$(a + b - 1)^2 - 4ab \geq 0.$$

This inequality, along with the requirement that $a + b \leq 1$ and both $a, b > 0$, is sufficient to guarantee that x gives a completion in Δ^3 , see Figure 1.

For $n > 2$, we take advantage of the factorization of rank-1 matrices as products of vectors to obtain the following more general result:

Theorem 2.2. *Let M be an $n \times n$ partial probability matrix, where $n \geq 2$, with nonnegative observed entries along the diagonal:*

$$M = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix}.$$

Then M is completable if and only if $\sum_{i=1}^n \sqrt{a_i} \leq 1$, or equivalently, $\|(a_1, \dots, a_n)\|_{1/2} \leq 1$. In the special case $\sum_{i=1}^n \sqrt{a_i} = 1$, M has a unique completion.

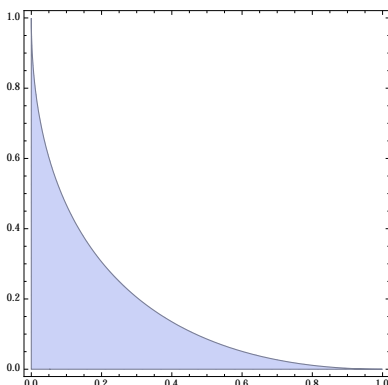


Figure 1: The colored region corresponds to completable probability masks of 2×2 matrices with diagonal entries a, b observed.

Proof. Recall that a rank-1 $n \times n$ probability matrix can be factored as $\mathbf{u}^T \mathbf{v}$ for $\mathbf{u} \in \mathbb{R}^n, \mathbf{v} \in \mathbb{R}^n$ where the sum of the entries in each vector is 1. For this problem, consider all possible values of \mathbf{u} in Δ^{n-1} , but do not restrict the values of \mathbf{v} to the simplex. Instead, let \mathbf{v} be formulated in terms of \mathbf{u} and the entries of the matrix. Explicitly, we have

$$\begin{pmatrix} u_1 \\ \vdots \\ u_n \end{pmatrix} \begin{pmatrix} a_1/u_1 & \cdots & a_n/u_n \end{pmatrix} = \begin{pmatrix} a_1 & & \\ & \ddots & \\ & & a_n \end{pmatrix}.$$

A probability completion will arise when $\sum v_i = 1$. Here we assume $a_i > 0$ and thus $u_i > 0$ for all i . We will consider the case when $a_i = 0$ for some i separately.

Let $f(\mathbf{u}) = \sum_{i=1}^n v_i = \sum_{i=1}^n a_i/u_i$ denote this quantity and compute the minimum of f on the simplex. For this computation, consider u_1, \dots, u_{n-1} as independent variables and $u_n = 1 - \sum u_i$.

$$\begin{aligned} f &= \left(\sum_{i=1}^{n-1} \frac{a_i}{u_i} \right) + \frac{a_n}{1 - \sum_{i=1}^{n-1} u_i} \\ \Rightarrow \frac{\partial f}{\partial u_i} &= -\frac{a_i}{u_i^2} + \frac{a_n}{(1 - \sum_{i=1}^{n-1} u_i)^2} \end{aligned}$$

Setting $\partial f / \partial u_i = 0$ for all i implies $a_i/u_i^2 = k$ is constant for all i . Since \mathbf{u} is in the simplex, k is equal to $(\sum \sqrt{a_i})^2$. The value of f (i.e. the sum of v_i) at this point is $(\sum \sqrt{a_i})^2$. If this is ≤ 1 , continuity of f implies that a completion exists somewhere between our minimum and the boundary, because within an ϵ of the boundary of Δ^{n-1} , we have

$$\sum v_i = \frac{a_1}{u_1} + \cdots + \frac{a_n}{u_n} \gg 1.$$

If this minimum value is > 1 , the function will not achieve 1 anywhere in the simplex, so no completion is possible.

Now assume $a_i = 0$ for $i \in I$. If $|I| \geq n - 1$, then the statement of the theorem is clearly satisfied. Assume $|I| \leq n - 2$. If $\sum_{i \in [n] \setminus I} \sqrt{a_i} \leq 1$, then the probability mask with the rows and columns in I removed has a rank 1 probability completion, and it can be extended to a rank 1 completion of the original matrix by replacing the entries in the removed rows and columns with zeros. On the other hand, a completion of the original matrix gives a completion of the reduced matrix with sum of entries ≤ 1 . By continuity, the reduced matrix also has a completion with sum of entries equal to 1. Hence $\sum_{i \in [n] \setminus I} \sqrt{a_i} \leq 1$.

Finally, when equality is attained, the solution must be unique, since $(\sum \sqrt{a_i})^{-1}(\sqrt{a_1}, \dots, \sqrt{a_n})$ is unique as a minimum in the simplex. \square

Corollary 2.3. *Let $\sum_{i=1}^n \sqrt{a_i} < 1$. For $n = 2$, the probability mask M has exactly two completions. If $n > 2$, then the set of completions of M is $(n - 2)$ -dimensional.*

Proof. If M contains no zeros, then every path from $(\sum \sqrt{a_i})^{-1}(\sqrt{a_1}, \dots, \sqrt{a_n})$ to the boundary of the simplex will contain exactly one completion. For $\mathbf{u} \in \Delta^1$, there are only two distinct paths to the boundary. For higher-dimensional simplices, this will induce a codimension 1 set inside the $(n - 1)$ -dimensional simplex, which gives an $(n - 2)$ -dimensional set.

If M has zeros, let $I = \{i : a_i = 0\}$ denote the indices of zero observed entries. The set of completions of M is

$$\bigcup_{R, C \subseteq I: R \cup C = I} \{\text{completions of } M \text{ with rows in } R \text{ and columns in } C \text{ zero}\}.$$

By the previous discussion, the dimension of the semialgebraic set of completions with the sum of entries ≤ 1 of the $(n - |I|) \times (n - |I|)$ submatrix of M corresponding to nonzero observed entries is $(n - |I| - 1)$. For each such completion, we can freely fix all but one of the row sums u_i with $i \in I \setminus R$ and columns sums v_j with $j \in I \setminus C$: Without loss of generality assume $|C| \leq |R|$. Row sums u_i with $i \in I \setminus R$ determine all row sums. The nonzero observed entries together with row sums determine column sums in columns $[n] \setminus I$. All but one of the rest of the columns sums can be chosen freely such that they sum to one. Hence the dimension of the set of completions corresponding to R and C is $(n - |I| - 1) + (|I| - |R| + |I| - |C| - 1) \leq (n - |I| - 1) + (|I| - 1) = n - 2$. The equality is obtained for any R and C with $|R| + |C| = |I|$. \square

Remark 2.4. *The analysis of Theorem 2.2 works to derive the constraint for the 2×2 diagonal probability mask as in Example 2.1:*

$$\begin{aligned} \sqrt{a} + \sqrt{b} \leq 1 &\Leftrightarrow a + b + 2\sqrt{ab} \leq 1 \Leftrightarrow 2\sqrt{ab} \leq 1 - a - b \\ &\Leftrightarrow 4ab \leq (1 - a - b)^2 \text{ and } 0 \leq 1 - a - b. \end{aligned}$$

Example 2.5. *The matrix $\text{diag}(1/4, 1/4, 1/4)$ does not have a completion to a rank-1 probability matrix, since $3\sqrt{1/4} = 3/2$. The set of 3×3 diagonal partial matrices that are completable to rank-1 probability matrices is shown in Figure 2.*

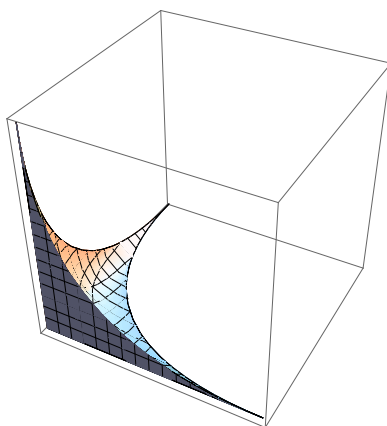


Figure 2: The colored region corresponds to completable probability masks of 3×3 matrices with diagonal entries observed.

2.2 Block Diagonal Matrices

In this section, we will use our result about to diagonal partial matrices to prove completability conditions for different types of partial matrices with increasing generality:

$$\text{Diagonal} \implies \text{Block Diagonal} \implies \text{Acyclic} \implies \text{Feasible.}$$

To discuss non-diagonal masks, we introduce graph notation used in various places in the matrix completion literature.

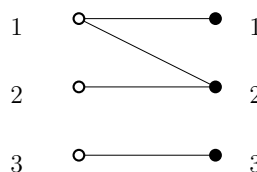
Definition 2.6. Let $M = \mathbf{u}^T \mathbf{v}$ be an $m \times n$ matrix of probabilities whose entries sum to 1, and let \mathbf{u} and \mathbf{v} be vectors in Δ^{m-1} and Δ^{n-1} respectively. A bipartite graph can be associated to M in the following way:

Graph	Matrix
White vertex r_i	i -th row
Black vertex c_j	j -th column
Edge $(r_i c_j)$	(i, j) -th entry
Weight $\omega(r_i)$	Sum of i -th row, or u_i
Weight $\omega(c_j)$	Sum of j -th column, or v_j
Edge weight $\omega(r_i c_j)$	Value of entry m_{ij}

The bipartite graph associated to a probability mask is the graph obtained by deleting the edges corresponding to unobserved entries, and omitting vertex weights.

Example 2.7. On the left is a partial probability matrix with entries, and at right is the corresponding bipartite graph.

$$\begin{pmatrix} x_{11} & x_{12} & \\ x_{21} & & \\ & & x_{33} \end{pmatrix}$$



Note that in this formulation, the question of completability is equivalent to the existence of a vertex labeling so that the black vertex weights and white vertex weights each sum to 1, and the edge weights satisfy $\omega(r_i c_j) = \omega(r_i)\omega(c_j)$. The diagonal case describes those masks whose graphs are the union of disjoint edges. We now consider more general bipartite graphs.

Lemma 2.8. *Let G be a bipartite graph with a connected component $K_{p,q}$, with edge weights $a_{1,1}, \dots, a_{p,q}$, so that the corresponding submatrix is rank 1. Let H be the graph $(G \setminus K_{p,q}) \cup K_{1,1}$, with the edge weight on $K_{1,1}$ given by $\omega(a_{1,1}) = \sum_{i=1}^p \sum_{j=1}^q a_{i,j}$. Then completability of G is equivalent to completability of H .*

Proof. (\Rightarrow) Begin with a vertex weighting on G . Replace the white vertices weighted u_1, \dots, u_p with a single white vertex weighted $\sum_{i=1}^p u_i$ and the black vertices weighted v_1, \dots, v_q by a single black vertex labeled $\sum_{i=1}^q v_i$. Since $K_{p,q}$ was disconnected from the other vertices, no other observed entries will be changed by this replacement.

(\Leftarrow) Begin with a vertex weighting on H . The fact that $a_{1,1}, \dots, a_{p,q}$ are a rank-1 $p \times q$ matrix implies that there is a rank-1 factorization $\mathbf{u}'^T \mathbf{v}'$. Scale the vector \mathbf{u}' with the constant $u_1 / \sum u'_i$; and scale \mathbf{v}' by the inverse. The resulting submatrix is the same but now $\sum u'_i = u_1$ and $\sum v'_i = v_1$. The new vertex weights give a completion of G . \square

This lemma establishes that Theorem 2.2 applies to block diagonal matrices as well. To extend to general acyclic matrices, we need to make a definition to account for exceptional cases involving zeros:

Definition 2.9. *Let M be a partial matrix, with corresponding graph G . We say that M is prunable if there is a set of vertices $W \subset V(G)$ such that every edge labeled zero is adjacent to some $w \in W$, but no edge with nonzero label is adjacent to any $w \in W$.*

Pruning refers to removing W and all incident edges from the graph, or equivalently, considering only the induced subgraph on $V \setminus W$. A careful pruning takes W so that the remaining graph has the largest possible number of components.

Proposition 2.10. *A matrix is completable to a rank-1 probability matrix only if it is prunable.*

Proof. Let W be the set of vertices whose incident edges are all labeled zero. Suppose some $e \in E(G)$ labeled zero is not adjacent to any $w \in W$. Then, if $e = (r_i, c_j)$, both r_i and c_j are connected to some other vertices c_k and r_l with nonzero edges. The 2×2 minor of M defined by rows i, l and columns j, k then has two nonzero entries along one diagonal and zero in the other. Therefore, M cannot be rank one. \square

The next lemma allows us to confine our conversation to matrices with nonzero entries:

Lemma 2.11. *Let G be a bipartite graph with a vertex v such that all edges incident to v have weight 0. In particular, the vertex v might be an isolated vertex. Completability of G is equivalent to completability of $G \setminus v$, except when $G \setminus v$ is connected.*

Proof. (\Rightarrow) Assume that the probability mask M corresponding to G is completable. Remove the row or column corresponding to v . Contract rows and columns corresponding to each completable block to one row and column respectively. Add the corresponding entries in a completion of M . After permuting rows and columns, we get a diagonal probability mask that has a nonnegative completion with the sum of entries ≤ 1 (by modifying the completion of M in the same way as we have modified M). By the proof of Theorem 2.2, the diagonal probability mask is completable as long as $n > 1$. By Lemma 2.8, also the probability mask corresponding to $G \setminus v$ is completable. In the case where $G \setminus v$ is connected, this reasoning does not hold, since the sum must be $= 1$.

(\Leftarrow) Begin with a probability matrix M which is a completion of $G \setminus v$. Simply add in a row or column of zeros corresponding to v to obtain the desired completion of G . \square

Example 2.12. *To illustrate why the exception in Lemma 2.11 is necessary, consider the following partial matrix:*

$$\begin{pmatrix} 0.15 & & \\ 0.05 & 0.1 & \\ & 0 & 0 \end{pmatrix}$$

If we prune off the bottom row, the remaining matrix is completable; however, if we prune off the last column, the matrix has only one rank-1 completion and its entries do not sum to 1.

We say that a block of a matrix is 1-closable (as in [KTTU12]), if the corresponding graph is a spanning tree. If all labels in a spanning tree are nonzero, then the rest of the entries in the block can be completed using fundamental cycles (see Section 4.1 for details). So, Lemma 2.11 will allow us to prune a partial matrix to the form where it can be completed to a block diagonal partial matrix using cycles. The result from the discussion so far is that for matrices that can be pruned to acyclic graphs with nonzero edge labels and $n > 1$ component, the question is reduced to the problem solved in Section 2.1.

In order to allow cycles in the graph, we recall that a Universal Gröbner basis for the ideal of relations among entries in a rank-1 matrix is indexed by the set of cycles of the bipartite graph; see Sections 4.1 and 5 for more detail. If the cycles in the graph satisfy these relations, they admit a rank-1 completion. We summarize all of our conditions in the following definition:

Definition 2.13. *A matrix M is said to be feasible, if M is prunable, and cycles in the graph of M satisfy the binomial relations from the Universal Gröbner basis.*

Theorem 2.14. *Let M be a feasible partial probability matrix such that after careful pruning, its graph G has s connected components. Let b_i be the sum of the weights in the i -th connected component of G **after rank-1 closure**. If $s = 1$, then M is completable if and only if $b_1 = 1$. For $s > 1$, M is completable to a probability matrix if and only if:*

$$\sum_{i=1}^s \sqrt{b_i} \leq 1.$$

Proof. By Lemma 2.11, if the pruned matrix has more than one component, completability of the pruned matrix is equivalent to completability of the original. In the exceptional case where every choice of vertices leaves one component, completability of the pruned matrix is also equivalent to the original. This is because every row and column removed was forced to be all zeros so cannot contribute to the total probability.

On the remaining nonzero matrix, blocks are completed using cycles, and then contracted from a block diagonal matrix to a diagonal matrix; this does not change completability by Lemma 2.8. Finally, we use Theorem 2.2. \square

Example 2.15. *The probability mask*

$$\begin{pmatrix} x_{11} & x_{12} & & \\ x_{21} & & & \\ & & & x_{33} \end{pmatrix} \quad (2.1)$$

with all observed entries nonnegative has a completion if and only if

$$\sqrt{x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11}} + \sqrt{x_{33}} \leq 1.$$

This is equivalent to the conditions

$$\begin{aligned} (x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11} + x_{33} - 1)^2 - 4(x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11})x_{33} &\geq 0, \\ x_{11} + x_{12} + x_{21} + x_{12}x_{21}/x_{11} + x_{33} &\leq 1. \end{aligned}$$

By clearing the denominators, we get polynomial inequalities in the observed entries whose solutions are all completable probability masks of form (2.1).

3 Semialgebraic Description

3.1 Examples

For an introduction to real algebraic geometry, see [BPR06].

Proposition 3.1. *Partial matrices of fixed type which can be completed to rank-1 probability matrices form a semialgebraic set.*

Proof. The independence model is a semialgebraic set defined by 2×2 -minors, nonnegativity constraints and entries summing to one. The statement of the proposition follows by Tarski-Seidenberg theorem. \square

The goal of this section is to find a semialgebraic description of this semialgebraic set. The difference from characterizations in Theorems 2.2 and 2.14 is that we aim to derive a description without square roots. For 2×2 partial matrices with diagonal entries, a semialgebraic description is given in Example 2.1 and its derivation from the inequality containing square roots is explained in Remark 2.4.

Example 3.2. Let us consider 3×3 probability masks with diagonal entries a, b, c . Denote the elementary symmetric polynomials by $S_1 = a + b + c$, $S_2 = ab + bc + ca$ and $S_3 = abc$. By consecutive squaring and reordering terms, we get:

$$\sqrt{a} + \sqrt{b} + \sqrt{c} = 1 \quad (3.1)$$

$$\Rightarrow 2(\sqrt{ab} + \sqrt{bc} + \sqrt{ca}) = 1 - S_1 \quad (3.2)$$

$$\Rightarrow 8(a\sqrt{bc} + b\sqrt{ac} + c\sqrt{ab}) = (1 - S_1)^2 - 4S_2 \quad (3.3)$$

$$\Rightarrow 128S_3(\sqrt{ab} + \sqrt{bc} + \sqrt{ca}) = ((1 - S_1)^2 - 4S_2)^2 - 64S_1S_3 \quad (3.4)$$

Substituting Equation (3.2) into Equation (3.4) gives:

$$64S_3(1 - S_1) = ((1 - S_1)^2 - 4S_2)^2 - 64S_1S_3 \quad (3.5)$$

$$\Leftrightarrow ((1 - S_1)^2 - 4S_2)^2 - 64S_3 = 0 \quad (3.6)$$

The degree four polynomial equation (3.6) with 35 terms in the region

$$a, b, c \geq 0, \quad (3.7)$$

$$1 - S_1 \geq 0, \quad (3.8)$$

$$(1 - S_1)^2 - 4S_2 \geq 0 \quad (3.9)$$

gives the equation (3.1). The left hand sides of inequalities (3.8) and (3.9) are given by right hand sides of equations (3.2) and (3.3). The region defined by

$$\sqrt{a} + \sqrt{b} + \sqrt{c} \leq 1$$

is the same as

$$((1 - S_1)^2 - 4S_2)^2 - 64S_3 \geq 0 \quad (3.10)$$

together with inequalities (3.7), (3.8) and (3.9). This is a semialgebraic description of 3×3 diagonal partial matrices which can be completed to rank-1 probability matrices.

Before we continue with studying semialgebraic descriptions of general partial matrices, we make a detour and characterize diagonal partial tensors that can be completed to rank-1 probability tensors. This is also a semialgebraic set and its semialgebraic description can be studied in a similar way to the partial matrix case.

3.2 Tensors

The reasoning for diagonal matrices translates nicely into higher-dimensional tensors as follows:

Theorem 3.3. Suppose we are given a partial probability tensor $T \in \Delta^{n^d-1} \subset (\mathbb{R}^n)^{\otimes d}$ with nonnegative observed entries a_i along the diagonal, i. e. we have $t_{i\dots i} = a_i$ for $1 \leq i \leq n$, and all other entries unobserved. Then T is completable if and only if

$$\sum_{i=1}^n a_i^{1/d} \leq 1.$$

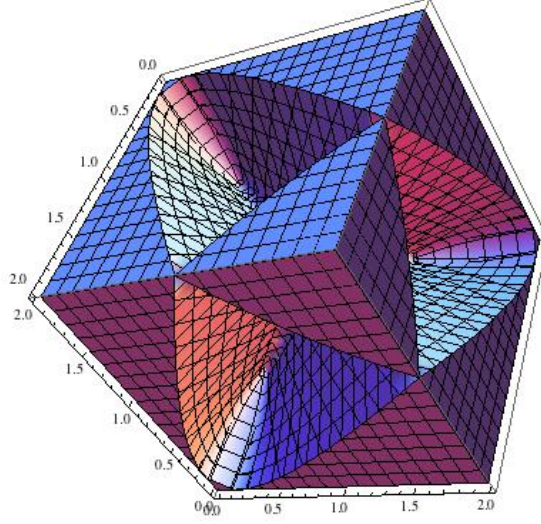


Figure 3: Region defined by the polynomial inequality (3.10) inside the cube $[0, 2]^3$.

Proof. The proof is analogous to the proof of Theorem 2.2, with a few adjustments to deal with the multiple parametrizing vectors. The tensor T can be factored as $\mathbf{u}^1 \otimes \dots \otimes \mathbf{u}^d$, where each $\mathbf{u}^i \in \Delta^{n-1} \subset \mathbb{R}^n$. The relations $a_i = u_i^1 \dots u_i^d$ imply that the coordinates of \mathbf{u}^d can be expressed as functions on the product of simplices $(\Delta^{n-1})^{d-1}$. Define $f : (\Delta^{n-1})^{d-1} \rightarrow \mathbb{R}$ by:

$$f(\mathbf{u}^1, \dots, \mathbf{u}^{d-1}) = \sum_{i=1}^n u_i^d = \sum_{i=1}^n \frac{a_i}{u_i^1 \dots u_i^{d-1}}.$$

Since every variable appears in the denominator of some term, f approaches infinity at the boundary of the product of simplices. A candidate vector \mathbf{u}^d will be available if and only if the minimum value of f on the product of simplices is below one. To find the minimum, compute partial derivatives; as in the proof of Theorem 2.2, we let $u_n^j = 1 - \sum_{k=1}^{n-1} u_k^j$ for each $j = 1, \dots, d-1$.

$$\frac{\partial f}{\partial u_i^j} = -\frac{a_i}{u_i^j \prod_{k=1}^{d-1} u_i^k} + \frac{a_n}{u_n^j \prod_{k=1}^{d-1} u_n^k}.$$

Setting the partial derivatives to zero gives us the following:

$$\frac{a_i}{u_i^j \prod_{k=1}^{d-1} u_i^k} = \frac{a_d}{u_n^j \prod_{k=1}^{d-1} u_n^k} \quad \forall i \in [n-1], \forall j \in [d-1].$$

Let c_j be the value on both sides of this equation. Picking two values of j , w.l.o.g., 1 and 2, this designation means:

$$\frac{a_i}{u_i^1 \prod_{k=1}^{d-1} u_i^k} / \frac{a_i}{u_i^2 \prod_{k=1}^{d-1} u_i^k} = \frac{u_i^2}{u_i^1} = \frac{c_1}{c_2}.$$

Applying to all indices, we have $u_i^j = \frac{c_1}{c_j} u_i^1$ for all i, j . Since $\sum_i u_i^j = 1 = \frac{c_1}{c_j} \sum_i u_i^1 = c_1/c_j$ for all j , implying that every $c_j = c_1$, and

$$\begin{aligned} \frac{a_i}{(u_i^j)^d} &= c_1 \quad \forall i \in [n], \forall j \in [d-1] \\ \Rightarrow u_i^j &= (a_i/c_1)^{1/d} = \kappa(a_i)^{1/d}. \end{aligned}$$

for some constant κ . Since the sum $\sum_i u_i^j = 1$, the value of $\kappa = (\sum_{i=1}^n a_i^{1/d})^{-1}$. Plugging in these values of u_i^j , we obtain

$$f = \sum_{i=1}^n \frac{a_i}{(\kappa(a_i)^{1/d})^{d-1}} = \sum_{i=1}^n \frac{a_i^{1/d}}{\kappa^{d-1}} = \left(\sum_{i=1}^n a_i^{1/d} \right)^d.$$

Since f is at its minimum here, the value must be less than one for a solution to exist, proving the theorem. \square

3.3 Semialgebraic Description

We will characterize the semialgebraic set of diagonal partial tensors which can be completed to rank-1 probability tensors. This is the positive part of the unit ball in the $L^{\frac{1}{d}}$ space.

Proposition 3.4. *There exists a unique irreducible polynomial f of degree d^{n-1} with constant term 1 that vanishes on the boundary of diagonal partial tensors which can be completed to rank-1 probability tensors. The semialgebraic description takes the form $f \geq 0$, coordinates ≥ 0 plus additional inequalities that separate our set from other chambers in the region defined by $f \geq 0$.*

The proof of Proposition 3.4 was suggested to us by Bernd Sturmfels. For analogous proof idea, see [NPS08, Lemma 2.1].

Proof. Denote the diagonal entries of the partial tensor by x_1, \dots, x_n . We will show that the defining polynomial of the $\frac{1}{d}$ -unit ball can be written as

$$p_{d,n} = \prod_{\substack{y_i \text{ s.t. } y_i^d = x_i^{1/d} \\ \text{for each } i}} \left((1 - y_1 - \dots - y_{n-1})^d - x_n \right). \quad (3.11)$$

We want to eliminate y_1, \dots, y_n from the ideal

$$I = \left\langle y_1^d - x_1, \dots, y_n^d - x_n, \sum_{i=1}^n y_i - 1 \right\rangle \subset \mathbb{Q}[x_1, \dots, x_n, y_1, \dots, y_n].$$

First replace y_n by $1 - y_1 - \dots - y_{n-1}$ in the equation $y_n^d - x_n$. We consider the field of rational functions $K = \mathbb{Q}(x_1, \dots, x_n)$. Solving the first $n-1$ equations $y_i^d - x_i$ is

equivalent to adjoining the d -th roots of x_i for $i \in \{1, \dots, n-1\}$ to the base field. This gives a Galois extension L of degree d^{n-1} over K . The Galois group of the extension L/K is $(\mathbb{Z}/d\mathbb{Z})^{n-1}$. The product (3.11) is the orbit of $(1 - y_1 - \dots - y_{n-1})^d - x_n$ under the action of the Galois group, and thus lies in the base field K . Every factor in the product (3.11) is integral over $\mathbb{Q}[x_1, \dots, x_n]$, hence the product (3.11) is a degree d^{n-1} polynomial in x_1, \dots, x_n . No subproduct is left invariant under the Galois group, so (3.11) is irreducible. \square

In Example 3.2, the polynomial f is the left hand side of Equation (3.6). Inequalities (3.7), (3.8) and (3.9) separate the set $\sqrt{a} + \sqrt{b} + \sqrt{c} \leq 1$ from other chambers of $f \geq 0$.

The semialgebraic description for an arbitrary probability mask can be constructed in five steps using the semialgebraic description for diagonal masks:

1. Take all elements of the universal Gröbner basis of $\Pi_{m \times n}$ (see Section 4.1) that contain only observed entries. With these equations we check that observed entries do not contradict the rank-1 condition.
2. For each $R \subseteq [m], C \subseteq [n]$ consider the semialgebraic set

$$\{x_{ij} \text{ is observed} : x_{ij} = 0 \text{ if } i \in R \text{ or } j \in C \text{ and } x_{ij} > 0 \text{ otherwise}\}. \quad (3.12)$$

For example, the partial matrix

$$\begin{pmatrix} 1 & \\ 0 & 1 \end{pmatrix}.$$

does not belong to the semialgebraic set (3.12) for any R and C

3. For fixed R and C , express all completable entries as rational functions in observed entries, see Equation (4.2). This gives a block diagonal matrix.
4. Construct the diagonal mask corresponding to the block diagonal mask, where each diagonal entry is equal to the sum of entries in the corresponding block. Clear denominators. Intersect the semialgebraic set for this diagonal mask with the semialgebraic set in Step 2.
5. Take the union of semialgebraic sets in Step 5 for all R and C and intersect with the variety in Step 1.

4 Completion Algorithms

4.1 Algorithm for Checking Completability

When the bipartite graph corresponding to a probability mask is connected and acyclic with all nonzero entries, there is a unique completion. The missing entries can be computed by the following relations among entries in a cycle:

$$\prod_{(i,j) \in E_1} x_{ij} - \prod_{(i,j) \in E_2} x_{ij} = 0, \quad (4.1)$$

where (E_1, E_2) is a partition of the edges in a cycle so that no two edges in either E_i are adjacent. In particular, since the graph G is connected, if there is a missing edge (r_k, c_l) there is a path from r_k to c_l in G ; ordering those edges in the path order from r_k to c_l , let S_1 be every other edge in the list starting with r_k , and let S_2 be the complement in the path. Then:

$$\omega(r_k c_l) = \prod_{(ij) \in S_1} x_{ij} / \prod_{(ij) \in S_2} x_{ij}. \quad (4.2)$$

This formula is implemented in [KTTU12, Algorithm 8].

Example 4.1. *Given the base indicated at left, we can uniquely complete as at right using polynomials (4.1):*

$$\begin{pmatrix} a & b & \\ c & & d \\ & e & \end{pmatrix} \rightarrow \begin{pmatrix} a & b & \frac{ad}{c} \\ c & \frac{bc}{a} & d \\ \frac{ae}{b} & e & \frac{ade}{bc} \end{pmatrix}$$

The entries are assumed to be nonzero; if an observed entry is zero, we omit an appropriate row and/or column and complete the rest. If a probability matrix only has one connected component, then completability amounts to checking that the unique rank-1 completion has entries summing to 1.

Combining Theorem 2.14 with the algorithm just described, we may now present an algorithm for checking completability of an arbitrary partial probability matrix.

Algorithm 4.2 (Completability of Arbitrary Partial Matrices). *Let M be a partial matrix with nonnegative entries whose sum is ≤ 1 . The following algorithm answers "Is M completable?":*

1. Translate M into the corresponding bipartite graph G including edge weights.
2. If all edge weights are nonzero, proceed to Step 3; otherwise:
 - (a) If G is not prunable, return "NO."
 - (b) If G is prunable, execute a careful pruning, as described in Definition 2.9.
3. In the remaining graph, suppose there are n connected components, C_1, \dots, C_n . If $n = 1$:
 - (a) Check that Equation (4.1) holds for every cycle in the graph. If a cycle fails, return "NO."
 - (b) Uniquely complete to $K_{m,n}$. If the edge weights after completion add up to 1, return "YES." Else, return "NO."

4. If $n > 1$:

- (a) Check that Equation (4.1) holds for every cycle in the graph. If a cycle fails, return "NO."
- (b) Add in edges to make each component a complete bipartite graph, with edge weights from Equation (4.2). Let $S = \sum_{i=1}^n \sqrt{b_i}$ where b_i is the sum of the entries in C_i after the last step. If $S > 1$, return "NO." Else, return "YES."

4.2 Algorithm for Completing Partial Matrices with Two Components

In the last section, an algorithm was presented that determines completability for an arbitrary mask. In the special case where Step 3 returns $n = 2$ connected components, there is a finite set of completions, with cardinality between 0 and 2 (see Section 5 for an explanation). The following algorithm returns this set of completions:

Algorithm 4.3. Begin with a graph G , and carry out Algorithm 4.2 until Step 3. Let H be the graph returned from Step 3 with two complete bipartite components C_1 and C_2 .

1. Choose any edge connecting C_1 and C_2 ; set it equal to X . Fill in the remaining entries (in terms of the known entries and X) using Equation (4.2).
2. Solve the quadratic equation $\sum p_{ij} = 1$ for X , where p_{ij} 's are the set of entries obtained in Step 1. Substitute the two values for X into the completed matrix from Step 3 to obtain two (usually distinct) completions.
3. Reintroduce any vertices removed in Step 2 of Algorithm 4.2 as rows/columns with all zero entries.

Note that since the entries involving X all have the same sign, and their sum is fixed, the solutions for X (if real) must be both positive or both negative.

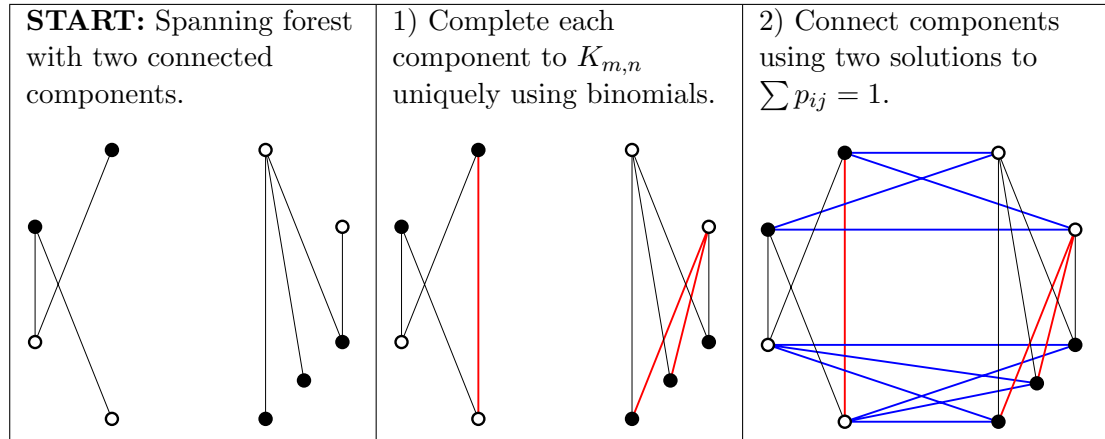


Table 2: Algorithm for completing probability matrix projections.

Example 4.4. Consider the projection of a 3×3 matrix into Δ_0^4 indicated by the partial matrix below:

$$\begin{pmatrix} .06 & .09 & \\ .08 & & \\ & & .15 \end{pmatrix}$$

Step 1: Add in the missing edges so that both connected components are complete bipartite.

$$\begin{pmatrix} .06 & .08 & \\ .09 & .12 & \\ & & .15 \end{pmatrix}$$

Step 2: Add in X to connect the components, and fill in the remaining edges.

$$\begin{pmatrix} .06 & .08 & X \\ .09 & .12 & 1.5X \\ .009/X & .012/X & .15 \end{pmatrix}$$

Step 3: Set $\sum p_{ij} = 1$ and solve for X :

$$\begin{aligned} (.06 + .08 + .09 + .12 + .15) + X + 1.5X + .009/X + .012/X &= 1 \\ \Rightarrow .5 + 2.5X + .021/X &= 1 \Rightarrow 2.5X^2 - .5X + .021 = 0 \end{aligned}$$

The two solutions for X yield the following two completions:

$$\begin{pmatrix} .06 & .08 & .06 \\ .09 & .12 & .09 \\ .15 & .2 & .15 \end{pmatrix} \qquad \begin{pmatrix} .06 & .08 & .14 \\ .09 & .12 & .21 \\ 9/140 & 3/35 & .15 \end{pmatrix}$$

If a set of entries is indeed a projection of a probability matrix, this algorithm will recover it. Though the generic fiber has two points, there will be a unique completion if and only if the discriminant of the quadratic polynomial in X is zero.

Example 4.5. Applying the algorithm to a random point $x \in \Delta_0^{m+n-2}$ does not necessarily produce a matrix in the probability simplex Δ^{mn-1} . Indeed, even in the smallest cases, this is seen to be false; the matrix below left is obviously in Δ_0^2 , but its fiber in $\mathcal{V}(\Pi_{m \times n})$ consists of the matrix at right and its complex conjugate:

$$\begin{pmatrix} \frac{1}{3} & \\ & \frac{1}{3} \end{pmatrix} \rightarrow \begin{pmatrix} \frac{1}{3} & \frac{1}{6} + \frac{i}{2\sqrt{3}} \\ \frac{1}{6} - \frac{i}{2\sqrt{3}} & \frac{1}{3} \end{pmatrix}$$

4.3 Completing Partial Matrices with More Than Two Components

When the graph has $n > 2$ components, and the quantity $\sum_{i=1}^n \sqrt{b_i} < 1$, there is an $(n - 2)$ -dimensional set of completions. For example, consider a diagonal partial 3×3 matrix with each observed entry equal to the same constant c . In Figure 4, each curve represents values of \mathbf{u} that parametrize a completion of the partial matrix with c on the diagonal, for various values of c . Here \mathbf{u} is projected onto the first two coordinates.

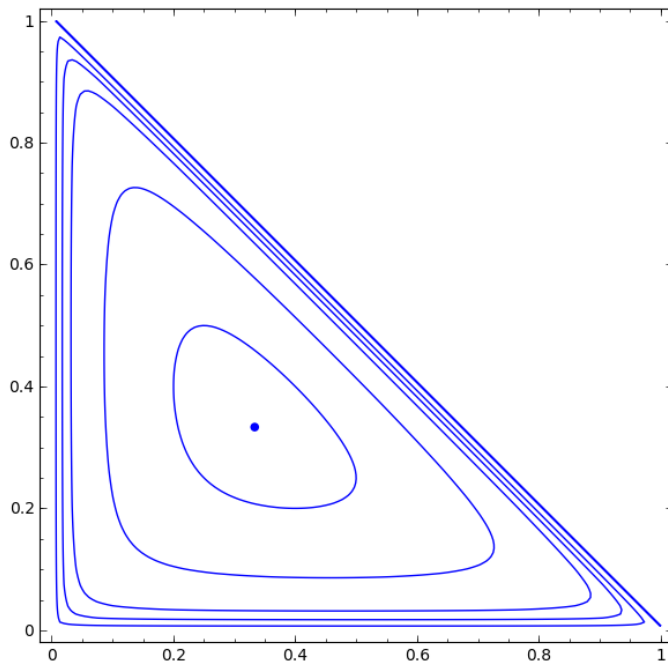


Figure 4: Solution curves for $c = 1/9, 1/10, 1/16, 1/36, 1/64$, and $1/150$, in the projection of the simplex onto the first two coordinates.

For practical applications, some completions are more useful than others. We may want to minimize a distance measure d from a fixed probability distribution. We will explain how to use Lagrange multipliers to solve this optimization problem if d is the Pearson χ^2 distance from the uniform distribution.

Let $S = \sum_{i=1}^n \sqrt{a_i}$. Let us parametrize a vector $\mathbf{u} \in \Delta^{n-1}$ by

$$\mathbf{u}(\mathbf{t}) = \left(\frac{\sqrt{a_1}}{S} + t_1, \frac{\sqrt{a_2}}{S} + t_2, \dots, \frac{\sqrt{a_{n-1}}}{S} + t_{n-1}, \frac{\sqrt{a_n}}{S} + t_n \right),$$

where $t_n = -t_1 - t_2 - \dots - t_{n-1}$. Then

$$f(\mathbf{u}(\mathbf{t})) = \sum_{i=1}^n v_i(\mathbf{t}) = \sum_{i=1}^n \frac{a_i}{u_i(\mathbf{t})} = \sum_{i=1}^n \frac{a_i}{\frac{\sqrt{a_i}}{S} + t_i} = \sum_{i=1}^n \frac{a_i S}{\sqrt{a_i} + t_i S}.$$

Proposition 4.6. *The semialgebraic set of completions of a diagonal probability mask $\text{diag}(a_1, a_2, \dots, a_n)$ is given by $f(\mathbf{u}(\mathbf{t})) = 1$ and $\mathbf{u}(\mathbf{t}) \geq 0$ (after clearing denominators).*

By the method of Lagrange multipliers, an element $(t_1, t_2, \dots, t_{n-1})$ in this semialgebraic set is a critical point for a distance function d if and only if the gradient of d is a constant multiple of the vector of partial derivatives $\frac{\partial f}{\partial t_i}$. To compute all the critical points of the function d on the variety given by $f(\mathbf{t}) = 1$, we need to solve the system of rational equations given by $f(\mathbf{t}) = 1$ and all the 2×2 minors of the matrix

$$L = \begin{pmatrix} \frac{\partial f}{\partial t_1} & \frac{\partial f}{\partial t_2} & \cdots & \frac{\partial f}{\partial t_{n-1}} \\ \frac{\partial d}{\partial t_1} & \frac{\partial d}{\partial t_2} & \cdots & \frac{\partial d}{\partial t_{n-1}} \end{pmatrix}.$$

Finally we need to check for all real solutions which satisfy $\mathbf{u}(\mathbf{t}) \geq 0$ which one minimizes the distance d .

Example 4.7. *Let us return to the matrix $A = \text{diag}(1/4, 1/25, 1/36)$, and find a completion that minimizes the Pearson χ^2 distance from the uniform distribution:*

$$d = \frac{1}{n^2} \sum_{i,j} \left(p_{ij} - \frac{1}{n^2} \right)^2 = \frac{1}{n^2} \sum_{i,j} \left(u_i v_j - \frac{1}{n^2} \right)^2.$$

We use the Pearson χ^2 distance, because it is widely used to test whether an observed distribution differs from a theoretical distribution. However, our method is not restricted to the Pearson χ^2 distance and can be used for any distance measure.

We construct the Lagrange matrix

$$L = \begin{pmatrix} \frac{\partial f}{\partial t_1} & \frac{\partial f}{\partial t_2} \\ \frac{\partial d}{\partial t_1} & \frac{\partial d}{\partial t_2} \end{pmatrix}$$

and find the critical points of d on the variety $f = 1$ by solving the system of rational equations $\{f = 1, \det(L) = 0\}$. We use `maple` to construct L and to solve the system of equations. This system has 18 solutions, out of which ten are real and four are feasible, i.e. they satisfy $\mathbf{u} \geq 0$. The minimum is achieved at

$$M = \begin{pmatrix} 0.250 & 0.049 & 0.215 \\ 0.204 & 0.040 & 0.176 \\ 0.032 & 0.006 & 0.028 \end{pmatrix} \text{ and } M^T.$$

The Pearson χ^2 distance from the uniform distribution is 0.683.

Any path from the local minimum to the boundary of the simplex will strike at least one solution. If any completion is acceptable, we can designate a simple path and find its points of intersection with the semialgebraic set of completions.

Proposition 4.8. *Let $A = \text{diag}(a_1, \dots, a_n)$, such that $n > 2$ and $S = \sum \sqrt{a_i} < 1$. Then, a completion of the matrix is given by:*

$$\mathbf{u} = \left(\frac{\sqrt{a_1}}{S} + t, \frac{\sqrt{a_2}}{S} - t, \frac{\sqrt{a_3}}{S}, \dots, \frac{\sqrt{a_n}}{S} \right),$$

where t is one of the solutions to the following quadratic equation:

$$\begin{aligned} \left(\frac{\sqrt{a_1} + \sqrt{a_2}}{S}\right) t^2 + \left(a_2 - a_1 - \left(\frac{\sqrt{a_1} + \sqrt{a_2}}{S}\right)^2\right) t \\ + \left(\frac{a_1\sqrt{a_2} + a_2\sqrt{a_1}}{S} - \frac{\sqrt{a_1 a_2}(\sqrt{a_1} + \sqrt{a_2})}{S^3}\right) = 0, \end{aligned} \quad (4.3)$$

both of which lie in the interval $[-\sqrt{a_1}/S, \sqrt{a_2}/S]$.

Proof. The trajectory traced for values of $t \in [-\sqrt{a_1}/S, \sqrt{a_2}/S]$, is a line segment on the simplex. Setting the sum of the coordinates of $(a_i/u_i)_{i=1,\dots,n}$ equal to 1 and clearing denominators gives the quadratic equation above. Since it passes through the local minimum, continuity implies existence of two solutions in the desired interval. \square

Example 4.9. Consider the matrix $A = \text{diag}(1/4, 1/25, 1/36)$. To obtain a completion, one may solve the quadratic equation (4.3), which turns into

$$\frac{77}{90} t^2 + \frac{28}{325} t - \frac{28}{845} = 0$$

giving solutions $t = -\frac{6}{715} \sqrt{586} - \frac{36}{715}$, or $\frac{6}{715} \sqrt{586} - \frac{36}{715}$. Using the latter value, we obtain the matrix completion:

$$\mathbf{u}^T \mathbf{v} = \begin{pmatrix} 0.250 & 0.374 & 0.105 \\ 0.027 & 0.040 & 0.011 \\ 0.066 & 0.098 & 0.028 \end{pmatrix}.$$

5 Algebraic Matroids

Underlying much of the discussion up to now has been an analysis of the algebraic matroid $\mathcal{M}(\Pi_{m \times n})$ associated to the algebraic variety of probability matrices. In particular, the need to check cycles in the completability test and the finiteness of the set of completions for graphs with two connected components are both related to the matroid structure. Later, we will discuss how the complicated matroid structures for the variety of rank- r matrices and the variety of higher-order rank-1 tensors make their analysis less accessible.

In this section, we make the connection to matroid theory explicit. An introduction to matroids and common notation may be found in [Oxl11]. A closer focus on algebraic matroids (particularly the determinantal matroid) may be found in [KRT13]; earlier, its application to low-rank matrix completion was explored in [KTTU12]. For details on the computation of algebraic matroids, one may refer to [Ros14].

5.1 Rank-One Determinantal Matroid

The determinantal matroid is the algebraic matroid associated to the determinantal ideal. In particular, the determinantal ideal $I_{r,m \times n}$ is generated by $\langle (r+1) \times (r+1) \rangle$ -minors of an $m \times n$ matrix of variables x_{ij} , and its vanishing set is the set of $m \times n$

matrices of rank $\leq r$. The matroid encoding the algebraic relationships among the x_{ij} 's is denoted $\mathbf{D}(m \times n, r)$. In the case of the determinantal matroid $\mathbf{D}(m \times n, 1)$, much is clearly understood. In particular, the matroid $\mathbf{D}(m \times n, 1)$ is the graphic matroid on $K_{m,n}$. Some consequences are in the following proposition:

Proposition 5.1. *The following facts hold about $\mathbf{D}(m \times n, 1)$:*

1. Rank of $\mathbf{D}(m \times n, 1) = m + n - 1$.
2. $\mathcal{B} = \{\text{spanning trees of } K_{m,n}\}$.
3. $\mathcal{C} = \{\text{simple cycles of } K_{m,n}\}$.

Circuit polynomials are the essentially unique relations among the elements of each circuit of the matroid. The circuit polynomials of $\mathbf{D}(m \times n, 1)$ are each of the form described in Equation 4.1.

5.2 Probability Matroid

The rank-1 probability matroid is the algebraic matroid associated to the ideal of algebraic relations among the entries $p_{ij} = P(X = x_i, Y = y_j)$ for X, Y independent discrete random variables. Let this ideal be denoted $\Pi_{m \times n}$; it is generated as

$$\Pi_{m \times n} = \left\langle 2 \times 2 \text{ minors}, \sum p_{ij} - 1 \right\rangle;$$

the associated variety is the section of $\mathbf{D}(m \times n, 1)$ with the hyperplane $\sum p_{ij} - 1$.

Proposition 5.2. *The dimension of $\mathcal{V}(\Pi_{m \times n})$ is $m + n - 2$.*

Proof. There is a parametrization of $\mathcal{V}(\Pi_{m \times n})$ analogous to the standard parametrization of the determinantal variety.

$$(a_1, \dots, a_{m-1}, b_1, \dots, b_{n-1}) \mapsto (a_i b_j : 1 \leq i \leq m, 1 \leq j \leq n),$$

where we set $a_m = 1 - \sum_{i=1}^{m-1} a_i$ and $b_n = 1 - \sum_{i=1}^{n-1} b_i$. This map is actually invertible, via:

$$(p_{ij}) \mapsto \left(\sum_j p_{1j}, \dots, \sum_j p_{m-1,j}, \sum_i p_{i1}, \dots, \sum_i p_{i,n-1} \right)$$

The existence of these maps implies that the dimension of the two spaces is equal. \square

The corollary for the matroid is that $\rho(\mathcal{M}(\Pi_{m \times n})) = m + n - 2$; however, the matroid can be much more tightly characterized:

Theorem 5.3. *The matroid $\mathcal{M}(\Pi_{m \times n})$ has bases given by $\mathcal{B} = \{\text{spanning forests with 2 connected components}\}$, and circuits given by $\mathcal{C} = \{\text{spanning trees and non-spanning simple cycles}\}$*

Proof. The bases of a matroid characterize it completely; we claim that spanning forests on two components are the bases of $\mathcal{M}(\Pi_{m \times n})$. From [KRT13], a base of an algebraic matroid is a subset of the coordinates for which the projection map is finite surjective. Then, the algorithm in Section 4.2 returns the finite fiber over any point, since we are not restricted to the simplex in the more general algebraic context. This gives the set of bases, and the set of circuits follows from the two types of relations in the ideal. \square

Remark 5.4. *The matroid-theoretic way of formulating Theorem 5.3 is that $\mathcal{M}(\Pi_{m \times n})$ is the truncation to rank $m+n-2$ of the determinantal matroid $\mathbf{D}(m \times n, 1)$, see [Wel76, Definition 4.1].*

The base degree, as defined in [Ros14], is the cardinality of a generic fiber in a projection onto the base's coordinates. If a base B has an isolated vertex, the base degree is 1; if both components have positive number of edges, the base degree is 2.

The circuit polynomials associated to cycles are the binomials inherited from the determinantal ideal. As for the spanning trees, the circuit polynomials are obtained either by elimination in $\Pi_{m \times n}$, or we complete the matrix using the binomials and then set the sum equal to 1. After clearing denominators, we obtain a polynomial with the circuit polynomial as a factor. This leads us to a bound on the degree of circuit polynomials:

Proposition 5.5. *Let s be the number of edges in a spanning tree C not adjacent to a leaf of C . The degree of a circuit polynomial corresponding to a spanning tree is bounded above by $1 + s$.*

Proof. We can obtain some multiple of the circuit polynomial for the spanning tree by completing the matrix using the binomials and then setting the sum equal to 1. Each term in the sum of entries is a Laurent monomial of total degree 1, such that every appearing variable has degree 1 or -1 . No edge adjacent to a leaf appears in the denominator of a term in this sum, so multiplying by the product of all other variables gives a polynomial expression of degree $1 + s$. Since the polynomial is only a *multiple* of the circuit polynomial, this is only an upper bound. \square

6 Generalizations

6.1 Low-Rank Matrices

One natural direction to generalize these results would be to fix $r > 1$, and find conditions for a matrix to be rank r or nonnegative rank r and have nonnegative entries that sum to 1. One obvious consequence of our results is that any matrix completable to rank 1 is trivially completable as a higher-rank matrix. It is harder to provide tighter conditions, however, even in the smallest examples.

Example 6.1 ($r = 2, m = n = 3$). *There are two polynomials constraining the entries of a 3×3 probability matrix of rank 2: the determinant must be zero, and the sum of the entries must be 1. The variety of matrices with these properties has dimension 7.*

In terms of the matroid, there are two distinct bases up to row permutation and transpose: the set of size 7 obtained by omitting adjacent edges, and the set obtained by omitting non-adjacent edges. To find completions, we substitute X and $R - X$ for the missing entries, where $R = 1 - (a + b + c + d + e + f + g)$:

$$\text{Base 1: } \begin{pmatrix} a & b & c \\ d & e & f \\ g & X & R - X \end{pmatrix} \qquad \text{Base 2: } \begin{pmatrix} a & b & c \\ d & X & f \\ g & e & R - X \end{pmatrix}.$$

Since the sum is now fixed at 1, we only need to check that there is a value of X in $[0, R]$ so that the determinant is 0. The determinant in base 1 gives a linear equation in X , while the determinant in base 2 is a quadratic; the solutions to each are:

$$X = \frac{g(bf - ce) + R(ae - bd)}{(ae - bd) + (af - cd)}$$

$$X = \frac{(aR + bd - cg) \pm \sqrt{(aR + bd - cg)^2 - 4a(b(dR - fg) + e(af - cd))}}{2a}$$

Substituting the values (.07, .09, .09, .12, .15, .04, .16) for the known coordinates of the matrix yields a completion for Base 1, but since the discriminant of the Base 2 determinant is negative, no probability completion is possible.

Since the determinantal matroid is not fully understood for $r > 1$, the results on completability to probability matrices of rank 1 will be difficult to generalize to higher rank. From the statistics viewpoint, it would be more interesting to study completability to probability matrices of nonnegative rank at most r , because the r -th mixture model of two discrete random variables is the semialgebraic set of matrices of nonnegative rank at most r . If a nonnegative matrix has rank 0, 1, or 2, then its nonnegative rank is equal to its rank. Hence, in Example 6.1, we simultaneously address the question of completing a partial matrix to a probability matrix of nonnegative rank 2.

If $r \geq 3$, then matrices of nonnegative rank at most r form a complicated semi-algebraic set. For $r = 3$, a semialgebraic description of this set is given in [KRS13, Theorem 3.1]. Partial matrices that are completable to probability matrices of nonnegative rank at most 3, are coordinate projections of this semialgebraic set. To find all probability completions of nonnegative rank 3 of a partial matrix, one has to find all probability completions of rank 3 and then take the intersection with the semialgebraic set of matrices of nonnegative rank at most 3.

6.2 General Tensors

Generalization to higher-dimensional tensors than the two-vector case brings several challenges. Theorem 3.3 gives a partial result characterizing diagonal tensors. However, the nice bipartite graph structure we had for matrices becomes k -partite hypergraphs with k -hyperedges; notions like connectivity and acyclicity will need to be modified. So, while any rank-1 matrix completability problem was reducible to a diagonal case, the

tensor case does not seem to be reducible in the same way. We record here the results for the smallest case distinct from matrices:

Example 6.2 ($2 \times 2 \times 2$ Tensors). *The variety of $2 \times 2 \times 2$ tensors whose entries sum to 1 is 3-dimensional. For this example, we will only consider the independent sets of the algebraic matroid corresponding to this variety. We use the octahedral symmetry group of the cube to look only at orbits of independent sets:*

1. (Size 1) Any singleton, e.g. p_{000} . The only condition is $p_{000} \leq 1$.
2. (Size 2) Three orbits of pairs:
 - (a) p_{000}, p_{001} : $p_{000} + p_{001} \leq 1$.
 - (b) p_{000}, p_{011} : $\sqrt{p_{000}} + \sqrt{p_{011}} \leq 1$.
 - (c) p_{000}, p_{111} : $\sqrt[3]{p_{000}} + \sqrt[3]{p_{111}} \leq 1$.
3. (Size 3) Three orbits of triples:
 - (a) $p_{000}, p_{001}, p_{010}$: $p_{000} + p_{001} + p_{010} + (p_{001}p_{010}/p_{000}) \leq 1$.
 - (b) $p_{000}, p_{001}, p_{110}$: $\sqrt{p_{000} + p_{001}} + \sqrt{p_{110} + p_{001}p_{110}/p_{000}} \leq 1$.
 - (c) $p_{000}, p_{101}, p_{011}$: The tensor is completable if and only if the equation

$$x^3 + (p_{000} + p_{101} + p_{011} - 1)x^2 + (p_{000}p_{101} + p_{000}p_{011} + p_{101}p_{011})x + p_{000}p_{101}p_{011} = 0$$
 has a root in the interval $[0, 1]$.

In this small example, most partial tensors reduced to a case we knew how to handle; however, 3c does not have a simple semialgebraic description. This equation was obtained by adding x as an entry then completing using minors, and summing the entries to 1.

Acknowledgements. We thank Bernd Sturmfels for suggesting this problem to us and providing detailed feedback on the first draft of this article. We also thank Vishesh Karwa and Aleksandra Slavković for suggesting the problem. We thank Thomas Kahle for sharing his ideas on the project. Kaie Kubjas thanks David Karpuk for helpful conversations. This collaboration was initiated while both authors were guests of the Max-Planck Institute for Mathematics, and was continued at the as2014 conference at Illinois Institute of Technology.

References

- [BPR06] Saugata Basu, Richard Pollack, and Marie-Françoise Roy. *Algorithms in real algebraic geometry*, volume 10 of *Algorithms and Computation in Mathematics*. Springer-Verlag, Berlin, second edition, 2006.
- [KRS13] Kaie Kubjas, Elina Robeva, and Bernd Sturmfels. Fixed points of the EM algorithm and nonnegative rank boundaries. arXiv:1312.5634, 2013.

- [KRT13] Franz J. Király, Zvi Rosen, and Louis Theran. Algebraic matroids with graph symmetry. arXiv:1312.3777, 2013.
- [KTTU12] Franz J. Király, Louis Theran, Ryota Tomioka, and Takeaki Uno. The algebraic combinatorial approach for low-rank matrix completion. arXiv:1211.4116, 2012.
- [NPS08] Jiawang Nie, Pablo A. Parrilo, and Bernd Sturmfels. Semidefinite representation of the k -ellipse. In Alicia Dickenstein, Frank-Olaf Schreyer, and Andrew J. Sommese, editors, *Algorithms in Algebraic Geometry*, volume 146 of *The IMA Volumes in Mathematics and its Applications*, pages 117–132. Springer, New York, 2008.
- [Oxl11] James Oxley. *Matroid theory*. Oxford Graduate Texts in Mathematics. Oxford University Press, Oxford, second edition, 2011.
- [Ros14] Zvi Rosen. Computing algebraic matroids. arXiv:1403.8148, 2014.
- [Wel76] D. J. A. Welsh. *Matroid theory*. Academic Press [Harcourt Brace Jovanovich, Publishers], London-New York, 1976. L. M. S. Monographs, No. 8.

Authors' addresses:

Kaie Kubjas, Aalto Science Institute, Aalto University, PO Box 15500, FI-00076 Aalto, Finland, kaie.kubjas@gmail.com

Zvi Rosen, Department of Mathematics, University of California, Berkeley, CA 94720, USA, zhrosen@berkeley.edu