11:1 2020

ASTAT

Algebraic Statistics

msp

msp.org/astat

See inside back cover or msp.org/astat for submission instructions.

Algebraic Statistics (ISSN 2693-3004 electronic, 2693-2997 printed) at Mathematical Sciences Publishers, 798 Evans Hall #3840, c/o University of California, Berkeley, CA 94720-3840 is published continuously online. Periodical rate postage paid at Berkeley, CA 94704, and additional mailing offices.

AStat peer review and production are managed by EditFlow® from MSP.

PUBLISHED BY
mathematical sciences publishers
nonprofit scientific publishing
http://msp.org/

Algebraic Statistics
Vol. 11, No. 1, 2020
https://doi.org/10.2140/astat.2020.11.1

EDITORIAL: A NEW BEGINNING

THOMAS KAHLE AND SONJA PETROVIĆ

Algebraic statistics and *Algebraic Statistics*

The creation of a field that bridges two disparate areas takes both ingenuity and the ability to generate excitement about new interdisciplinary ideas. For that field to continuously evolve over two decades, expanding to include virtually every aspect of the ground fields, as well as a growing number of neighboring research areas, takes a continued and dedicated community effort. *Algebraic Statistics* (AStat) is being established as a journal to be run by and devoted to such a community, representing interdisciplinary researchers in the field coming from all backgrounds.

While algebra has always played a prominent role in statistics, the publication of a couple of seminal works in the late 1990s defined the new direction by connecting modern computational algebraic geometry and commutative algebra to two critical problems in statistics: sampling from discrete conditional distributions [3] and experimental design [7]. With the onset of the 2000s, the use of these techniques in statistics really took off, generating a large body of research papers and several textbooks [1; 2; 4; 5; 6; 8; 9]. In the last decade, the field has seen a massive influx of new people bringing new ideas and perspectives to problems at the intersection of nonlinear algebra, interpreted in broadest possible sense, and statistics.

The term "algebraic statistics" has thus evolved in meaning to include an ever-expanding list of topics. We understand it as an umbrella term for using algebra (multilinear algebra, commutative algebra, and computational algebra), geometry and combinatorics to obtain insights in mathematical statistics as well as for diverse applications of these tools to data science.

The community of algebraic statisticians is quite an active one, organizing many conferences, symposia, seminars, and special sessions at regional and international meetings, and striving for involvement and representation within both nonlinear algebra and statistics. The predecessor community-run journal, which existed for a decade and published ten volumes, has now been discontinued due to a dispute in ownership with a third party interested in a profit oriented future for the journal. The core of the algebraic statistics community strongly supports the establishment of this new journal. It is a leap forward, a fresh start that takes into account historical lessons learned and seeks to grow and expand the research scope. For this endeavor, we are happy to team up with MSP as a publishing partner that is committed to support academic scholarship and to ensuring the long-term success of our research community.

MSC2020: 62R01.
Keywords: algebraic statistics.

The first volume

The first volume, in two issues, contains eleven papers with a mix that represents algebraic statistics well. Mathematical themes include Gröbner bases, both the standard and non-commutative versions, toric and tropical varieties, numerical nonlinear algebra, holonomic gradient descent, and algebraic combinatorics. On the side of statistics, there are models for diverse types of data, parameter estimation under the likelihood principle, covariance estimation, and time series. Applications covered include computational neuroscience, clustering analysis, engineering, material science, and geology.

(1) The paper "Maximum likelihood estimation of toric Fano varieties" showcases likelihood geometry. Its main result explains how properties of likelihood estimation depend on algebraic and geometric features of the underlying toric models.

(2) Linear covariance models are models for Gaussian random variables with linear constraints on the covariance matrix. The paper "Estimating linear covariance models with numerical nonlinear algebra" addresses the problem of maximum likelihood estimation in these models, the related complexity challenges, and introduces an accompanying `Julia` package.

(3) "Expected value of the one-dimensional earth mover's distance" gives explicit formulas for the expected value of a distance between a pairs of one-dimensional discrete probability distributions using algebraic combinatorics, and discusses applications of it in clustering analysis.

(4) In "Inferring properties of probability kernels from the pairs of variables they involve" the authors discuss how inference about inherently continuous and uncountable probability kernels can be encoded in discrete structures such as lattices.

(5) In computational neuroscience, neural codes model patterns of neuronal response to stimuli. The field provides many open problems for mathematics and statistics. "Minimal embedding dimensions of connected neural codes" address a problem from receptive field coding: the embedding of neural codes in low dimension.

(6) The holonomic gradient method in *Holonomic gradient method for two way contingency tables* is a numerical procedure to approximate otherwise inaccessible likelihood integrals. It is here applied in a discrete situation of contingency tables.

(7) *Algebraic analysis of rotation data* studies a well-known model for rotation data using the tools from non-commutative algebra and the holonomic gradient descent method. It also discusses applications to several areas of science and engineering.

(8) *Maximum likelihood degree of the two-dimensional linear Gaussian covariance model* provides explicit formulas for the number of solutions of likelihood equations in special cases of the same problem as in paper (2).

(9) *Tropical gaussians: a brief survey* takes a tour through the analogues of Gaussian distributions over the tropical semiring. This has applications in, for example, economics and phylogenetics.

(10) *The norm and saturation of a binomial ideal, and applications to Markov bases* connects back to the beginnings of algebraic statistics: Markov bases. Here the focus is on the complexity of Markov bases.

(11) Finally, *Compatibility of distributions in probabilistic models: An algebraic frame and some characterizations* studies the problem when and how two distributions for two sets of variables can be put together to a distribution for the union of the variables and exhibits discrete and algebraic structures in this problem.

Call for submissions

We see AStat as a primary forum serving the broad community in a focused way. As an interdisciplinary endeavor, by definition, a concerted effort will be made for AStat to serve various constituents interested in and interacting with algebraic statistics. Specifically, in our definition, AStat is devoted to algebraic aspects of statistical theory, methodology and applications, seeking to publish a wide range of research and review papers that address one of the following:

- algebraic, geometric and combinatorial insights into statistical models or the behavior of statistical procedures;
- development of new statistical models and methods with interesting algebraic or geometric properties;
- novel applications of algebraic and geometric methods in statistics.

We invite the community to send their best work in algebraic statistics to be considered for publication here. This includes contributions which connect statistical theory, methodology, or application to the world of algebra, geometry, and combinatorics in ways that may not be labeled as traditional.

References

[1] S. Aoki, H. Hara, and A. Takemura, *Markov bases in algebraic statistics*, Springer, 2012.

[2] C. Bocci and L. Chiantini, *An introduction to algebraic statistics with tensors*, Unitext **118**, Springer, 2019.

[3] P. Diaconis and B. Sturmfels, "Algebraic algorithms for sampling from conditional distributions", *Ann. Statist.* **26**:1 (1998), 363–397.

[4] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on algebraic statistics*, Oberwolfach Seminars **39**, Birkhäuser, 2009.

[5] P. Gibilisco, E. Riccomagno, M. P. Rogantin, and H. P. Wynn, *Algebraic and geometric methods in statistics*, Cambridge Univ. Press, 2010.

[6] L. Pachter and B. Sturmfels, *Algebraic statistics for computational biology*, Cambridge Univ. Press, 2005.

[7] G. Pistone, E. Riccomagno, and H. P. Wynn, *Algebraic statistics: computational commutative algebra in statistics*, Monographs on Statistics and Applied Probability **89**, CRC, Boca Raton, FL, 2001.

[8] S. Sullivant, *Algebraic statistics*, Graduate Studies in Mathematics **194**, American Mathematical Society, 2018.

[9] S. Watanabe, *Algebraic geometry and statistical learning theory*, Cambridge Monographs on Applied and Computational Mathematics **25**, Cambridge University Press, 2009.

THOMAS KAHLE (thomas.kahle@ovgu.de), Otto-von-Guericke Universität Magdeburg
SONJA PETROVIĆ (sonja.petrovic@iit.edu), Illinois Institute of Technology
Managing Editors

Algebraic Statistics
Vol. 11, No. 1, 2020
https://doi.org/10.2140/astat.2020.11.5

MAXIMUM LIKELIHOOD ESTIMATION OF TORIC FANO VARIETIES

CARLOS AMÉNDOLA, DIMITRA KOSTA AND KAIE KUBJAS

We study the maximum likelihood estimation problem for several classes of toric Fano models. We start by exploring the maximum likelihood degree for all 2-dimensional Gorenstein toric Fano varieties. We show that the ML degree is equal to the degree of the surface in every case except for the quintic del Pezzo surface with two ordinary double points and provide explicit expressions that allow one to compute the maximum likelihood estimate in closed form whenever the ML degree is less than 5. We then explore the reasons for the ML degree drop using A-discriminants and intersection theory. Finally, we show that toric Fano varieties associated to 3-valent phylogenetic trees have ML degree one and provide a formula for the maximum likelihood estimate. We prove it as a corollary to a more general result about the multiplicativity of ML degrees of codimension zero toric fiber products, and it also follows from a connection to a recent result about staged trees.

1. Introduction

Maximum likelihood estimation (MLE) is a standard approach to parameter estimation, and a fundamental computational task in statistics. Given observed data and a model of interest, the maximum likelihood estimate is the set of parameters that is most likely to have produced the data. Algebraic techniques have been developed for the computation of maximum likelihood estimates for algebraic statistical models [1; 2; 24; 26; 27].

The *maximum likelihood degree* (ML degree) of an algebraic statistical model is the number of complex critical points of the likelihood function over the Zariski closure of the model [9]. It measures the complexity of the maximum likelihood estimation problem on a model. In [27], an algebraic algorithm is presented for computing all critical points of the likelihood function, with the aim of identifying the local maxima in the probability simplex. In the same article, an explicit formula for the ML degree of a projective variety which is a generic complete intersection is derived and this formula serves as an upper bound for the ML degree of special complete intersections. Moreover, a geometric characterization of the ML degree of a smooth variety in the case when the divisor corresponding to the rational function is a normal crossings divisor is given in [9]. In the same paper an explicit combinatorial formula for the ML degree of a toric variety is derived by relaxing the restrictive smoothness assumption and allowing some mild singularities. For an introduction to the geometry behind the MLE for algebraic statistical models for discrete data the interested reader is referred to [29], which includes most of the current results on the MLE problem from the perspective of algebraic geometry as well as statistical motivation.

MSC2020: primary 62F10; secondary 13P25, 14M25, 14Q15 .
Keywords: algebraic statistics, maximum likelihood estimation, maximum likelihood degree, Fano varieties, toric varieties, toric fiber product.

This article is concerned with the problem of MLE on toric Fano varieties. Toric varieties correspond to log-linear models in statistics. Since the seminal papers by L. A. Goodman in the 1970s [21; 22], log-linear models have been widely used in statistics and areas like natural language processing when analyzing cross-classified data in multidimensional contingency tables [6]. The ML degree of a toric variety is bounded above by its degree. Toric Fano varieties provide several interesting classes of toric varieties for investigating the ML degree drop. We focus on studying the maximum likelihood estimation for 2-dimensional Gorenstein toric Fano varieties, the Veronese $(2, 2)$ with different scalings and toric Fano varieties associated to 3-valent phylogenetic trees.

Two-dimensional Gorenstein toric Fano varieties correspond to reflexive polygons and by the classification results there are exactly 16 isomorphism classes of such polygons, see for example [33]. Out of these 16 isomorphism classes five correspond to smooth del Pezzo surfaces and 11 correspond to del Pezzo surfaces with singularities. Our first main result Theorem 3.1 states that the ML degree is equal to the degree of the surface in all cases except for the quintic del Pezzo surface with two ordinary double points. Furthermore, in Table 2, we provide explicit expressions that allow the maximum likelihood estimate to be computed in closed form whenever the ML degree is less than five.

We also explore reasons and bounds for the ML degree drop of a toric variety building on the work of Améndola et al [3]. The critical points of the likelihood function on a toric variety lie in the intersection of the toric variety with a linear space of complementary dimension. By Bézout's theorem, the sum of degrees of irreducible components of this intersection is bounded above by the degree of the toric variety, and hence the ML degree of a toric variety is bounded by its degree. However, not all the points in the intersection contribute towards the ML degree, i.e. the points with a zero coordinate or the sum of coordinates equal to zero are not counted towards the ML degree. In the case of the quintic del Pezzo surface with two ordinary double points, the ML degree drops by two because there are two points in the intersection of the toric variety and the linear space whose coordinates sum to zero, see Example 4.4. These two points do not depend on the observed data by Lemma 4.6. Although we do not see this phenomenon with two-dimensional Gorenstein toric Fano varieties, the ML degree of a toric variety can drop also because the toric variety and the hyperplane intersect nontransversally, and we will see in Sections 4 and 5 that this is often the case.

Buczyńska and Wiśniewski proved that certain varieties associated to 3-valent phylogenetic trees are toric Fano varieties [7]. In phylogenetics, these varieties correspond to the CFN model in the Fourier coordinates. These varieties are examples of codimension zero toric fiber products as defined by Sullivant in [35]. Our second main result is Theorem 5.5 that states that the MLE, ML degree as well as critical points of the likelihood function behave multiplicatively in the case of codimension zero toric fiber product of toric ideals. As a corollary, we obtain that the ML degree of the Buczyńska–Wiśniewski phylogenetic variety associated to a 3-valent tree is one and we get a closed form for the MLE. This result holds for the CFN model only in the Fourier coordinates, as the ML degree of the actual model in the probability coordinates can be much higher. We observe that the result about the CFN model in the Fourier coordinates can be alternatively deduced from the recent work of Duarte, Marigliano and Sturmfels [13], since Buczyńska–Wiśniewski phylogenetic varieties give staged tree models. It follows from the work of

Huh [28] and Duarte, Marigliano and Sturmfels [13] that the ML estimator of a variety of ML degree one is given by a Horn map, i.e. an alternating product of linear forms of specific form, and such models allow a special characterization using discriminantal triples. We discuss the Horn map and the discriminantal triple for Buczyńska–Wiśniewski phylogenetic varieties on 3-valent trees in Example 5.16.

The outline of this paper is the following. In Section 2, we recall preliminaries on maximum likelihood estimation, log-linear models and toric Fano varieties. In Section 3, we study the maximum likelihood estimation for two-dimensional Gorenstein toric Fano varieties. In Section 4, we explore the ML degree drop using A-discriminants and the intersection theory. Finally, Section 5 is dedicated to phylogenetic models and codimension zero toric fiber products.

2. Preliminaries

2.1. ***Maximum likelihood estimation.*** Consider the complex projective space $\mathbb{P}^{n-1}$ with coordinates $(p_1, \ldots, p_n)$. Let X be a discrete random variable taking values on the state space $[n]$. The coordinate p_i represents the probability of the i-th event $p_i = P(X = i)$ where $i = 1, \ldots, n$. Therefore $p_1 + \ldots + p_n = 1$. The set of points in $\mathbb{P}^{n-1}$ with positive real coordinates is identified with the probability simplex

$$\Delta_{n-1} = \{(p_1, \ldots, p_n) \in \mathbb{R}^n : p_1, \ldots, p_n \geq 0 \text{ and } p_1 + \ldots + p_n = 1\} .$$

An algebraic statistical model $\mathcal{M}$ is the intersection of a Zariski closed subset $V \subseteq \mathbb{P}^{n-1}$ with the probability simplex Δ_{n-1}. The data is given by a nonnegative integer vector $(u_1, \ldots, u_n) \in \mathbb{N}^n$, where u_i is the number of times the i-th event is observed.

The maximum likelihood estimation problem aims to find a model point $p \in \mathcal{M}$ which maximizes the likelihood of observing the data u. This amounts to maximizing the corresponding likelihood function

$$L_u(p_1, \ldots, p_n) = \frac{p_1^{u_1} \cdots p_n^{u_n}}{(p_1 + \ldots + p_n)^{(u_1 + \ldots + u_n)}}$$

over the model $\mathcal{M}$. Statistical computations are usually implemented in the affine n-plane $p_1 + \ldots + p_n = 1$. However, including the denominator makes the likelihood function a well-defined rational function on the projective space $\mathbb{P}^{n-1}$, enabling one to use projective algebraic geometry to study its restriction to the variety V.

he likelihood function might not be concave; it can have many local maxima, making the problem of finding or certifying a global maximum difficult. In algebraic statistics, one tries to find all critical points of the likelihood function, with the aim of identifying all local maxima [9; 26; 27]. This corresponds to solving a system of polynomial equations called likelihood equations. These equations characterize the critical points of the likelihood function L_u. We recall that the number of complex solutions to the likelihood equations, which equals the number of complex critical points of the likelihood function L_u over the variety V, is called the maximum likelihood degree (ML degree) of the variety V.

2.2. ***Log-linear models.*** In this article we are studying maximum likelihood estimation of log-linear models. From the algebraic perspective, a log-linear model is a toric variety intersected with a probability

simplex, hence log-linear models are sometimes called toric models. The likelihood function over a log-linear model is concave, although it can have more than one complex critical point over the corresponding toric variety intersected with the plane $p_1 + \ldots + p_n = 1$ (there is exactly one critical point in the positive orthant). This means that in practice, algorithms like iterative proportional fitting (IPF) are used to find the MLE over a log-linear model. The closed form of the solution is in general not rational and to find its algebraic degree one needs to compute the ML degree. It is an open problem whether there is a connection between the convergence rate of IPF and the ML degree of a log-linear model [12, Section 7.3]. The study of the ML degree paired with homotopy continuation methods may speed up the MLE computation with respect to IPF in certain instances, as explored in [3, Section 8].

Definition 2.1. Let $A = (a_{ij}) \in \mathbb{Z}^{(d-1)\times n}$ be an integer matrix. The log-linear model associated to A is

$$\mathcal{M}_A = \{p \in \Delta_{n-1} : \log p \in \text{rowspan}(A)\}.$$

Alternatively, a log-linear model can be defined as the intersection of a toric variety and the probability simplex. Recall that $\theta^{a_j} := \theta_1^{a_{1,j}}\theta_2^{a_{2,j}}\cdots\theta_d^{a_{d-1,j}}$ for $j = 1, \ldots, n$.

Definition 2.2. Let $A = (a_{ij}) \in \mathbb{Z}^{(d-1)\times n}$ be an integer matrix. The toric variety $V_A \subseteq \mathbb{R}^n$ is the Zariski closure of the image of the parametrization map

$$\psi : (\mathbb{C}^*)^d \to (\mathbb{C}^*)^n, (s, \theta_1, \ldots, \theta_{d-1}) \mapsto \left(s\theta^{a_1}, \ldots, s\theta^{a_n}\right).$$

The ideal of V_A is denoted by I_A and called the toric ideal associated to A.

Often the columns of A are lattice points of a lattice polytope $Q \subseteq \mathbb{R}^{d-1}$. In this case we say that V_A is the toric variety corresponding to Q. The log-linear model $\mathcal{M}_A$ is the intersection of the toric variety V_A with the probability simplex Δ_{n-1}. We omit A from the notation whenever it is clear from the context. We conclude this subsection with a characterization of the MLE for log-linear models.

Proposition 2.3 (Corollary 7.3.9 in [36]). *Let A be a $(d-1)\times n$ nonnegative integer matrix and let $u \in \mathbb{N}^n$ be a data vector of size $u_+ = u_1 + \cdots + u_n$. The maximum likelihood estimate over the model $\mathcal{M}_A$ for the data u is the unique solution $\hat{p}$, if it exists, to*

$$\hat{p}_1 + \ldots + \hat{p}_n = 1, \quad A\hat{p} = \frac{1}{u_+}Au \quad \text{and} \quad \hat{p} \in \mathcal{M}_A.$$

Proposition 2.3 is also known as Birch's Theorem. Often we consider V_A as a projective variety in $\mathbb{P}^{n-1}$. The projective version of Proposition 2.3 is given in Section 4. We usually use the affine version when we want to compute the ML degree or find critical points of the likelihood function and the projective version when studying the ML degree drop.

2.3. ***Toric Fano varieties.*** In this section we will provide a brief introduction to toric Fano varieties, the main objects of study in this article. Fano varieties are a class of varieties with a special positive divisor class giving an embedding of each variety into projective space. They were introduced by Giro Fano [16] and have been extensively studied in birational geometry in the context of the minimal model program (see [31], [30]).

Definition 2.4. A complex projective algebraic variety X with ample anticanonical divisor class $-K_X$ is called a *Fano variety*.

Two-dimensional Fano varieties are also known as *del Pezzo surfaces* named after the Italian mathematician Pasquale del Pezzo, who encountered this class of varieties when studying surfaces of degree d embedded in $\mathbb{P}^d$. Throughout this paper we will use the terminology del Pezzo surface to refer to a two-dimensional Fano variety. We note that we do not use the terminology Fano surface, as a Fano surface usually refers to a surface of general type whose points index the lines on a nonsingular cubic threefold, which is not a Fano variety [15].

We will consider Fano varieties that are also toric varieties as defined in Definition 2.2. We first focus on the characterization of two-dimensional Gorenstein toric Fano varieties, i.e. normal toric Fano varieties whose anticanonical divisor K_X is not only an ample divisor but also a Cartier divisor. Isomorphism classes of Gorenstein toric Fano varieties are in bijection with isomorphism classes of reflexive polytopes, which were introduced in [5].

Definition 2.5. A lattice polytope is reflexive if it contains the origin in its interior and its dual polytope is also a lattice polytope.

In particular, toric del Pezzo surfaces are in bijection with two-dimensional reflexive polytopes. The classification of two-dimensional reflexive polytopes can be found for example in [33].

Proposition 2.6 (Section 4 in [33]). *There are exactly* 16 *isomorphism classes of two-dimensional reflexive polytopes, those depicted in Figure 1.*

In Figure 1, the reflexive polytopes are labeled by the number of lattice points on the boundary. On the other hand, the self-intersection number K_S^2 of the canonical class of a del Pezzo surface is the *degree* of the del Pezzo surface, which we denote by d. Here we adopt the approach of [10, Chapter 8.3], where the reflexive polytope is the one corresponding to the anticanonical embedding of the del Pezzo surface. According to [10, Chapter 8.3, Ex. 8.3.8], for each of the 16 reflexive polytopes we obtain exactly the corresponding toric del Pezzo surface. Furthermore, in [11, Chapter 8] the degree of each of these surfaces is given and coincides with the number of lattice points on the boundary. In this way, the projective

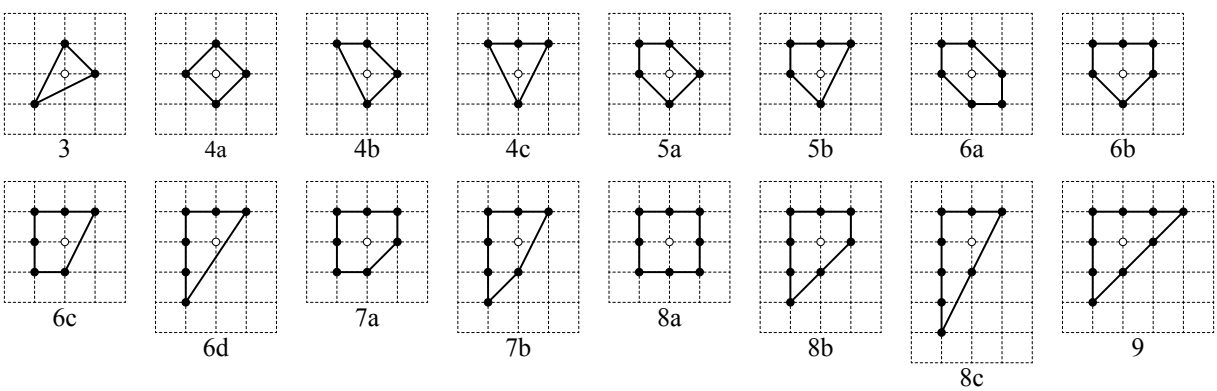


Figure 1. Isomorphism classes of two-dimensional reflexive polytopes.

varieties corresponding to the polytopes labeled by 6a, 7a, 8a, 8b and 9 are smooth and the projective varieties corresponding to the rest of the polytopes in Figure 1 have singularities. The dual of the polytope labeled by number x and letter y is in the isomorphism class of the polytope labeled by number $12-x$ and letter y. This is related to the so-called "12 theorem" for reflexive polytopes of dimension 2 [20].

Remark 2.7. As explained above, in the manner that toric varieties were defined in Definition 2.2, the degree of the toric variety corresponding to a polytope Q and the number of lattice points on the boundary of Q coincide. However, sometimes in the literature (see for instance [8, Example, p. 123]) the dual polytope is used to characterize the isomorphism class of a toric del Pezzo surface. In our setting, the corresponding polytope for $\mathbb{P}^2$ is the polytope 9 in Figure 1 which gives the anticanonical embedding, i.e. the degree 3 Veronese embedding into $\mathbb{P}^9$ using the linear system of cubics.

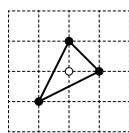
3. MLE of two-dimensional Gorenstein toric Fanos

In this section we determine the ML degree of two-dimensional Gorenstein toric Fano varieties. When the ML degree is less than or equal to three, we reduce the likelihood equations to relatively simple expressions that can be used to compute a closed form for the maximum likelihood estimates. We use the cubic del Pezzo surface as an example to illustrate the MLE derivation. To avoid statistical difficulties, in all of this section we have translated reflexive polygons by a positive vector such that the resulting polygons lie minimally in the positive orthant.

Theorem 3.1. *Let S_d be a two-dimensional Gorenstein toric Fano variety. In Table 1 we determine the ML degree of S_d and show that it is equal to the degree d of the surface in all cases except for the quintic surface S_{5a}. Table 2 provides explicit expressions for computing the maximum likelihood estimate of the algebraic statistical models corresponding to the cubic S_3, the quartics S_{4a}, S_{4b}, S_{4c} and the quintic S_{5a} toric two-dimensional Fano variety.*

Table 1 is constructed using Proposition 2.3 and `Macaulay2` [23]. The results described in Table 1 are in accordance with [29, Theorem 3.2], which states that the ML degree of a projective toric variety is bounded above by its degree. We see in Table 1 that the ML degree drops to three in the case of a quintic del Pezzo surface S_{5a} corresponding to the reflexive polytope 5a in Figure 1. The next section provides an explanation of the ML degree drop in the case of the quintic S_{5a} using the notion of A-discriminant.

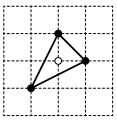
Example 3.2 (singular cubic del Pezzo surface). Consider the reflexive polytope



The corresponding projective toric variety is a cubic surface S_3 in $\mathbb{P}^3$ with three singular points. Its ideal is generated by $I_{S_3} = \langle p_4^3 - p_1 p_2 p_3 \rangle$.

Ideal of degree-d del Pezzo S_d	mldeg
S_3: $p_1p_2p_3 - p_4^3$	3
S_{4a}: $p_2p_4 - p_1p_5,\ p_3^2 - p_1p_5$	4
S_{4b}: $p_2p_4 - p_3^2,\ p_2p_3 - p_1p_5$	4
S_{4c}: $p_2p_4 - p_3^2,\ p_2^2 - p_1p_5$	4
S_{5a}: $p_3p_5 - p_4p_6,\ p_2p_5 - p_6^2,\ p_2p_4 - p_3p_6,\ p_1p_4 - p_6^2,\ p_1p_3 - p_2p_6$	3
S_{5b}: $p_3p_5 - p_4p_6,\ p_2p_5 - p_6^2,\ p_2p_4 - p_3p_6,\ p_1p_4 - p_2p_6,\ p_2^2 - p_1p_3$	5
S_{6a}: $p_4p_6 - p_5p_7,\ p_3p_6 - p_7^2,\ p_2p_6 - p_1p_7,\ p_3p_5 - p_4p_7,\ p_2p_5 - p_7^2$ $p_1p_5 - p_6p_7,\ p_2p_4 - p_3p_7,\ p_1p_4 - p_7^2,\ p_1p_3 - p_2p_7$	6
S_{6b}: $p_5p_6 - p_1p_7,\ p_4p_6 - p_2p_7,\ p_3p_5 - p_4p_7,\ p_2p_5 - p_7^2,\ p_2p_4 - p_3p_7$ $p_1p_4 - p_7^2,\ p_1p_3 - p_2p_7,\ p_2^2 - p_3p_6,\ p_1p_2 - p_6p_7$	6
S_{6c}: $p_6^2 - p_5p_7,\ p_4p_6 - p_3p_7,\ p_3p_6 - p_2p_7,\ p_4p_5 - p_2p_7,\ p_3p_5 - p_2p_6$ $p_2p_4 - p_1p_7,\ p_3^2 - p_1p_7,\ p_2p_3 - p_1p_6,\ p_2^2 - p_1p_5$	6
S_{6d}: $p_6^2 - p_5p_7,\ p_5p_6 - p_4p_7,\ p_3p_6 - p_2p_7,\ p_5^2 - p_4p_6,\ p_3p_5 - p_2p_6$ $p_3p_4 - p_2p_5,\ p_3^2 - p_1p_6,\ p_2p_3 - p_1p_5,\ p_2^2 - p_1p_4$	6
S_{7a}: $p_5p_7 - p_4p_8,\ p_4p_7 - p_3p_8,\ p_2p_7 - p_1p_8,\ p_5p_6 - p_3p_8,\ p_4p_6 - p_3p_7$ $p_2p_6 - p_1p_7,\ p_4p_5 - p_2p_8,\ p_3p_5 - p_1p_8,\ p_4^2 - p_1p_8,\ p_3p_4 - p_1p_7$ $p_2p_4 - p_1p_5,\ p_3^2 - p_1p_6,\ p_7^2 - p_6p_8,\ p_2p_3 - p_1p_4$	7
S_{7b}: $p_7^2 - p_6p_8,\ p_6p_7 - p_5p_8,\ p_4p_7 - p_3p_8,\ p_3p_7 - p_2p_8,\ p_6^2 - p_5p_7$ $p_4p_6 - p_2p_8,\ p_3p_6 - p_2p_7,\ p_4p_5 - p_2p_7,\ p_3p_5 - p_2p_6,\ p_3p_4 - p_1p_8$ $p_2p_4 - p_1p_7,\ p_3^2 - p_1p_7,\ p_2p_3 - p_1p_6,\ p_2^2 - p_1p_5$	7
S_{8a}: $p_8^2 - p_7p_9,\ p_6p_8 - p_5p_9,\ p_5p_8 - p_4p_9,\ p_3p_8 - p_2p_9,\ p_2p_8 - p_1p_9$ $p_6p_7 - p - 4p_9,\ p_5p_7 - p_4p_8,\ p_3p_7 - p_1p_9,\ p_2p_7 - p_1p_8,\ p_6^2 - p_3p_9$ $p_5p_6 - p_2p_9,\ p_4p_6 - p_1p_9,\ p_5^2 - p_1p_9,\ p_4p_5 - p_1p_8,\ p_3p_5 - p_2p_6$ $p_2p_5 - p_1p_6,\ p_4^2 - p_1p_7,\ p_3p_4 - p_1p_6,\ p_2p_4 - p_1p_5,\ p_2^2 - p_1p_3$	8
S_{8b}: $p_8^2 - p_7p_9,\ p_7p_8 - p_6p_9,\ p_5p_8 - p_4p_9,\ p_4p_8 - p_3p_9,\ p_2p_8 - p_1p_9$ $p_7^2 - p_6p_8,\ p_5p_7 - p_3p_9,\ p_4p_7 - p_3p_8,\ p_2p_7 - p_1p_8,\ p_5p_6 - p_3p_8$ $p_4p_6 - p_3p_7,\ p_2p_6 - p_1p_7,\ p_5^2 - p_2p_9,\ p_4p_5 - p_1p_9,\ p_3p_5 - p_1p_8$ $p_4^2 - p_1p_8,\ p_3p_4 - p_1p_7,\ p_2p_4 - p_1p_5,\ p_3^2 - p_1p_6,\ p_2p_3 - p_1p_4$	8
S_{8c}: $p_8^2 - p_7p_9,\ p_7p_8 - p_6p_9,\ p_6p_8 - p_5p_9,\ p_4p_8 - p_3p_9,\ p_3p_8 - p_2p_9$ $p_7^2 - p_5p_9,\ p_6p_7 - p_5p_8,\ p_4p_7 - p_2p_9,\ p_3p_7 - p_2p_8,\ p_6^2 - p_5p_7$ $p_4p_6 - p_2p_8,\ p_3p_6 - p_2p_7,\ p_4p_5 - p_2p_7,\ p_3p_5 - p_2p_6,\ p_4^2 - p_1p_9$ $p_3p_4 - p_1p_8,\ p_2p_4 - p_1p_7,\ p_3^2 - p_1p_7,\ p_2p_3 - p_1p_6,\ p_2^2 - p_1p_5$	8
S_9: $p_9^2 - p_8p_{10},\ p_8p_9 - p_7p_{10},\ p_6p_9 - p_5p_{10},\ p_5p_9 - p_4p_{10},\ p_3p_9 - p_2p_{10}$ $p_8^2 - p_7p_9,\ p_6p_8 - p_4p_{10},\ p_5p_8 - p_4p_9,\ p_3p_8 - p_2p_9,\ p_6p_7 - p_4p_9$ $p_5p_7 - p_4p_8,\ p_3p_7 - p_2p_8,\ p_6^2 - p_3p_{10},\ p_5p_6 - p_2p_{10},\ p_4p_6 - p_2p_9$ $p_3p_6 - p_1p_{10},\ p_2p_6 - p_1p_9,\ p_5^2 - p_2p_9,\ p_4p_5 - p_2p_8,\ p_3p_5 - p_1p_9$ $p_2p_5 - p_1p_8,\ p_4^2 - p_2p_7,\ p_3p_4 - p_1p_8,\ p_2p_4 - p_1p_7,\ p_3^2 - p_1p_6$ $p_2p_3 - p_1p_5,\ p_2^2 - p_1p_4$	9

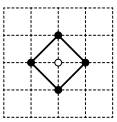
Table 1. ML degrees of 2-dimensional Gorenstein toric Fanos.



$$\text{MLE:}\quad \hat{s}=\frac{(-3x+c)^3}{(x+a)(x+b)},\quad \hat{\theta}_1=\frac{x+a}{-3x+c},\quad \hat{\theta}_2=\frac{x+b}{-3x+c}$$

$$\text{Eq:}\qquad 28x^3+[a+b-27c]x^2+[ab+9c^2]x-c^3=0,$$

$$\text{where}\quad a=\frac{u_1-u_3}{u_+},\quad b=\frac{u_2-u_3}{u_+},\quad c=\frac{3u_3+u_4}{u_+}$$

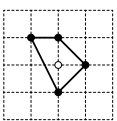


$$\text{MLE:}\quad \hat{s}=\frac{(-x^2+8x+8-(3b^2-4a^2+4))^3}{16(x-1)^2(x+b)(-7x^2+(8a+8)x+(3b^2-4a^2-4-8a))},$$

$$\hat{\theta}_1=\frac{-7x^2+(8a+8)x+(3b^2-4a^2-4-8a)}{2(-x^2+8x-(3b^2-4a^2+4))},\quad \hat{\theta}_2=\frac{4(x+b)(x-1)}{-x^2+8x-(3b^2-4a^2+4)}$$

$$\text{Eq:}\quad 15x^4-16x^3+(8a^2-22b^2-56)x^2+(16(4-4a^2+5b^2))x+8(4a^2-5b^2-2)-(4a^2-3b^2)^2=0$$

$$\text{where}\quad a=\frac{u_1-u_4}{u_+},\quad b=\frac{u_2-u_3}{u_+},\quad c=\frac{2u_3+2u_4+u_5}{u_+}.$$



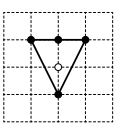
$$\text{MLE:}\qquad \hat{s}=\frac{(4x^2+2(c-1)x+ab)^3}{(1-x)(x^2+(c+1)x+ab+c)(-2x^2-(2c+a)x+a-ab)},$$

$$\hat{\theta}_1=\frac{-2x^2-(a+2c)x+(a-ab)}{4x^2+2(c-1)x+ab},\quad \hat{\theta}_2=\frac{x^2+(c+1)x+(ab+c)}{4x^2+2(c-1)x+ab}$$

$$\text{Eq:}\quad 17x^4+(17c-16)x^3+(3+9ab-8c+4c^2)x^2+(4abc-5ab-c)x+a^2b^2=0$$

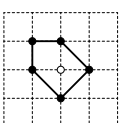
$$\text{where}\quad a=\frac{u_1+2u_4+u_5}{u_+},\quad b=\frac{u_3+2u_4+u_5}{u_+},\quad c=\frac{u_2-3u_4-u_5}{u_+}$$

$$\text{MLE:}\quad \hat{s}=\frac{4x^2(b-x)}{-3x^2-(2a+2)x+(4a+c)b},\quad \hat{\theta}_1=\frac{-3x^2-(2a+2)x+(4a+c)b}{8x^2},\quad \hat{\theta}_2=\frac{2x}{b-x}$$



$$\text{Eq:}\quad -55x^4+12x^3+(c(4a+c)+b(5b-8))x^2-(4b(ab+ac+c))x+(4a+c)b^2c=0$$

$$\text{where}\quad a=\frac{u_2-u_4}{u_+},\quad b=\frac{2u_1+u_5}{u_+},\quad c=\frac{2u_3+4u_4+u_5}{u_+}$$



$$\text{MLE:}\qquad \hat{s}=\frac{(-x^2+cx)(x^2+(a+b)x)}{2x^2-(b+2c)x+bc},$$

$$\hat{\theta}_1=\frac{2x^2-(b+2c)x+bc}{x^2+(a+b)x},\quad \hat{\theta}_2=\frac{2x^2-(b+2c)x+bc}{-x^2+cx}$$

$$\text{Eq:}\qquad -5x^3+(3-5a)x^2+(-a-b(b+5c))x+b^2c=0$$

$$\text{where}\quad a=\frac{u_2-u_3-3u_4-2u_6}{u_+},\quad b=\frac{u_3+u_5+2(u_4+u_6)}{u_+},\quad c=\frac{u_1+u_3+u_6+2u_4}{u_+}$$

Table 2. Explicit forms for the MLE for 2-dimensional Gorenstein toric Fanos, with corresponding polytopes Q_d. "Eq" stands for the polynomial equation of degree $d = \text{mldeg}$.

We are interested in the algebraic statistical model given by the matrix

$$A = \begin{bmatrix} 2 & 1 & 0 & 1 \\ 1 & 2 & 0 & 1 \end{bmatrix}.$$

This nonnegative integer matrix A gives the parametrization map

$$f : \mathbb{C}^3 \to \mathbb{C}^3 \ , \ (s, \theta_1, \theta_2) \mapsto (s\theta_1^2\theta_2, s\theta_1\theta_2^2, s, s\theta_1\theta_2).$$

After applying Birch's theorem, we can write the unique maximum likelihood estimate $\hat{s}, \hat{\theta}$ for the data u as $(\hat{s}, \hat{\theta}_1, \hat{\theta}_2) = (\hat{p}_4^3/(\hat{p}_1\hat{p}_2), \hat{p}_1/\hat{p}_4, \hat{p}_2/\hat{p}_4)$, where

$$\hat{p}_1 = x + a, \quad \hat{p}_2 = x + b, \quad \hat{p}_4 = -3x + c,$$

with $a = \dfrac{u_1 - u_3}{u_+}$, $b = \dfrac{u_2 - u_3}{u_+}$, $c = \dfrac{3u_3 + u_4}{u_+}$ and x is given by

$$28x^3 + [(a + b) - 27c]x^2 + [ab + 9c^2]x - c^3 = 0.$$

Remark 3.3. When the ML degree of the del Pezzo surface is greater than or equal to five, the maximum likelihood estimate $\hat{p}_i$, $i = 1, \ldots, n$ satisfies an equation of degree five or higher. By the Abel–Ruffini theorem there is no algebraic solution for a general polynomial equation of degree five or higher, therefore one would expect that it is not possible to obtain a closed form solution for the maximum likelihood estimate in these cases. However, one can then turn to numerical algebraic geometry methods to compute the MLE (see e.g. [26]).

4. ML degree drop

In order to understand why the ML degree is lower than the degree for the quintic del Pezzo surface 5a, it is useful to think of different embeddings of a toric variety via scalings and how these affect the ML degree. For a full analysis see [3].

Let $Q \subseteq \mathbb{R}^{d-1}$ be a lattice polytope with n lattice points $a_j \in \mathbb{Z}^{d-1}$. Define A to be the $(d-1) \times n$ matrix with the columns $a_1, \ldots, a_n$. A scaling $c \in (\mathbb{C}^*)^n$ can be used to define the parametrization $\psi^c : (\mathbb{C}^*)^d \to (\mathbb{C}^*)^n$ as

$$\psi^c(s, \theta_1, \ldots, \theta_{d-1}) = (c_1 s\theta^{a_1}, \ldots, c_n s\theta^{a_n}).$$

We denote by V^c the Zariski closure of the image of the monomial map ψ^c. The usual parametrization of the toric variety is when $c = (1, \ldots, 1)$. We then denote the corresponding toric variety by $V = V^{(1,\ldots,1)}$.

Definition 4.1. The *ML degree drop* of a scaled toric variety V^c is the difference $\deg(V) - \mathrm{mldeg}(V^c)$.

Define $f_c = \sum_{i=1}^n c_i\theta^{a_i}$ where $c = (c_1, \ldots, c_n) \in (\mathbb{C}^*)^n$.

Definition 4.2. To any matrix A as above, one can associate the variety

$$\nabla_A = \overline{\left\{ c \in (\mathbb{C}^*)^n \mid \exists \theta \in (\mathbb{C}^*)^{d-1} \text{ such that } f_c(\theta) = \frac{\partial f_c}{\partial \theta_i}(\theta) = 0 \text{ for all } i \right\}}.$$

This is the Zariski closure of the set of scalings $c \in (\mathbb{C}^*)^n$ such that the hypersurface $\{f_c = 0\}$ has a singular point in $(\mathbb{C}^*)^{d-1}$. If $\left\{\sum \lambda_j a_j : \lambda_j \in \mathbb{Z}, \sum \lambda_j = 1\right\} = \mathbb{Z}^{d-1}$ — that is, if the affine lattice generated by A is the full integer lattice — then the variety ∇_A is a hypersurface. In this case, it is defined by an irreducible polynomial denoted Δ_A, called the *A-discriminant* [19, Chapter 8].

The main object that determines whether the ML degree drops is the polynomial:

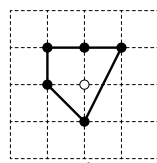
$$E_A(c) = \prod_{\Gamma \text{ face of } Q} \Delta_{\Gamma \cap A}(c) \tag{4-1}$$

where the product is taken over all nonempty faces $\Gamma \subset Q$ including Q itself and $\Gamma \cap A$ is the matrix whose columns correspond to the lattice points contained in Γ. Under certain conditions this is precisely the *principal A-determinant* [19, Chapter 10].

Theorem 4.3 (Theorem 2 in [3]). *Let $V^c \subset \mathbb{P}^{n-1}$ be the scaled toric variety defined by the monomial parametrization with scaling $c \in (\mathbb{C}^*)^n$ fixed. Then* $\operatorname{mldeg}(V^c) < \deg(V)$ *if and only if* $E_A(c) = 0$.

Example 4.4. We will explain why for $c = (1, 1, \ldots, 1)$, the ML degree of the quintic del Pezzo 5b is 5 (and thus equal to its degree), while the ML degree of the quintic del Pezzo 5a is strictly less than 5.

Let us consider first the case of the quintic del Pezzo 5b (see figure on the right). We can label its lattice points and arrange them in the matrix


5b

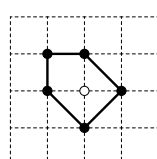
$$A = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 2 \\ 1 & 2 & 0 & 1 & 2 & 2 \end{bmatrix}$$

We have to check that for $c = (1, \ldots, 1)$, $E_A(c) \neq 0$. By (4-1), the polynomial $E_A(c)$ is a product of $\Gamma \cap A$-discriminants. Vertices a_i have $\Delta_{a_i} = 1$. Analogously, edges of lattice length one cannot have nontrivial discriminant (the lattice length of an edge is the number of lattice points contained in the edge minus one). The only potential edge e that may be relevant here is the one of lattice length 2. The corresponding $f_{c,e} = c_{02}y^2 + c_{12}xy^2 + c_{22}x^2y^2$ has a nontrivial singularity if and only if $c_{02} + c_{12}x + c_{22}x^2$ does, thus $\Delta_e(c) = c_{12}^2 - 4c_{02}c_{22}$. Note it is nonzero for $c = (1, \ldots, 1)$.

It only remains to check that $(1, \ldots, 1) \notin \nabla_A$. The following `M2` computation verifies that for $c = (1, \ldots, 1)$, $f_c = y + y^2 + x + xy^2 + x^2y^2 + xy$ has no singularities:

```
R = QQ[x,y]
J = ideal(y+y^2+x+x*y^2+x^2*y^2+x*y, 1+y^2+2*x*y^2+y,
1+2*y+2*x*y+2*x^2*y+x)
gens gb J
```

The last command returns that the Gröbner basis for J is $\{1\}$. Now, for the quintic del Pezzo 5a, we identify the matrix


5a

$$A = \begin{bmatrix} 0 & 0 & 1 & 1 & 1 & 2 \\ 1 & 2 & 0 & 1 & 2 & 1 \end{bmatrix}.$$

All edges are of lattice length one, so we again focus on ∇_A. However, now $c = (1, \ldots, 1) \in \nabla_A$, as the following code verifies.

```
I = ideal(y+y^2+x+x*y^2+x^2*y+x*y, 1+y^2+2*x*y+y, 1+2*y+2*x*y+x^2+x)
gens gb I
```

In this case we get 2 points, the solutions of $x+y=0$, $y^2-y-1=0$, as singularities for $f_c = y+y^2+x+xy^2+x^2y+xy$. The corresponding points of the variety are

$$\begin{aligned}&(1/2(1+\sqrt{5}),\ 1/2(3+\sqrt{5}),\ -1/2(1+\sqrt{5}),\ -1/2(3+\sqrt{5}),\ -2-\sqrt{5},\ 2+\sqrt{5}),\\ &(1/2(1-\sqrt{5}),\ 1/2(3-\sqrt{5}),\ -1/2(1-\sqrt{5}),\ -1/2(3-\sqrt{5}),\ -2+\sqrt{5},\ 2-\sqrt{5}).\end{aligned} \tag{4-2}$$

According to Theorem 4.3, the ML degree must drop for 5a.

Remark 4.5. The singular locus of the quintic del Pezzo S_{5a} consists of the two points $(0,0,1,0,0,0)$ and $(0,0,0,0,0,1)$ which are both rational double points. These points are different from the two points (4-2) that cause the ML degree drop.

Theorem 4.3 characterizes scaling factors c such that the ML degree of V^c is less than the degree of V. All critical points of the likelihood function of V^c lie in the intersection of V with a linear space. In the rest of this section, we will investigate the ML degree drop for a given toric variety V^c by studying this intersection.

Let $L_c(p) = \sum_{i=1}^n c_i p_i$ and $L_{c,i}(p) = \sum_{j=1}^n A_{ij} c_j p_j$ for $i = 1,\dots,d-1$. These polynomials are implicit versions of the polynomials f_c and $\theta_i \frac{\partial f_c}{\partial \theta_i}$ for $i=1,\dots,d-1$. By [3, Proposition 7] the ML degree of V^c is the number of points p in $V\backslash\mathcal{V}(p_1\cdots p_n(c_1p_1+\ldots+c_np_n))$ satisfying

$$(Au)_i L_c(p) = u_+ L_{c,i}(p) \quad \text{for } i=1,\dots,d-1 \tag{4-3}$$

for generic vectors u. Define $\mathcal{L}'_{c,u}$ to be the intersection of V with the solution set of (4-3) and $\mathcal{L}_{c,u}$ to be the intersection of $V\backslash\mathcal{V}(p_1\cdots p_n(c_1p_1+\ldots+c_np_n))$ with the solution set of (4-3). By [18, Example 12.3.1], the sum of degrees of the irreducible components of $\mathcal{L}'_{c,u}$ is at most $\deg V$.

The obvious reason for the ML degree drop comes from removing these irreducible components of $\mathcal{L}'_{c,u}$ that belong to $\mathcal{V}(p_1\cdots p_n(c_1p_1+\ldots+c_np_n))$. We will see in Lemma 4.6 that the irreducible components of $\mathcal{L}'_{c,u}$ that are removed do not depend on u but only on c and the variety V. In the case of the toric del Pezzo surface 5a, the ML degree drop is completely explained by this reason. The degree of this del Pezzo surface is five. The variety $\mathcal{L}'_{c,u}$ consists of two zero-dimensional components of degrees three and two. The degree two component consists of two points (4-2) that lie in the variety $\mathcal{V}(p_1\cdots p_n(c_1p_1+\ldots+c_np_n))$ and hence is removed.

Lemma 4.6. *The points in $\mathcal{L}'_{c,u}\backslash\mathcal{L}_{c,u}$ are independent of u. They are exactly the points $p\in V$ that satisfy $L_c(p)=L_{c,1}(p)=\ldots=L_{c,d-1}(p)=0$.*

Proof. Any $p\in V$ satisfying $L_c(p)=L_{c,1}(p)=\ldots=L_{c,d-1}(p)=0$ is in $\mathcal{L}'_{c,u}\backslash\mathcal{L}_{c,u}$ for any u. Conversely, by the proof of [3, Theorem 13], if $p\in\mathcal{L}'_{c,u}\backslash\mathcal{L}_{c,u}$, then $L_c(p)=0$. It follows from equations (4-3) that then also $L_{c,i}(p)=0$ for $i=1,\dots,d-1$. But then p satisfies (4-3) for any u. □

The more complicated reason for the ML degree drop can be the nontransversal intersection of V and the linear subspace defined by (4-3). We recall that two projective varieties $A, B\subseteq\mathbb{P}^n$ intersect

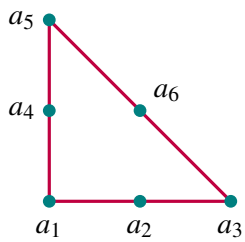


Figure 2. Polytope Q corresponding to the smooth Fano variety $\mathbb{P}^2$.

transversally at $p \in A \cap B$ if p is a smooth point of A, B and

$$T_p A + T_p B = T_p \mathbb{P}^n.$$

The intersection of A and B is generically transverse if it is transverse at a general point of every component of $A \cap B$. If the intersection of V and the linear subspace defined by (4-3) is not generically transverse, the sum of degrees of the irreducible components of $\mathcal{L}'_{c,u}$ can be less than $\deg(V)$, in which case also the ML degree of the toric variety V^c is less than the degree of the toric variety V. One could think that the intersection of V and the linear subspace defined by (4-3) is generically transverse for generic vectors u, but since the linear subspace defined by (4-3) depends on the variety V, then the intersection is not necessarily generically transverse. We will see several such examples later in this section and in Section 5. We note that the sum of degrees of the irreducible components can be less that $\deg V$ even if the degrees are counted with multiplicity as in [18, Example 12.3.1].

Corollary 4.7. *The ML degree drop* $\deg(V) - \mathrm{mldeg}(V^c)$ *is bounded below by the sum of degrees of the irreducible components of the intersection of* V *and the linear subspace defined by* $L_c(p) = L_{c,1}(p) = \ldots = L_{c,d-1}(p) = 0$. *If the intersection of* V *and the linear subspace defined by* (4-3) *is generically transverse, then this bound is exact.*

In Corollary 4.7, we consider only irreducible components whose ideals are different from $\langle p_1, \ldots, p_n \rangle$ as we work over the projective space.

Proof. The sum of degrees of the irreducible components of $\mathcal{L}'_{c,u}$ is at most $\deg(V)$ by [18, Example 12.3.1] and the number of elements of $\mathcal{L}_{c,u}$ is $\mathrm{mldeg}(V^c)$. By Lemma 4.6, we obtain $\mathcal{L}_{c,u}$ from $\mathcal{L}'_{c,u}$ by removing all the irreducible components that satisfy $L_c(p) = L_{c,1}(p) = \ldots = L_{c,d-1}(p) = 0$. Hence the difference of the sum of degrees of the irreducible components of $\mathcal{L}'_{c,u}$ and the number of elements of $\mathcal{L}_{c,u}$ is the sum of degrees of the irreducible components of the intersection of V and the linear subspace defined by $L_c(p) = L_{c,1}(p) = \ldots = L_{c,d-1}(p) = 0$. If V and the linear subspace defined by (4-3) intersect generically transversely, then the sum of degrees of the irreducible components of $\mathcal{L}'_{c,u}$ is equal to $\deg V$. □

To understand the above observations, we analyze the different ML degree drops corresponding to the quadratic Veronese embedding of $\mathbb{P}^2$ given by the Fano polytope in Figure 2.

In [3, Example 26], it was shown that different scalings $c \in \mathbb{R}^6$ produce ML degrees ranging from 1 to 4, under several combinations of the vector c lying on each of the discriminants making up the principal

A-determinant, defined by

$$E_A(c) = \Delta_A(c) \cdot \Delta_{[a_1\, a_2\, a_3]}(c) \cdot \Delta_{[a_3\, a_5\, a_6]}(c) \cdot \Delta_{[a_1\, a_4\, a_5]}(c) \tag{4-4}$$
$$= \det(C) \det\begin{bmatrix} 2c_{00} & c_{10} \\ c_{10} & 2c_{20} \end{bmatrix} \det\begin{bmatrix} 2c_{20} & c_{11} \\ c_{11} & 2c_{02} \end{bmatrix} \det\begin{bmatrix} 2c_{00} & c_{01} \\ c_{01} & 2c_{02} \end{bmatrix}, \quad \text{where } C = \begin{bmatrix} 2c_{00} & c_{10} & c_{01} \\ c_{10} & 2c_{20} & c_{11} \\ c_{01} & c_{11} & 2c_{02} \end{bmatrix}.$$

The different combinations are presented in [3, Table 2], which we reproduce here in Table 3 (while fixing some typos). In each line, we go further than identifying a possible drop and actually explain the exact drops observed.

Naively, each appearance of a 0 in a row of Table 3 makes the ML degree drop by 1. But this cannot be, since the last row has all four zeros and the ML degree cannot drop to 0. We will see in the explanation of the last two rows that it is in general impossible to predict the exact drop just from knowing in what discriminants the vector c lies.

C	Δ_A	$\Delta_{[a_1\, a_2\, a_3]}$	$\Delta_{[a_3\, a_5\, a_6]}$	$\Delta_{[a_1\, a_4\, a_5]}$	mldeg
$\begin{bmatrix} 2 & 1 & 1 \\ 1 & 2 & 1 \\ 1 & 1 & 2 \end{bmatrix}$	$\neq 0$	$\neq 0$	$\neq 0$	$\neq 0$	**4**
$\begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 3 \\ 1 & 3 & 2 \end{bmatrix}$	$\neq 0$	0	$\neq 0$	$\neq 0$	**3**
$\begin{bmatrix} 2 & 2 & 1 \\ 2 & 2 & 2 \\ 1 & 2 & 2 \end{bmatrix}$	$\neq 0$	0	0	$\neq 0$	**2**
$\begin{bmatrix} -2 & 2 & 2 \\ 2 & -2 & 2 \\ 2 & 2 & -2 \end{bmatrix}$	$\neq 0$	0	0	0	**1**
$\begin{bmatrix} 17 & 22 & 27 \\ 22 & 29 & 36 \\ 27 & 36 & 45 \end{bmatrix}$	0	$\neq 0$	$\neq 0$	$\neq 0$	**3**
$\begin{bmatrix} 2 & 3 & 3 \\ 3 & 5 & 5 \\ 3 & 5 & 5 \end{bmatrix}$	0	$\neq 0$	0	$\neq 0$	**2**
$\begin{bmatrix} 2 & 2 & 2 \\ 2 & 2 & 2 \\ 2 & 2 & 2 \end{bmatrix}$	0	0	0	0	**1**

Table 3. ML degrees for different scalings c_{ij} in the matrix C.

- **Row 1** This corresponds to the generic case. The intersection $\mathcal{L}'_{c,u}$ is transverse and zero-dimensional with 4 points, corresponding to the ML degree. There are no points in $\mathcal{L}'_{c,u} \backslash \mathcal{L}_{c,u}$ and there is no drop.
- **Row 2** When computing the points in $V_A \cap \{L_c(p) = L_{c,1}(p) = \ldots = L_{c,d-1}(p) = 0\}$ we obtain the unique projective point $[1 : -1 : 1 : 0 : 0 : 0]$, which makes the A-discriminant of the edge $[a_1, a_2, a_3]$ of Q vanish. Removing this point gives the ML degree of $3 = 4 - 1$.
- **Row 3** Now we have one more point in the removal set: apart from the one in the above row, there is also $[0 : 0 : 1 : 0 : 1 : -1]$ on the zero locus of the A-discriminant of the edge $[a_3, a_5, a_6]$. The drop is accounted for exactly these two points and we have ML degree $2 = 4 - 2$.
- **Row 4** There are three points in $\mathcal{L}'_{c,u} \backslash \mathcal{L}_{c,u}$ that lie on the zero loci of edge A-discriminants: $[1 : 1 : 1 : 0 : 0 : 0]$ for the edge $[a_1, a_2, a_3]$, $[0 : 0 : 1 : 0 : 1 : 1]$ for the edge $[a_3, a_5, a_6]$ and $[1 : 0 : 0 : 1 : 1 : 0]$ for the edge $[a_1, a_4, a_5]$. They explain the drop in ML degree $1 = 4 - 3$.
- **Row 5** The only removal point is $[1 : -2 : 4 : 1 : 1 : -2]$, which does *not* lie on zero loci of any of the A-discriminants of the edges of Q, but only on the zero locus of the A-discriminant of the whole of Q. Removing this point gives the ML degree of $3 = 4 - 1$.
- **Row 6** This is the first time that the removal ideal $I_A + \langle L_c(p), L_{c,1}(p), \ldots, L_{c,d-1}(p) \rangle$ is not radical. While there is only one point, $[0 : 0 : 1 : 0 : 1 : -1]$, its multiplicity is 2. We used the `Macaulay2` package `SegreClasses` [25] to compute the multiplicity. The intersection $\mathcal{L}'_{c,u}$ is zero-dimensional but consists of two components of degree 2. The first component is prime and corresponds to the two points in the ML degree 2. The toric variety and the linear space defined by (4-3) intersect transversely at both points of the first component. The second component is primary, but not prime. Its radical $\langle p_6 + p_5, p_4, p_3 + p_5, p_2, p_1 \rangle$ is a zero-dimensional ideal of degree 1, corresponding to the above point that lies on the zero locus of the A-discriminant of the edge $[a_3, a_5, a_6]$.

  Although the intersection of the toric variety and the linear space defined by (4-3) is dimensionally transverse, it is not transverse at the point defined by the second component. We also observe that while $\Delta_A(c) = 0$, there is no singular point of f_c, which means c lies strictly in the closure in Definition 4.2 (see Remark 4.8 below).
- **Row 7** Now the removal ideal is 1-dimensional of degree 2. It is given by

$$\langle p_2^2 + p_2 p_3 + p_3 p_4\,,\ p_1 + p_2 + p_4\,,\ p_2 + p_5 - p_2 - p_3\,,\ p_2 + p_3 + p_6 \rangle .$$

  Its variety intersects $V_{\Gamma \cap A}$ in one point for each edge Γ of Q. In other words, the reason why all discriminants $\Delta_{[a_1\, a_2\, a_3]}$, $\Delta_{[a_3\, a_5\, a_6]}$, $\Delta_{[a_1\, a_4\, a_5]}$ vanish is that the removal set intersects the planes $p_1 = p_2 = p_4 = 0\,,\ p_2 = p_3 = p_6 = 0$ and $p_4 = p_5 = p_6 = 0$ respectively, and one can find in each a point with complementary support. Furthermore, it intersects the open set where none of the p_i are zero, which explains why $\Delta_A = 0$ too. Unfortunately, this alone does not explain why the ML degree is 1.

  By looking at the intersection ideal of the toric variety V_A with the equations (4-3), we realize that the intersection is not transverse (not even dimensionally transverse). Indeed, there are two

components: a zero-dimensional component of degree 1 (corresponding to the MLE) and a one-dimensional component of degree 2 (so the sum of the degrees is $1+2=3<4$). This last component matches the removal ideal above. At the 0-dimensional component the toric variety intersects the linear space defined by (4-3) transversely. At a generic point of the 1-dimensional component the intersection is not transverse. Both components have multiplicity one and hence also the sum of degrees counted with multiplicity is less than four.

Remark 4.8. If for some scaling c, the A-discriminant of at least one edge is zero and the A-discriminant of at least one edge is nonzero, then there is no singular point $\theta \in (\mathbb{C}^*)^2$ of f_c. Indeed, such a point $\theta = (\theta_1, \theta_2) \in (\mathbb{C}^*)^2$ would need to satisfy

$$\begin{bmatrix} 2c_{00} & c_{10} & c_{01} \\ c_{10} & 2c_{20} & c_{11} \\ c_{01} & c_{11} & 2c_{02} \end{bmatrix} \begin{bmatrix} 1 \\ \theta_1 \\ \theta_2 \end{bmatrix} = \begin{bmatrix} 0 \\ 0 \\ 0 \end{bmatrix}. \tag{4-5}$$

Say $\Delta_{[a_3\, a_5\, a_6]} = \det \left[\begin{smallmatrix} 2c_{20} & c_{11} \\ c_{11} & 2c_{02} \end{smallmatrix}\right] = 0$. In order for the system (4-5) to be consistent, $\left[\begin{smallmatrix} c_{10} & 2c_{20} & c_{11} \\ c_{01} & c_{11} & 2c_{02} \end{smallmatrix}\right]$ should have rank 1. In particular, $\det \left[\begin{smallmatrix} c_{10} & c_{01} \\ 2c_{20} & c_{11} \end{smallmatrix}\right] = 0$ and $\det \left[\begin{smallmatrix} c_{10} & c_{01} \\ c_{11} & 2c_{02} \end{smallmatrix}\right] = 0$, which in turn means that both $\left[\begin{smallmatrix} 2c_{00} & c_{10} & c_{01} \\ c_{10} & 2c_{20} & c_{11} \end{smallmatrix}\right]$ and $\left[\begin{smallmatrix} 2c_{00} & c_{10} & c_{01} \\ c_{01} & c_{11} & 2c_{02} \end{smallmatrix}\right]$ have rank 1. We conclude that C itself must have rank 1, so that c must lie in the discriminants of all the edges (as in Row 7), contradicting that one of them was nonzero. This means that if we are in the case of Row 6, even though $\Delta_A(c) = 0$, the scaling $c \in \nabla_A$ is added in the closure that appears in Definition 4.2.

Conjecture 4.9. *The intersection of V and the linear space given by (4-3) is generically transverse at the irreducible component of $\mathcal{L}'_{c,u}$ that gives the MLE.*

This conjecture holds for all the examples considered in this section. At other irreducible components the intersection may or may not be transverse.

Remark 4.10. For toric models,

$$\operatorname{mldeg}(V) = \chi_{\text{top}}(V \backslash H) = \chi_{\text{top}}(V) - \chi_{\text{top}}(V \cap H) \tag{4-6}$$

where $H = \{p \mid p_1 \cdots p_n(p_1 + \cdots + p_n) = 0\}$ and χ_{top} is the topological Euler characteristic. The first equality was proved by Huh and Sturmfels [29, Theorem 3.2], and the second equality is the excision formula. The Euler characteristic of toric del Pezzo surfaces can be computed by $\chi_{\text{top}}(V) = 2 + \operatorname{rkPic}(V)$. The rank of the Picard group $\operatorname{rkPic}(V)$ can be computed taking into account that the minimal resolution of every del Pezzo surface is a product of two projective lines $\mathbb{P}^1 \times \mathbb{P}^1$ (polytope 8a in Figure 1), the quadric cone $\mathbb{P}(1, 1, 2) \subset \mathbb{P}^3$ (polytope 8c in Figure 1), or the blow-up of a projective plane in $9 - d$ points in almost general position; namely at most three of which are collinear, at most six of which lie on a conic, and at most eight of them on a cubic having a node at one of the points. We refer the reader to [11] for a more detailed study of this classical subject of algebraic geometry. It would be interesting to explore the ML degree drop further from this perspective.

5. Toric fiber products and phylogenetic models

In [7], Buczyńska–Wiśniewski study an infinite family of toric Fano varieties that correspond to 3-valent phylogenetic trees. These Fano varieties are of index 4 with Gorenstein terminal singularities. In phylogenetics, these varieties correspond to the CFN model in the Fourier coordinates. We define them through the corresponding polytopes.

Definition 5.1. Let $\mathcal{T}$ be a 3-valent tree, i.e. every vertex of $\mathcal{T}$ has degree 1 or 3. Consider all labelings of the edges of $\mathcal{T}$ with 0's and 1's such that at every inner vertex the sum of labels on the incident edges is even. Define $P_{\mathcal{T}} \subseteq \mathbb{R}^E$ to be the convex hull of such labelings. Let $I_{\mathcal{T}}$ be the homogeneous ideal and $V_{\mathcal{T}}$ be the projective toric variety corresponding to $P_{\mathcal{T}}$.

Example 5.2. If $\mathcal{T}$ is tripod, then $P_{\mathcal{T}} = \operatorname{conv}((0,0,0),(1,1,0),(1,0,1),(0,1,1))$ is a simplex and $V_{\mathcal{T}} = \mathbb{P}^3$ is the 3-dimensional projective space.

Example 5.3. If $\mathcal{T}$ is the unique 3-valent four leaf tree, then

$$P_{\mathcal{T}} = \operatorname{conv}((0,0,0,0,0),(1,1,0,0,0),(0,0,0,1,1),(1,1,0,1,1),\\ (1,0,1,1,0),(1,0,1,0,1),(0,1,1,1,0),(0,1,1,0,1)).$$

The ideal $I_{\mathcal{T}}$ is generated by the two quadratic polynomials $x_{00000}x_{11011} - x_{11000}x_{00011}$ and $x_{10110}x_{01101} - x_{01110}x_{10101}$.

The aim of the rest of the section is to show that if $\mathcal{T}$ is a 3-valent tree then the ML degree of the variety $V_{\mathcal{T}}$ is one. We will also give a closed form for its maximum likelihood estimate. This result will be a special case of a more general result about ML degrees of codimension-0 toric fiber products of toric ideals. A toric fiber product can be defined for any two ideals that are homogenous by the same multigrading [35], however, we use the definition specific to toric ideals [14, Section 2.3]. Besides phylogenetic models considered in this section, codimension-0 toric fiber products appear in general group-based models in the Fourier coordinates and reducible hierarchical models (see [35, Section 3] for details on applications).

Let $r \in \mathbb{N}$ and $s_i, t_i \in \mathbb{N}$ for $1 \leq i \leq r$. We consider toric ideals corresponding to vector configurations $\mathcal{B} = \{b^i_j : i \in [r],\ j \in [s_i]\} \subseteq \mathbb{Z}^{d_1}$ and $\mathcal{C} = \{c^i_k : i \in [r], k \in [t_i]\} \subseteq \mathbb{Z}^{d_2}$. These toric ideals are denoted by $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$, they live in the polynomial rings $\mathbb{R}[x^i_j : i \in [r],\ j \in [s_i]]$ and $\mathbb{R}[y^i_k : i \in [r], k \in [t_i]]$, and they are required to be homogeneous with respect to the multigrading by $\mathcal{A} = \{a^i : i \in [r]\} \subseteq \mathbb{Z}^d$. We assume that there exists $\omega \in \mathbb{Q}^d$ such that $\omega a^i = 1$ for all i, so that $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$ are homogeneous also with respect to the standard grading. Toric ideals $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$ being homogeneous with respect to the multigrading by $\mathcal{A}$ implies that there exist linear maps $\pi_1 : \mathbb{Z}^{d_1} \to \mathbb{Z}^d$ and $\pi_2 : \mathbb{Z}^{d_2} \to \mathbb{Z}^d$ such that $\pi_1(b^i_j) = a^i$ and $\pi_2(c^i_k) = a^i$. We define the vector configuration

$$\mathcal{B} \times_{\mathcal{A}} \mathcal{C} = \{(b^i_j, c^i_k) : i \in [r],\ j \in [s_i],\ k \in [t_i]\}.$$

The toric fiber product of $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$ with respect to the multigrading by $\mathcal{A}$ is defined as

$$I_{\mathcal{B}} \times_{\mathcal{A}} I_{\mathcal{C}} := I_{\mathcal{B} \times_{\mathcal{A}} \mathcal{C}},$$

and it lives in the polynomial ring $\mathbb{R}[z^i_{jk} : i \in [r],\ j \in [s_i],\ k \in [t_i]]$.

Example 5.4. Let $\mathcal{T}$ be a 3-valent tree with $n \geq 4$ leaves. Write $\mathcal{T}$ as a union of two trees $\mathcal{T}_1$ and $\mathcal{T}_2$ that share an interior edge e. Take $r = 2$. Let b^1_j be the vertices of $P_{\mathcal{T}_1}$ that have label 0 on edge e and let b^2_j be the vertices of $P_{\mathcal{T}_1}$ that have label 1 on edge e. Define similarly c^1_k and c^2_k for $P_{\mathcal{T}_2}$. Let $\mathcal{A} = \{(0, 1), (1, 0)\}$, so that π_1 maps b^1_j to $(0, 1)$ and b^2_j to $(1, 0)$. Then $I_{\mathcal{T}}$ is the toric fiber product of $I_{\mathcal{T}_1}$ and $I_{\mathcal{T}_2}$ with respect to multigrading by $\mathcal{A}$. In [35, Section 3.4] the toric fiber product construction is explained in full generality for group-based phylogenetic models in the Fourier coordinates.

Given a vector u indexed by the elements of $\mathcal{B} \times_{\mathcal{A}} \mathcal{C}$, we denote its entries by u^i_{jk} for $i \in [r]$, $j \in [s_i]$, $k \in [t_i]$. We define $u^i_{++} = \sum_{j\in[s_i],k\in[t_i]} u^i_{jk}$, $u^i_{j+} = \sum_{k\in[t_i]} u^i_{jk}$ and $u^i_{+k} = \sum_{j\in[s_i]} u^i_{jk}$. We denote the corresponding vectors by $u_{\mathcal{A}}$, $u_{\mathcal{B}}$ and $u_{\mathcal{C}}$ since they are indexed by the elements of $\mathcal{A}$, $\mathcal{B}$ and $\mathcal{C}$. These vectors $u_{\mathcal{A}} = (u^i_{++})_{i\in[r]}$, $u_{\mathcal{B}} = (u^i_{j+})_{i\in[r],j\in[s_i]}$ and $u_{\mathcal{C}} = (u^i_{+k})_{i\in[r],k\in[t_i]}$ are marginal sums. We also define $u^+_{++} = \sum_{i\in[r],j\in[s_i],k\in[t_i]} u^i_{jk}$, $(u_{\mathcal{B}})^+_+ = \sum_{i\in[r]} \sum_{j\in[s_i]} (u^i_{j+})$ and $(u_{\mathcal{C}})^+_+ = \sum_{i\in[r]} \sum_{k\in[t_i]} (u^i_{+k})$. Similarly, if p^i_{jk} is a joint probability distribution indexed by the elements of $\mathcal{B} \times_{\mathcal{A}} \mathcal{C}$, the sum of the joint probability distribution over $\mathcal{A}$ (resp. $\mathcal{B}$, $\mathcal{C}$) is a marginal probability distribution and we denote it by $p_{\mathcal{A}} = (p^i_{++})_{i\in[r]}$ (resp. $p_{\mathcal{B}} = (p^i_{j+})_{i\in[r],j\in[s_i]}$, $p_{\mathcal{C}} = (p^i_{+k})_{i\in[r],k\in[t_i]}$).

Theorem 5.5. *Let $\mathcal{A}$ consist of linearly independent vectors. Then the ML degree of $I_{\mathcal{B}} \times_{\mathcal{A}} I_{\mathcal{C}}$ is equal to the product of the ML degrees of $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$. For a data vector u, every critical point of the likelihood function has the form*

$$\overline{p}^i_{jk} := \frac{(\overline{p_{\mathcal{B}}})^i_j (\overline{p_{\mathcal{C}}})^i_k}{(\overline{p_{\mathcal{A}}})^i}, \tag{5-1}$$

where $\overline{p_{\mathcal{A}}}$, $\overline{p_{\mathcal{B}}}$ and $\overline{p_{\mathcal{C}}}$ are critical points of the likelihood function for the models $I_{\mathcal{A}}$, $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$ and data vectors $u_{\mathcal{A}}$, $u_{\mathcal{B}}$ and $u_{\mathcal{C}}$, respectively. Since the elements of $\mathcal{A}$ are linearly independent, $\overline{p_{\mathcal{A}}}$ is in fact the normalized $u_{\mathcal{A}}$. Moreover, we obtain the maximum likelihood estimate of $I_{\mathcal{B}} \times_{\mathcal{A}} I_{\mathcal{C}}$ by taking $\overline{p_{\mathcal{A}}}$, $\overline{p_{\mathcal{B}}}$ and $\overline{p_{\mathcal{C}}}$ to be the maximum likelihood estimates of the models $I_{\mathcal{A}}$, $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$.

Theorem 5.5 generalizes known results about reducible hierarchical [34, Proposition 4.14] and discrete graphical models. In particular, one recovers the rational formula for the MLE in the case of decomposable graphical models (indeed, they have ML degree one) [34, Proposition 4.18] and general discrete graphical models [17, Theorem 1]. The latter result is for mixed graphical interaction models that allow both discrete and continuous random variables. Theorem 5.5 generalizes the case when all variables are discrete.

To prove Theorem 5.5, we first have to recall how to obtain a generating set for $I_{\mathcal{B}} \times_{\mathcal{A}} I_{\mathcal{C}}$ from the generating sets for $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$. Let

$$f = x^{i_1}_{j^1_1} x^{i_2}_{j^1_2} \cdots x^{i_d}_{j^1_d} - x^{i_1}_{j^2_1} x^{i_2}_{j^2_2} \cdots x^{i_d}_{j^2_d} \in \mathbb{R}[x^i_j].$$

For $k = (k_1, k_2, \ldots, k_d) \in [t_{i_1}] \times [t_{i_2}] \times \cdots \times [t_{i_d}]$ define

$$f_k = z^{i_1}_{j^1_1 k_1} z^{i_2}_{j^1_2 k_2} \cdots z^{i_d}_{j^1_d k_d} - z^{i_1}_{j^2_1 k_1} z^{i_2}_{j^2_2 k_2} \cdots z^{i_d}_{j^2_d k_d} \in \mathbb{R}[z^i_{jk}].$$

Let $T_f = \prod_{l=1}^{d}[t_{i_l}]$. For a set of binomials F, we define

$$\text{Lift}(F) = \{f_k : f \in F, k \in T_f\}.$$

We also define

$$\text{Quad} = \{z^i_{j_1k_1}z^i_{j_2k_2} - z^i_{j_1k_2}z^i_{j_2k_1} : i \in [r],\, j_1,\, j_2 \in [s_i],\, k_1,\, k_2 \in [t_i]\}.$$

Proposition 5.6 ([35], Corollary 14). *Let $\mathcal{A}$ consist of linearly independent vectors. Let F be a generating set of $I_\mathcal{B}$ and let G be a generating set of $I_\mathcal{C}$. Then $I_\mathcal{B} \times_\mathcal{A} I_\mathcal{C}$ is generated by*

$$\mathit{Lift}(F) \cup \mathit{Lift}(G) \cup \mathit{Quad}.$$

Example 5.7. The 3-valent four leaf tree $\mathcal{T}_4$ is the union of two tripods $\mathcal{T}_3$ that share an interior edge e. By Proposition 5.6, a generating set for $I_{\mathcal{T}_4}$ is given by the lifts of generating sets for $I_{\mathcal{T}_3}$ and by quadrics with respect to the edge e. Since $I_{\mathcal{T}_3} = \langle 0\rangle$, its lift is $\{0\}$. The set Quad consists of $x_{00000}x_{11011} - x_{11000}x_{00011}$ and $x_{10110}x_{01101} - x_{10101}x_{01110}$ that are generators of $I_{\mathcal{T}_4}$.

Example 5.8. The 3-valent five leaf tree $\mathcal{T}_5$ is the union of the 3-valent four leaf tree $\mathcal{T}_4$ and tripod $\mathcal{T}_3$ that share an interior edge e. The fifth index of a variable $x_{e_1e_2e_3e_4e_5}$ in the coordinate ring of $\mathcal{T}_4$ and the first index of a variable $x_{e_5e_6e_7}$ in the coordinate ring of $\mathcal{T}_3$ correspond to the edge e. Recall that a generating set of $I_{\mathcal{T}_4}$ is $F = \{x_{00000}x_{11011} - x_{11000}x_{00011},\, x_{10110}x_{01101} - x_{01110}x_{10101}\}$ and a generating set of $I_{\mathcal{T}_3}$ is $G = \{0\}$. Both elements of F have four lifts corresponding to $k = (000, 110)$, $k = (000, 101)$, $k = (011, 110)$ and $k = (011, 101)$. Hence Lift(F) consists of

$$\begin{aligned}
&x_{0000000}x_{1101110} - x_{1100000}x_{0001110},\, x_{0000000}x_{1101101} - x_{1100000}x_{0001101},\\
&x_{0000011}x_{1101110} - x_{1100011}x_{0001110},\, x_{0000011}x_{1101101} - x_{1100011}x_{0001101},\\
&x_{1011000}x_{0110110} - x_{0111000}x_{1010110},\, x_{1011000}x_{0110101} - x_{0111000}x_{1010101},\\
&x_{1011011}x_{0110110} - x_{0111011}x_{1010110},\, x_{1011011}x_{0110101} - x_{0111011}x_{1010101},
\end{aligned}$$

and Lift$(G) = \{0\}$. The set Quad consists of 12 polynomials.

To prove Theorem 5.5, we also need the following lemmas.

Lemma 5.9. *For any $u \in \mathbb{R}^{\mathcal{B}\times_\mathcal{A}\mathcal{C}}$, we have $(\mathcal{B} \times_\mathcal{A} \mathcal{C})u = (\mathcal{B}u_\mathcal{B}, \mathcal{C}u_\mathcal{C})$.*

Proof. Assume that the r-th row comes from the $\mathcal{B}$ part of the matrix $\mathcal{B} \times_\mathcal{A} \mathcal{C}$. Then the r-th row of $\mathcal{B} \times_\mathcal{A} \mathcal{C}$ multiplied with u gives

$$\sum_{i\in[r],\, j\in[s_i],k\in[t_i]} (b^i_j, c^i_k)_r u^i_{jk} = \sum_{i\in[r],\, j\in[s_i],k\in[t_i]} (b^i_j)_r u^i_{jk} = \sum_{i\in[r],\, j\in[s_i]} (b^i_j)_r u^i_{j+}.$$

This is the r-th row of $\mathcal{B}$ multiplied with $u_\mathcal{B}$. □

Lemma 5.10. *The following equations hold*:

$$\overline{p}_\mathcal{B} = \overline{p_\mathcal{B}} \quad \textit{and} \quad \overline{p}_\mathcal{C} = \overline{p_\mathcal{C}}.$$

In particular, the entries of $\overline{p}$ sum to one i.e. $\overline{p}^+_{++} = \sum_{i\in[r]}\sum_{j\in[s_i]}\sum_{k\in[t_i]}\overline{p}^i_{jk} = 1$.

Proof. By the definition of $\overline{p}$, we have

$$(\overline{p}_{\mathcal{B}})^i_j = \sum_{k\in[t_i]} \frac{(\overline{p_{\mathcal{B}}})^i_j (\overline{p_{\mathcal{C}}})^i_k}{(\overline{p_{\mathcal{A}}})^i} = \frac{(\overline{p_{\mathcal{B}}})^i_j (\overline{p_{\mathcal{C}}})^i_+}{(\overline{p_{\mathcal{A}}})^i}.$$

Hence we need to show that $(\overline{p_{\mathcal{C}}})^i_+ = (\overline{p_{\mathcal{A}}})^i$. By Birch's theorem for $I_{\mathcal{C}}$, we have $\mathcal{C}\overline{p_{\mathcal{C}}} = \mathcal{C}\frac{u_{\mathcal{C}}}{(u_{\mathcal{C}})^+_+}$ and hence $\pi_2(\mathcal{C})\overline{p_{\mathcal{C}}} = \pi_2(\mathcal{C})\frac{u_{\mathcal{C}}}{(u_{\mathcal{C}})^+_+}$ where π_2 is applied to C columnwise. The second equation is equivalent to $\sum_{i\in[r]} t_i a^i (\overline{p_{\mathcal{C}}})^i_+ = \sum_{i\in[r]} t_i a^i \frac{(u_{\mathcal{C}})^i_+}{(u_{\mathcal{C}})^+_+}$. Since a^i are linearly independent, this implies $(\overline{p_{\mathcal{C}}})^i_+ = \frac{(u_{\mathcal{C}})^i_+}{(u_{\mathcal{C}})^+_+} = \frac{u^i_{++}}{u^+_{++}} = (\overline{p_{\mathcal{A}}})^i$ for all $i \in [r]$. □

Proof of Theorem 5.5. We start by showing that every vector of the form (5-1) satisfies the conditions of Birch's theorem, i.e. $\overline{p}^+_{++} = 1$, $(\mathcal{B}\times_{\mathcal{A}}\mathcal{C})\overline{p} = (\mathcal{B}\times_{\mathcal{A}}\mathcal{C})\frac{u}{u^+_{++}}$ and $\overline{p} \in V(I_{\mathcal{B}}\times_{\mathcal{A}} I_{\mathcal{C}})$, and hence is a critical point of the likelihood function of $I_{\mathcal{B}}\times_{\mathcal{A}} I_{\mathcal{C}}$. The entries of $\overline{p}$ sum to one by Lemma 5.10. Secondly, we have

$$(\mathcal{B}\times_{\mathcal{A}}\mathcal{C})\overline{p} = (\mathcal{B}\overline{p}_{\mathcal{B}}, \mathcal{C}\overline{p}_{\mathcal{C}}) = (\mathcal{B}\overline{p_{\mathcal{B}}}, \mathcal{C}\overline{p_{\mathcal{C}}}) = (\mathcal{B}\frac{u_{\mathcal{B}}}{(u_{\mathcal{B}})^+_+}, \mathcal{C}\frac{u_{\mathcal{C}}}{(u_{\mathcal{C}})^+_+}) = (\mathcal{B}\frac{u_{\mathcal{B}}}{u^+_{++}}, \mathcal{C}\frac{u_{\mathcal{C}}}{u^+_{++}}) = (\mathcal{B}\times_{\mathcal{A}}\mathcal{C})\frac{u}{u^+_{++}}.$$

The first and last equalities hold by Lemma 5.9. The second equality holds by Lemma 5.10 while the third equality follows from Birch's theorem for $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$. Thirdly, we have to show $\overline{u} \in I_{\mathcal{B}}\times_{\mathcal{A}} I_{\mathcal{C}}$. For $f_k \in \mathrm{Lift}(F)$, we have

$$\begin{aligned} f_k(\overline{p}) &= \frac{(\overline{p_{\mathcal{B}}})^{i_1}_{j_1^1}(\overline{p_{\mathcal{C}}})^{i_1}_{k_1}}{(\overline{p_{\mathcal{A}}})^{i_1}} \frac{(\overline{p_{\mathcal{B}}})^{i_2}_{j_2^1}(\overline{p_{\mathcal{C}}})^{i_2}_{k_2}}{(\overline{p_{\mathcal{A}}})^{i_2}} \cdots \frac{(\overline{p_{\mathcal{B}}})^{i_d}_{j_d^1}(\overline{p_{\mathcal{C}}})^{i_d}_{k_d}}{(\overline{p_{\mathcal{A}}})^{i_d}} - \frac{(\overline{p_{\mathcal{B}}})^{i_1}_{j_1^2}(\overline{p_{\mathcal{C}}})^{i_1}_{k_1}}{(\overline{p_{\mathcal{A}}})^{i_1}} \frac{(\overline{p_{\mathcal{B}}})^{i_2}_{j_2^2}(\overline{p_{\mathcal{C}}})^{i_2}_{k_2}}{(\overline{p_{\mathcal{A}}})^{i_2}} \cdots \frac{(\overline{p_{\mathcal{B}}})^{i_d}_{j_d^2}(\overline{p_{\mathcal{C}}})^{i_d}_{k_d}}{(\overline{p_{\mathcal{A}}})^{i_d}} \\ &= \left((\overline{p_{\mathcal{B}}})^{i_1}_{j_1^1}(\overline{p_{\mathcal{B}}})^{i_2}_{j_2^1} \cdots (\overline{p_{\mathcal{B}}})^{i_d}_{j_d^1} - (\overline{p_{\mathcal{B}}})^{i_1}_{j_1^2}(\overline{p_{\mathcal{B}}})^{i_2}_{j_2^2} \cdots (\overline{p_{\mathcal{B}}})^{i_d}_{j_d^2} \right) \frac{(\overline{p_{\mathcal{C}}})^{i_1}_{k_1}}{(\overline{p_{\mathcal{A}}})^{i_1}} \frac{(\overline{p_{\mathcal{C}}})^{i_2}_{k_2}}{(\overline{p_{\mathcal{A}}})^{i_2}} \cdots \frac{(\overline{p_{\mathcal{C}}})^{i_d}_{k_d}}{(\overline{p_{\mathcal{A}}})^{i_d}} \\ &= f(\overline{p_{\mathcal{B}}}) \cdot \frac{(\overline{p_{\mathcal{C}}})^{i_1}_{k_1}}{(\overline{p_{\mathcal{A}}})^{i_1}} \frac{(\overline{p_{\mathcal{C}}})^{i_2}_{k_2}}{(\overline{p_{\mathcal{A}}})^{i_2}} \cdots \frac{(\overline{p_{\mathcal{C}}})^{i_d}_{k_d}}{(\overline{p_{\mathcal{A}}})^{i_d}} = 0. \end{aligned}$$

An element of Quad gives

$$(z^i_{j_1k_1}z^i_{j_2k_2} - z^i_{j_1k_2}z^i_{j_2k_1})(\overline{p}) = \frac{(\overline{p_{\mathcal{B}}})^i_{j_1}(\overline{p_{\mathcal{C}}})^i_{k_1}}{(\overline{p_{\mathcal{A}}})^i} \frac{(\overline{p_{\mathcal{B}}})^i_{j_2}(\overline{p_{\mathcal{C}}})^i_{k_2}}{(\overline{p_{\mathcal{A}}})^i} - \frac{(\overline{p_{\mathcal{B}}})^i_{j_1}(\overline{p_{\mathcal{C}}})^i_{k_2}}{(\overline{p_{\mathcal{A}}})^i} \frac{(\overline{p_{\mathcal{B}}})^i_{j_2}(\overline{p_{\mathcal{C}}})^i_{k_1}}{(\overline{p_{\mathcal{A}}})^i} = 0.$$

Hence $\overline{p}$ is a critical point of the likelihood function of $I_{\mathcal{B}}\times_{\mathcal{A}} I_{\mathcal{C}}$.

Conversely, let $\overline{p}$ be any critical point of the likelihood function of $I_{\mathcal{B}}\times_{\mathcal{A}} I_{\mathcal{C}}$. Then the entries of $\overline{p}_{\mathcal{B}}$ and $\overline{p}_{\mathcal{C}}$ sum to one. Also

$$(\mathcal{B}\frac{u_{\mathcal{B}}}{(u_{\mathcal{B}})^+_+}, \mathcal{C}\frac{u_{\mathcal{C}}}{(u_{\mathcal{C}})^+_+}) = (\mathcal{B}\frac{u_{\mathcal{B}}}{u^+_{++}}, \mathcal{C}\frac{u_{\mathcal{C}}}{u^+_{++}}) = (\mathcal{B}\times_{\mathcal{A}}\mathcal{C})\frac{u}{u^+_{++}} = (\mathcal{B}\times_{\mathcal{A}}\mathcal{C})\overline{p} = (\mathcal{B}\overline{p}_{\mathcal{B}}, \mathcal{C}\overline{p}_{\mathcal{C}}).$$

For every f in a generating set for $I_{\mathcal{B}}$

$$f(\overline{p}_{\mathcal{B}}) = \sum_{k\in T_f} c_k f_k(\overline{p}) = 0,$$

where c_k are integer coefficients. By Birch's theorem, $\overline{p}_{\mathcal{B}}$ is a critical point of the likelihood function of $I_{\mathcal{B}}$ and similarly $\overline{p}_{\mathcal{C}}$ is a critical point of the likelihood function of $I_{\mathcal{C}}$. It is left to show that $\overline{p}$ is the only element in $I_{\mathcal{B}} \times_{\mathcal{A}} I_{\mathcal{C}}$ with marginals $\overline{p}_{\mathcal{B}}$ and $\overline{p}_{\mathcal{C}}$. Indeed, for fixed $i \in [r]$, the matrix of $\overline{p}^i_{jk}$ for $j \in [s_i], k \in [t_i]$ has rank 1, because Quad contains all 2×2-minors for this matrix. Hence the marginals $\overline{p}^i_{j+}$ and $\overline{p}^i_{+k}$ completely determine this matrix.

Finally, we get the maximum likelihood estimate of $I_{\mathcal{B}} \times_{\mathcal{A}} I_{\mathcal{C}}$ by taking $\overline{p_{\mathcal{B}}}$ and $\overline{p_{\mathcal{C}}}$ to be the maximum likelihood estimates of the models $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$. By Lemma 5.10, the margins $\hat{p}_{\mathcal{B}}$ and $\hat{p}_{\mathcal{C}}$ of the maximum likelihood estimate of the model $I_{\mathcal{B}} \times_{\mathcal{A}} I_{\mathcal{C}}$ are equal to $\overline{p_{\mathcal{B}}}$ and $\overline{p_{\mathcal{C}}}$, in particular $\overline{p_{\mathcal{B}}}$ and $\overline{p_{\mathcal{C}}}$ are nonnegative. By Proposition 2.3, for each toric model there is a unique nonnegative critical point of the likelihood function and it is the maximum likelihood estimate. Hence $\overline{p_{\mathcal{B}}}$ and $\overline{p_{\mathcal{C}}}$ have to be the maximum likelihood estimates for the models $I_{\mathcal{B}}$ and $I_{\mathcal{C}}$. □

Let $\mathcal{T}$ be an n-leaf 3-valent tree. We denote the coordinates of a vector $u \in \mathbb{R}^{2^{n-1}}$ by u_l where l corresponds to a labeling of the edges of $\mathcal{T}$. Let $\mathcal{T}'$ be a subtree of $\mathcal{T}$ and denote the restriction of the labeling l to $\mathcal{T}'$ by $l|_{T'}$. We denote by $u_{T'}$ the marginal sum of u with the edges of $\mathcal{T}$ not in $\mathcal{T}'$ marginalized out, i.e. the vector $u_{T'}$ is indexed by the labelings of $\mathcal{T}'$ and if l' is a labeling of $\mathcal{T}'$ then $(u_{T'})_{l'} = \sum_{l|_{T'}=l'} u_l$. If the subtree is an edge e, then the marginal sum is defined in the same way and denoted by u_e. As before, we denote the sum of entries of u by u_+.

Corollary 5.11. *For any 3-valent tree $\mathcal{T}$, the ML degree of $V_{\mathcal{T}}$ is one. If $\mathcal{T}$ is tripod and u is a data vector, then the maximum likelihood estimate is*

$$\hat{p} = \frac{u}{u_+}.$$

It $\mathcal{T}$ has more than three leaves, let $\mathcal{T}_1, \mathcal{T}_2, \ldots, \mathcal{T}_{n-2}$ be the tripods contained in $\mathcal{T}$ and let $e_1, e_2, \ldots, e_{n-3}$ be the inner edges of $\mathcal{T}$. For data vector u, the maximum likelihood estimate is

$$\hat{p}_l = \frac{\prod_{j=1,\ldots,n-2} (\widehat{p_{\mathcal{T}_j}})_{l|_{\mathcal{T}_j}}}{\prod_{j=1,\ldots,n-3} (\widehat{p_{e_j}})_{l|_{e_j}}}, \tag{5-2}$$

where $\widehat{p_{e_j}}$ is the normalized u_{e_j}, and $\widehat{p_{\mathcal{T}_j}}$ is the maximum likelihood estimate for the tree $\mathcal{T}_j$ and the data vector $u_{\mathcal{T}_j}$.

The ML degree of a group-based phylogenetic model in probability coordinates is not known to be related to the ML degree in the Fourier coordinates. In particular, the ML degree can be much larger in the probability coordinates than the one in the Fourier coordinates [27, Section 6]. In probability coordinates, numerical methods are needed already in the smallest cases to determine the maximum likelihood estimate and the critical points of the likelihood function [32].

Example 5.12. Let $\mathcal{T}$ be the 3-valent four leaf tree, let $\mathcal{T}_1$ and $\mathcal{T}_2$ be the tripods contained in $\mathcal{T}$, and let e be the inner edge of $\mathcal{T}$. We consider the data vector

$$u = (u_{00000}, u_{11000}, u_{00011}, u_{11011}, u_{10110}, u_{10101}, u_{01110}, u_{01101}) = (17, 5, 27, 5, 16, 5, 19, 6)$$

with a total of 100 observations. Then

$$u_{\mathcal{T}_1} = (u_{000++}, u_{110++}, u_{101++}, u_{011++}) = (44, 10, 21, 25),$$
$$u_{\mathcal{T}_2} = (u_{++000}, u_{++110}, u_{++101}, u_{++011}) = (22, 35, 11, 32),$$
$$u_e = (u_{++0++}, u_{++1++}) = (54, 46).$$

Since the $V_{\mathcal{T}_1} = V_{\mathcal{T}_2} = \mathbb{P}^3$, we have

$$\widehat{p_{\mathcal{T}_1}} = (\frac{44}{100}, \frac{10}{100}, \frac{21}{100}, \frac{25}{100}), \widehat{p_{\mathcal{T}_2}} = (\frac{22}{100}, \frac{35}{100}, \frac{11}{100}, \frac{32}{100}) \text{ and } \widehat{p_e} = (\frac{54}{100}, \frac{46}{100}).$$

Then by (5-2)

$$\hat{p}_{00000} = \frac{(\widehat{p_{\mathcal{T}_1}})_{000}(\widehat{p_{\mathcal{T}_2}})_{000}}{(\widehat{p_e})_0} = \frac{44 \cdot 22 \cdot 100}{100^2 \cdot 54} = \frac{121}{675}.$$

Similarly, we can find other coordinates of the maximum likelihood estimate. This gives

$$\hat{p} = \left(\frac{121}{675}, \frac{11}{270}, \frac{176}{675}, \frac{8}{135}, \frac{147}{920}, \frac{231}{4600}, \frac{35}{184}, \frac{11}{184}\right).$$

We obtain the same result when using Birch's theorem.

Recent work on rational maximum likelihood estimators establishes that a class of tree models known as staged trees have ML degree 1 [13, Proposition 12]. In light of Corollary 5.11, it is natural to ask if there is any relation between staged tree models and 3-valent phylogenetic tree models. We find that this is the case in the proposition below. In fact, we believe that any codimension zero toric fiber product can be viewed as a generalized staged tree and this is left as a future research direction. Conversely, Ananiadi and Duarte study when staged trees are codimension-0 toric fiber products in the recent paper [4].

Consider a rooted tree $\mathcal{T}$ with at least two edges emanating from every non-leaf vertex of $\mathcal{T}$. Consider a labeling of the edges of $\mathcal{T}$ by the elements of a set S. The *floret* associated with a vertex v is the multiset of labels of edges emanating from v. The tree $\mathcal{T}$ is called a *staged tree* if any two florets are equal or disjoint. The set of florets is denoted by F.

Definition 5.13 (Definition 10 in [13]). Let J denote the set of all paths from root to leaf in $\mathcal{T}$. For a path $j \in J$ and a label $s \in S$, let μ_{sj} denote the number of times an edge labeled by s appears on the path j. A *staged tree model* is the image of the parameter space

$$\Theta = \left\{ (\theta_s)_{s\in S} \in (0,1)^S : \sum_{s\in f} \theta_s = 1 \text{ for all } f \in F \right\}$$

under the map $p_j = \prod \theta_s^{\mu_{sj}}$.

Proposition 5.14. *All 3-valent phylogenetic tree models as defined in Definition 5.1 are staged tree models.*

Proof. The staged tree begins with a tripod that can be chosen arbitrarily. The first stage has 4 edges corresponding to the four labelings of this tripod. The tripod corresponding to any subsequent stage must

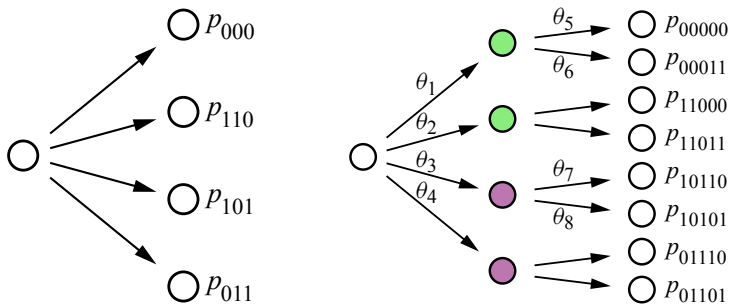


Figure 3. Staged tree for the tripod (left) and the 4-leaf tree (right).

share an edge with a tripod corresponding to a previous stage. The florets are binary and correspond to the two possible labelings of the common edge. The parameters θ_s for a given stage are marginal probabilities for the tripod divided by the marginal probabilities for the edge shared with a tripod corresponding to a previous stage. In this way, the staged tree corresponding to a phylogenetic model on a 3-valent n-leaf tree has one stage of edges for every tripod in the n-leaf tree. □

Example 5.15. The staged trees corresponding to the phylogenetic models on the tripod and the 3-valent 4-leaf tree are depicted in Figures 3 and 3. The vertices filled with the same color have the same florets. Unfilled vertices have all different florets. In Figure 3, the parameters θ_i are equal to

$$\theta_1 = p_{000++}, \quad \theta_2 = p_{110++}, \quad \theta_3 = p_{101++}, \quad \theta_4 = p_{011++},$$
$$\theta_5 = \frac{p_{++000}}{p_{++0++}}, \quad \theta_6 = \frac{p_{++011}}{p_{++0++}}, \quad \theta_7 = \frac{p_{++110}}{p_{++1++}}, \quad \theta_8 = \frac{p_{++101}}{p_{++1++}}.$$

Next, we study the ML degree drop for small phylogenetic models. One can see from Table 4 that if $\mathcal{T}$ is a 3-valent tree with at least four leaves then the degree of the phylogenetic variety is strictly larger than the sum of degrees of the components of the intersection $\mathcal{L}'_{c,u}$ of the phylogenetic variety with the affine space defined in (4-3). This implies that this intersection is not generically transverse. Since $c = (1, 1, \ldots, 1)$ in this example, we drop the subscript from L_c and $L_{c,i}$.

In the case of the 4-leaf tree, the intersection of $V_\mathcal{T}$ and the linear subspace defined by $L(p) = L_1(p) = \ldots = L_5(p) = 0$ has one component of dimension 1 and degree 1. It is the same component as in Table 4 that does not contribute to the ML degree. Since the intersection is not generically transverse, the degree 1 of this component gives only a lower bound to the difference $\deg V_\mathcal{T} - \mathrm{mldeg} V_\mathcal{T} = 3$.

In the case of the 5-leaf tree, the intersection of $V_\mathcal{T}$ and the linear subspace defined by $L(p) = L_1(p) = \ldots = L_7(p) = 0$ has three components each of dimension 3 and degrees 1, 3, 3. These components are the three components in Table 4 that do not contribute to the ML degree of $V_\mathcal{T}$. Since the intersection is not generically transverse, the sum $1 + 3 + 3 = 7$ of degrees of the components gives only a lower bound to the difference $\deg V_\mathcal{T} - \mathrm{mldeg} V_\mathcal{T} = 33$.

In both cases, the intersection is transverse only at the zero-dimensional component of degree 1 that gives the MLE.

tree	dim	deg	dim int.	deg int.	# comp.	(dim, deg) comp.	$\sum$deg comp.
tripod	3	1	0	1	1	(0,1)	1
4-leaf	5	4	1	1	2	(0,1),(1,1)	2
5-leaf	7	34	3	7	4	(0,1),(3,1),(3,3),(3,3)	8

Table 4. Properties of phylogenetic ideals on 3-valent trees. The data presented in this table are: the dimension of the variety $V_{\mathcal{T}}$; the degree of the variety $V_{\mathcal{T}}$; the dimension of the intersection of $V_{\mathcal{T}}$ and the linear subspace defined by $L(p) = L_1(p) = \ldots = L_{d-1}(p) = 0$; the degree of the intersection of $V_{\mathcal{T}}$ and the linear subspace defined by $L(p) = L_1(p) = \ldots = L_{d-1}(p) = 0$; the number of irreducible components of this intersection; the dimension and the degree of each such irreducible component in the form (dim, deg); the sum of the degrees of the irreducible components.

Huh [28] proved that if the MLE of a statistical model is a rational function of the data, then it has to be an alternating product of linear forms of specific form. In particular, the MLE is given by

$$\hat{p}_j = \lambda_j \prod_{i=1}^{m} \Big(\sum_{k=1}^{n+1} h_{ik} u_k \Big)^{h_{ij}},$$

where $\lambda = (\lambda_0, \lambda_1, \ldots, \lambda_n)$ and H is an $m \times (n+1)$ integer matrix whose columns sum to zero. Such a map is called Horn uniformization and the matrix H is called a Horn matrix. Duarte, Marigliano and Sturmfels [13, Theorem 1] proved that then there exists a determinantal triple $(A, \Delta, \mathbf{m})$ such that the statistical model is the image under the monomial map $\phi_{\Delta,\mathbf{m}}$ of an orthant of the dual of the toric variety Y_A.

Example 5.16. Let $\mathcal{T}$ be the 3-valent 4-leaf tree. Then $\lambda = (1, 1, \ldots, 1)$ and

$$H = \begin{array}{c} \\ 000{+}{+} \\ 110{+}{+} \\ 101{+}{+} \\ 011{+}{+} \\ {+}{+}{+}{+}{+} \\ {+}{+}000 \\ {+}{+}011 \\ {+}{+}110 \\ {+}{+}101 \\ {+}{+}0{+}{+} \\ {+}{+}1{+}{+} \end{array} \begin{array}{c} \begin{array}{cccccccc} p_{00000} & p_{00011} & p_{11000} & p_{11011} & p_{10110} & p_{10101} & p_{01110} & p_{01101} \end{array} \\ \begin{bmatrix} 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 1 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 \\ -1 & -1 & -1 & -1 & -1 & -1 & -1 & -1 \\ 1 & 0 & 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 1 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 0 & 1 \\ -1 & -1 & -1 & -1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & -1 & -1 & -1 & -1 \end{bmatrix} \end{array}.$$

If $\mathcal{T}$ is a 3-valent n-leaf tree, then H has $6(n-2)-1$ rows and 2^{n-1} columns. Each column contains $n-2$ ones and the same number of minus ones, all other entries are zeros. The vector λ has all its entries equal to $(-1)^{n-2}$.

The rows of the matrix

$$A = \begin{bmatrix} 1 & 1 & 1 & 1 & 1 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 0 & 1 & 0 \\ 0 & 0 & 0 & 0 & 0 & 0 & 0 & 1 & 1 & 0 & 1 \\ 2 & 2 & 1 & 1 & 1 & 0 & 0 & 1 & 1 & 1 & 1 \\ 1 & 1 & 2 & 2 & 1 & 1 & 1 & 0 & 0 & 1 & 1 \end{bmatrix}$$

give a basis of the left kernel of the Horn matrix H. The discriminant

$$\Delta_A = -x_5x_{10}x_{11} + x_1x_6x_{11} + x_1x_7x_{11} + x_2x_6x_{11} + x_2x_7x_{11} + x_3x_8x_{10} + x_3x_9x_{10} + x_4x_8x_{10} + x_4x_9x_{10}$$

vanishes on the dual of the toric variety

$$Y_A = \overline{\left\{\left(t_1t_4^2t_5 : t_1t_4^2t_5 : t_1t_4t_5^2 : t_1t_4t_5 : t_2t_5 : t_2t_5 : t_3t_4 : t_3t_4 : t_2t_4t_5 : t_3t_4t_5\right) \in \mathbb{RP}^3 : t_1, t_2, t_3, t_4, t_5 \in \mathbb{R}^*\right\}}.$$

The toric variety Y_A is of dimension 4 and degree 4. Let $\mathbf{m} = -x_5x_{10}x_{11}$. Then

$$\frac{1}{\mathbf{m}}\Delta_A = 1 - \frac{x_1x_6}{x_5x_{10}} - \frac{x_1x_7}{x_5x_{10}} - \frac{x_2x_6}{x_5x_{10}} - \frac{x_2x_7}{x_5x_{10}} - \frac{x_3x_8}{x_5x_{11}} - \frac{x_3x_9}{x_5x_{11}} - \frac{x_4x_8}{x_5x_{11}} - \frac{x_4x_9}{x_5x_{11}}.$$

This gives the monomial map

$$\phi_{(\Delta_A,\mathbf{m})} = \left(\frac{x_1x_6}{x_5x_{10}}, \frac{x_1x_7}{x_5x_{10}}, \frac{x_2x_6}{x_5x_{10}}, \frac{x_2x_7}{x_5x_{10}}, \frac{x_3x_8}{x_5x_{11}}, \frac{x_3x_9}{x_5x_{11}}, \frac{x_4x_8}{x_5x_{11}}, \frac{x_4x_9}{x_5x_{11}}\right).$$

The model $\mathcal{M}$ is the image of

$$Y_{A,\sigma}^* = \left\{x \in Y_A^* : x_1, x_2, x_3, x_4, x_6, x_7, x_8, x_9 > 0, x_5, x_{10}, x_{11} < 0\right\}$$

under the map $\phi_{(\Delta_A,\mathbf{m})}$. The maximum likelihood estimator is given by $\phi_{(\Delta_A,\mathbf{m})} \circ H : \Delta_7 \to \mathcal{M}$.

If the "3-valent" hypothesis is dropped, the ML degree does not need to be one. We conclude with an example of a toric fiber product where the ML degree is greater than one.

Example 5.17. Consider the tree $\mathcal{T}$ with five leaves that has two inner nodes of degrees three and four. Then $\mathcal{T}$ is the union of a tripod $\mathcal{T}_1$ and a four-leaf claw tree $\mathcal{T}_2$ with two edges identified. The ideal $I_\mathcal{T}$ is a toric fiber product of $I_{\mathcal{T}_1}$ and $I_{\mathcal{T}_2}$. The ML degree of $I_{\mathcal{T}_1}$ is 1 by Corollary 5.11. The ideal of $I_{\mathcal{T}_2}$ is generated by

$$x_{1001}x_{0110} - x_{0000}x_{1111}, x_{0101}x_{1010} - x_{0000}x_{1111}, x_{1100}x_{0011} - x_{0000}x_{1111}.$$

It is an ideal of codimension 3 and degree 8. Its maximum likelihood degree is 5. Hence also the ML degree of $I_\mathcal{T}$ is 5 by Theorem 5.5.

Similarly, if $\mathcal{T}$ is the six-leaf tree with two inner nodes of degree four then the ML degree of $I_\mathcal{T}$ is 25.

Acknowledgements

The authors thank Ivan Cheltsov, Alexander Davie, Martin Helmer, Milena Hering, Steffen Lauritzen, Anna Seigal, Bernd Sturmfels, Hendrik Suess and Seth Sullivant for valuable comments and suggestions. Part of this work was completed while D. Kosta was supported by a Daphne Jackson Trust Fellowship

funded jointly by EPSRC and the University of Edinburgh. C. Améndola was partially supported by the Deutsche Forschungsgemeinschaft (DFG) in the context of the Emmy Noether junior research group KR 4512/1-1. K. Kubjas was supported by the European Union's Horizon 2020 research and innovation programme: Marie Skłodowska-Curie grant agreement No. 748354, research carried out at LIDS, MIT and Team PolSys, LIP6, Sorbonne University.

References

[1] D. Agostini, D. Alberelli, F. Grande, and P. Lella, "The maximum likelihood degree of Fermat hypersurfaces.", *J. Algebr. Stat.* **6**:2 (2015).

[2] E. S. Allman, H. B. Cervantes, R. Evans, S. Hoşten, K. Kubjas, D. Lemke, J. A. Rhodes, and P. Zwiernik, "Maximum likelihood estimation of the Latent Class Model through model boundary decomposition", *J. Algebr. Stat.* **10**:1 (2019), 51–84.

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

[4] L. Ananiadi and E. Duarte, "Gröbner bases for staged trees", *arXiv preprint arXiv:1910.02721* (2019).

[5] V. V. Batyrev, "Dual polyhedra and mirror symmetry for Calabi-Yau hypersurfaces in toric varieties", *J. Algebraic Geom.* **3**:3 (1994), 493–535.

[6] Y. M. M. Bishop, S. E. Fienberg, and P. W. Holland, *Discrete multivariate analysis: theory and practice*, Springer, 2007.

[7] W. Buczyńska and J. a. A. Wiśniewski, "On geometry of binary symmetric models of phylogenetic trees", *J. Eur. Math. Soc. (JEMS)* **9**:3 (2007), 609–635.

[8] C. Casagrande, "The number of vertices of a Fano polytope", *Ann. Inst. Fourier* **56**:1 (2006), 121–130.

[9] F. Catanese, S. Hoşten, A. Khetan, and B. Sturmfels, "The maximum likelihood degree", *Amer. J. Math.* **128**:3 (2006), 671–697.

[10] D. A. Cox, J. B. Little, and H. K. Schenck, *Toric varieties*, vol. 124, Graduate Studies in Mathematics, American Mathematical Society, Providence, RI, 2011.

[11] I. V. Dolgachev, *Classical algebraic geometry*, Cambridge University Press, Cambridge, 2012. A modern view.

[12] M. Drton, B. Sturmfels, and S. Sullivant, *Lectures on algebraic statistics*, vol. 39, Oberwolfach Seminars, Birkhäuser Verlag, Basel, 2009.

[13] E. Duarte, O. Marigliano, and B. Sturmfels, "Discrete Statistical Models with Rational Maximum Likelihood Estimator", *arXiv preprint arXiv:1903.06110* (2019).

[14] A. Engström, T. Kahle, and S. Sullivant, "Multigraded commutative algebra of graph decompositions", *J. Algebraic Combin.* **39**:2 (2014), 335–372.

[15] G. Fano, "Sul sistema ∞^2 di rette contenuto in una varietà cubica generale dello spazio a quattro dimensioni", *Atti R. Acc. Sci. Torino* **39** (1904), 778–792.

[16] G. Fano, "Su alcune varietà algebriche a tre dimensioni razionali, e aventi curve-sezioni canoniche", *Comment. Math. Helv.* **14** (1942), 202–211.

[17] M. Frydenberg and S. L. Lauritzen, "Decomposition of maximum likelihood in mixed graphical interaction models", *Biometrika* **76**:3 (1989), 539–555.

[18] W. Fulton, *Intersection theory*, vol. 2, 2nd ed., Ergebnisse der Math. (3), Springer, 1998.

[19] I. M. Gelfand, M. M. Kapranov, and A. V. Zelevinsky, *Discriminants, resultants, and multidimensional determinants*, Mathematics: Theory & Applications, Birkhäuser Boston, Inc., Boston, MA, 1994.

[20] L. Godinho, F. von Heymann, and S. Sabatini, "12, 24 and beyond", *Adv. Math.* **319** (2017), 472–521.

[21] L. A. Goodman, "The Analysis of Multidimensional Contingency Tables: Stepwise Procedures and Direct Estimation Methods for Building Models for Multiple Classifications", *Technometrics* **13** (1971), 33–61.

[22] L. A. Goodman, "Causal analysis of data from panel studies and other kinds of surveys", *Amer. J. of Sociology* **78** (1973), 1135–1191.

[23] D. R. Grayson and M. E. Stillman, "Macaulay 2, a software system for research in algebraic geometry", 2002, Available at http://www.math.uiuc.edu/Macaulay2.

[24] E. Gross and J. I. Rodriguez, "Maximum likelihood geometry in the presence of data zeros", pp. 232–239 in *Proceedings of the 39th International Symposium on Symbolic and Algebraic Computation*, ACM, 2014.

[25] C. Harris and M. Helmer, "Segre class computation and practical applications", *Math. Comp.* **89**:321 (2020), 465–491.

[26] J. Hauenstein, J. I. Rodriguez, and B. Sturmfels, "Maximum likelihood for matrices with rank constraints", *J. Algebr. Stat.* **5**:1 (2014), 18–38.

[27] S. Hoşten, A. Khetan, and B. Sturmfels, "Solving the likelihood equations", *Found. Comput. Math.* **5**:4 (2005), 389–407.

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

[29] J. Huh and B. Sturmfels, "Likelihood geometry", pp. 63–117 in *Combinatorial algebraic geometry*, vol. 2108, Lecture Notes in Math., Springer, 2014.

[30] V. A. Iskovskikh and Y. G. Prokhorov, "Fano varieties", pp. 1–247 in *Algebraic geometry, V*, vol. 47, Encyclopaedia Math. Sci., Springer, 1999.

[31] J. Kollár, "Minimal models of algebraic threefolds: Mori's program", *Astérisque* 177-178 (1989), Exp. No. 712, 303–326. Séminaire Bourbaki, Vol. 1988/89.

[32] D. Kosta and K. Kubjas, "Maximum Likelihood Estimation of Symmetric Group-Based Models via Numerical Algebraic Geometry", *Bull. Math. Biol.* **81**:2 (2019), 337–360.

[33] M. Kreuzer and H. Skarke, "On the classification of reflexive polyhedra", *Comm. Math. Phys.* **185**:2 (1997), 495–508.

[34] S. L. Lauritzen, *Graphical models*, vol. 17, Oxford Statistical Science Series, The Clarendon Press, Oxford University Press, New York, 1996. Oxford Science Publications.

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

[36] S. Sullivant, *Algebraic statistics*, vol. 194, Graduate Studies in Mathematics, Amer. Math. Soc., 2018.

Received 2019-05-20. Revised 2020-01-17. Accepted 2020-03-24.

CARLOS AMÉNDOLA: carlos.amendola@tum.de
Department of Mathematics, Technical University of Munich, Garching, Germany

DIMITRA KOSTA: dimitra.kosta@glasgow.ac.uk
School of Mathematics and Statistics, University of Glasgow, United Kingdom

KAIE KUBJAS: kaie.kubjas@aalto.fi
Department of Mathematics and Systems Analysis, Aalto University, Finland

Algebraic Statistics
Vol. 11, No. 1, 2020
https://doi.org/10.2140/astat.2020.11.31

ESTIMATING LINEAR COVARIANCE MODELS WITH NUMERICAL NONLINEAR ALGEBRA

BERND STURMFELS, SASCHA TIMME AND PIOTR ZWIERNIK

Numerical nonlinear algebra is applied to maximum likelihood estimation for Gaussian models defined by linear constraints on the covariance matrix. We examine the generic case as well as special models (e.g., Toeplitz, sparse, trees) that are of interest in statistics. We study the maximum likelihood degree and its dual analogue, and we introduce a new software package `LinearCovarianceModels.jl` for solving the score equations. All local maxima can thus be computed reliably. In addition we identify several scenarios for which the estimator is a rational function.

1. Introduction

In many statistical applications, the covariance matrix Σ has a special structure. A natural setting is that one imposes linear constraints on Σ or its inverse Σ^{-1}. Here we study models for Gaussians whose covariance matrix Σ lies in a given linear space. Such linear Gaussian covariance models were introduced by 1]. He was motivated by the Toeplitz structure of Σ in time series analysis. Recent applications of such models include repeated time series, longitudinal data, and a range of engineering problems [26]. Other occurrences are Brownian motion tree models [29], as well as pairwise independence models, where some entries of Σ are set to zero.

The literature on estimating a covariance matrix is extremely rich. Its development has been particularly dynamic in high-dimensional statistics under a sparsity assumption on Σ or its inverse; see [12] for an overview. Although the sample covariance matrix is known to have poor statistical properties, for many Gaussian models the maximum likelihood estimator (MLE) remains an important reference point.

Maximum likelihood estimation for linear covariance models is a nonlinear algebraic optimization problem over a spectrahedral cone, namely the convex cone of positive definite matrices Σ that satisfy the linear constraints of interest. The objective function is not convex and can have multiple local maxima. Yet, if the sample size is large relative to the dimension, then the problem is essentially convex. This was shown in [32]. In general, however, the MLE problem is poorly understood, and there is a need for accurate methods that reliably identify all local maxima.

Nonlinear algebra furnishes such a method, namely solving the score equations [30, §7.1] using numerical homotopy continuation [27]. This is guaranteed to find all critical points of the likelihood function and hence all local maxima. A key step is the knowledge of the maximum likelihood degree

MSC2010: 14Q99, 62F10.
Keywords: linear covariance model, maximum likelihood, dual maximum likelihood, numerical nonlinear algebra.

(ML degree). This is the number of complex critical points. The ML degree of a linear covariance model is an invariant of a linear space of symmetric matrices which is of interest in its own right.

Our presentation is organized as follows. In Section 2 we introduce various models to be studied, ranging from generic linear equations to colored graph models. In Section 3 we discuss the maximum likelihood estimator as well as the dual maximum likelihood estimator. Starting from [30, Proposition 7.1.10], we derive a convenient form of the score equations. The natural point of entry for an algebraic geometer is the study of generic linear constraints. This is our topic in Section 4. We compute a range of ML degrees, and we compare them to the dual degrees in [28, §2.2].

In Section 5 we present our software `LinearCovarianceModels.jl` [31]. This is written in Julia and is easy to use. It computes the ML degree and the dual ML degree for a given subspace $\mathcal{L}$, and it determines all complex critical points for a given sample covariance matrix S. Among these, it identifies the real and positive definite solutions, and it then selects those that are local maxima. The package rests on the software `HomotopyContinuation.jl` of 2].

Section 6 discusses instances where the likelihood function has multiple local maxima. This is meant to underscore the strength of our approach. We then turn to models where the maximum is unique and the MLE is a rational function. In Section 7 we examine Brownian motion tree models. Here the linear constraints are determined by a rooted phylogenetic tree. We study the ML degree and dual ML degree. We show that the latter equals one for binary trees, and we derive the explicit rational formula for their MLE. A census of these degrees is found in Table 5.

2. Models

Let $\mathbb{S}^n$ be the $\binom{n+1}{2}$-dimensional real vector space of $n \times n$ symmetric matrices $\Sigma = (\sigma_{ij})$. The subset $\mathbb{S}^n_+$ of positive definite matrices is a full-dimensional open convex cone. Consider any linear subspace $\mathcal{L}$ of $\mathbb{S}^n$ whose intersection with $\mathbb{S}^n_+$ is nonempty. Then $\mathbb{S}^n_+ \cap \mathcal{L}$ is a relatively open convex cone. In optimization, where one uses the closure, this is known as a *spectrahedral cone*. In statistics, the intersection $\mathbb{S}^n_+ \cap \mathcal{L}$ is a *linear covariance model*. These are the models we study in this paper. In what follows we discuss various families of linear spaces $\mathcal{L}$ that are of interest to us.

Generic linear constraints. Fix a positive integer $m \leq \binom{n+1}{2}$, and suppose that $\mathcal{L}$ is a generic linear subspace of $\mathbb{S}^n$. Here "generic" is meant in the sense of algebraic geometry; i.e., $\mathcal{L}$ is a point in the Grassmannian that lies outside a certain algebraic hypersurface. This hypersurface has measure zero, so a random subspace will be generic with probability one. For a geometer, it is natural to begin with the generic case, since its complexity controls the complexity of any special family of linear spaces. In particular, the ML degree for a generic $\mathcal{L}$ depends only on m and n, and this furnishes an upper bound for the ML degree of the special families below.

Diagonal covariance matrices. Here we take $m \leq n$, and we assume that $\mathcal{L}$ is a linear space that consists of diagonal matrices. Restricting to covariance matrices that are diagonal is natural when modeling independent Gaussians. We use the term *generic diagonal model* when $\mathcal{L}$ is a generic point in the $(n-m)m$-dimensional Grassmannian of m-dimensional subspaces inside the diagonal $n \times n$ matrices.

Brownian motion tree models. A tree is a connected graph with no cycles. A rooted tree is obtained by fixing a vertex, called the root, and directing all edges away from the root. Fix a rooted tree T with n leaves. Every vertex v of T defines a *clade*, namely the set of leaves that are descendants of v. For the Brownian motion tree model on T, the space $\mathcal{L}$ is spanned by the rank-one matrices $e_A e_A^T$, where $e_A \in \{0,1\}^n$ is the indicator vector of A. Hence, if $\mathcal{C}$ is the set of all clades of T, then

$$\Sigma = \sum_{A \in \mathcal{C}} \theta_A e_A e_A^T, \quad \text{where } \theta_A \text{ are model parameters.} \tag{1}$$

The linear equations for the subspace $\mathcal{L}$ are $\sigma_{ij} = \sigma_{kl}$ whenever the least common ancestors $\mathrm{lca}(i,j)$ and $\mathrm{lca}(k,l)$ agree in the tree T. Assuming $\theta_A \geq 0$, the union of the models for all trees T is characterized by the ultrametric condition $\sigma_{ij} \geq \min\{\sigma_{ik}, \sigma_{jk}\} \geq 0$. Matrices of this form also play an important role in hierarchical clustering [15, §14.3.12], phylogenetics [13], and random walks on graphs [9].

Maximum likelihood estimation for this class of models is generally complicated, but recently there has been progress [16; 29] on exploiting the nice structure of the matrices Σ above. In Section 7 we study computational aspects of the MLE and, more importantly, provide a significant advance by considering the dual MLE.

Covariance graph models. We consider models $\mathcal{L}$ that arise from imposing zero restrictions on entries of Σ. This was studied in [5; 10]. This is similar to Gaussian graphical models where zero restrictions are placed on the inverse Σ^{-1}. We encode the sparsity structure with a graph whose edges correspond to nonzero off-diagonal entries of Σ. Zero entries in Σ correspond to pairwise marginal independences. These arise in statistical modeling in the context of causal inference [8]. Models with zero restrictions on the covariance matrix are known as covariance graph models. Maximum likelihood in these Gaussian models can be carried out using iterative conditional fitting [5; 10], which is implemented in the `ggm` package in R [22].

Toeplitz matrices. Suppose $X = (X_1, \ldots, X_n)$ follows the autoregressive model of order 1, that is, $X_t = \rho X_{t-1} + \epsilon_t$, where $\rho \in \mathbb{R}$ and $\epsilon_t \sim N(0, \sigma)$ for some σ. Assume that the ϵ_t are mutually uncorrelated. Then $\mathrm{cov}(X_t, X_{t-k}) = \rho^k$, and hence Σ is a Toeplitz matrix. More generally, covariance matrices from stationary time series are Toeplitz. Multichannel and multidimensional processes have covariance matrices of block Toeplitz form [3; 24]. Similarly, if X follows the moving average process of order q, then $\mathrm{cov}(X_t, X_{t-k}) = \gamma_k$ if $k \leq q$ and is zero otherwise; see, for example, [14, §3.3]. Thus, in time series analysis, we encounter matrices like

$$\begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 & \gamma_4 \\ \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 & \gamma_3 \\ \gamma_2 & \gamma_1 & \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_4 & \gamma_3 & \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix} \quad \text{or} \quad \begin{bmatrix} \gamma_0 & \gamma_1 & 0 & 0 & 0 \\ \gamma_1 & \gamma_0 & \gamma_1 & 0 & 0 \\ 0 & \gamma_1 & \gamma_0 & \gamma_1 & 0 \\ 0 & 0 & \gamma_1 & \gamma_0 & \gamma_1 \\ 0 & 0 & 0 & \gamma_1 & \gamma_0 \end{bmatrix}. \tag{2}$$

We found that the ML degree for such models is surprisingly low. This means that nonlinear algebra can reliably estimate Toeplitz matrices that are fairly large.

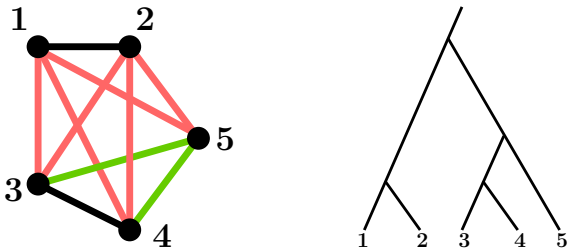


Figure 1. A covariance graph model with edge symmetries and the rooted tree for the corresponding Brownian motion tree model.

Colored covariance graph models. A generalization of covariance graph models is obtained by following [17], which introduces graphical models with vertex and edge symmetries. Models of this type also generalize the Toeplitz matrices and the Brownian motion tree models. Following the standard convention, we use the same colors for edges or vertices when the corresponding entries of Σ are equal. The black color is considered neutral and encodes no restrictions.

The Brownian motion tree model corresponds to a colored model over the complete graph, where edge symmetries are encoded by the tree; see Figure 1. Also, both matrices in (2) represent covariance graph models with edge and vertex symmetries.

3. Maximum likelihood estimator and its dual

Now that we have seen motivating examples, we formally define the MLE problem for a linear covariance model $\mathcal{L}$. Suppose we observe a random sample $X^{(1)}, \ldots, X^{(N)}$ in $\mathbb{R}^n$ from $N_n(0, \Sigma)$. The sample covariance matrix is $S = (1/N) \sum_{i=1}^{N} X^{(i)} X^{(i)T}$. The matrix S is positive semidefinite. Our aim is to maximize the function

$$\ell(\Sigma) = \log\det \Sigma^{-1} - \operatorname{tr}(S\Sigma^{-1}) \quad \text{subject to } \Sigma \in \mathcal{L}. \tag{3}$$

Following [30, Proposition 7.1.10], this equals the log-likelihood function times $N/2$.

We fix the standard inner product $\langle A, B\rangle = \operatorname{tr}(AB)$ on the space $\mathbb{S}^n$ of symmetric matrices. The orthogonal complement $\mathcal{L}^\perp$ to a subspace $\mathcal{L} \subset \mathbb{S}^n$ is defined as usual.

Proposition 3.1. *Finding all the critical points of the log-likelihood function $\ell(\Sigma)$ amounts to solving the following system of linear and quadratic equations in $2 \cdot \binom{n+1}{2}$ unknowns*:

$$\Sigma \in \mathcal{L}, \qquad K\Sigma = I_n, \qquad KSK - K \in \mathcal{L}^\perp. \tag{4}$$

Proof. The matrix Σ is a critical point ℓ if and only if, for every $U \in \mathcal{L}$, the derivative of ℓ at Σ in the direction U vanishes. This directional derivative equals

$$-\operatorname{tr}(\Sigma^{-1}U) + \operatorname{tr}(S\Sigma^{-1}U\Sigma^{-1}).$$

This formula follows by multivariate calculus from two facts: (i) the derivative of the matrix mapping $\Sigma \mapsto \Sigma^{-1}$ is the linear transformation $U \mapsto \Sigma^{-1} U \Sigma^{-1}$; (ii) the derivative of the function $\Sigma \mapsto \log\det\Sigma$ is the linear functional $U \mapsto \mathrm{tr}(\Sigma^{-1}U)$.

Using the identity $K = \Sigma^{-1}$, vanishing of the directional derivative is equivalent to

$$-\langle K, U\rangle + \langle KSK, U\rangle = 0.$$

The condition $\langle KSK - K, U\rangle = 0$ for all $U \in \mathscr{L}$ is equivalent to $KSK - K \in \mathscr{L}^{\perp}$. □

Example 3.2 (3×3 Toeplitz matrices). Let $\mathscr{L}$ be the space of Toeplitz matrices

$$\Sigma = \begin{bmatrix} \gamma_0 & \gamma_1 & \gamma_2 \\ \gamma_1 & \gamma_0 & \gamma_1 \\ \gamma_2 & \gamma_1 & \gamma_0 \end{bmatrix}.$$

This space has dimension 3 in $\mathbb{S}^3 \simeq \mathbb{R}^6$. Fix a sample covariance matrix $S = (s_{ij})$ with real entries. We need to solve the system (4). This consists of $3+9+3 = 15$ equations in $6+6 = 12$ unknowns, namely the entries of the covariance matrix $\Sigma = (\sigma_{ij})$ and its inverse $K = (k_{ij})$. The condition $\Sigma \in \mathscr{L}$ gives three linear polynomials

$$\sigma_{11} - \sigma_{33}, \qquad \sigma_{12} - \sigma_{23}, \qquad \sigma_{22} - \sigma_{33}.$$

The condition $K\Sigma = I_3$ translates into nine bilinear polynomials

$$\begin{aligned}
&\sigma_{11}k_{11} + \sigma_{12}k_{12} + \sigma_{13}k_{13} - 1, && \sigma_{12}k_{11} + \sigma_{22}k_{12} + \sigma_{23}k_{13}, && \sigma_{13}k_{11} + \sigma_{23}k_{12} + \sigma_{33}k_{13}, \\
&\sigma_{11}k_{12} + \sigma_{12}k_{22} + \sigma_{13}k_{23}, && \sigma_{12}k_{12} + \sigma_{22}k_{22} + \sigma_{23}k_{23} - 1, && \sigma_{13}k_{12} + \sigma_{23}k_{22} + \sigma_{33}k_{23}, \\
&\sigma_{11}k_{13} + \sigma_{12}k_{23} + \sigma_{13}k_{33}, && \sigma_{12}k_{13} + \sigma_{22}k_{23} + \sigma_{23}k_{33}, && \sigma_{13}k_{13} + \sigma_{23}k_{23} + \sigma_{33}k_{33} - 1.
\end{aligned}$$

Finally, the condition $KSK - K \in \mathscr{L}^{\perp}$ translates into three quadratic polynomials

$$\begin{aligned}
&k_{11}^2 s_{11} + k_{12}^2 s_{11} + k_{13}^2 s_{11} + 2k_{11}k_{12}s_{12} + 2k_{12}k_{22}s_{12} + 2k_{13}k_{23}s_{12} + 2k_{11}k_{13}s_{13} \\
&\qquad + 2k_{12}k_{23}s_{13} + 2k_{13}k_{33}s_{13} + k_{12}^2 s_{22} + k_{22}^2 s_{22} + k_{23}^2 s_{22} + 2k_{12}k_{13}s_{23} + 2k_{22}k_{23}s_{23} \\
&\qquad\qquad + 2k_{23}k_{33}s_{23} + k_{13}^2 s_{33} + k_{23}^2 s_{33} + k_{33}^2 s_{33} - k_{11} - k_{22} - k_{33}, \\
&k_{23}s_{13} + k_{12}k_{33}s_{13} + k_{12}k_{22}s_{22} + k_{22}k_{23}s_{22} + k_{13}k_{22}s_{23} + k_{12}k_{23}s_{23} \\
&\qquad + k_{23}^2 s_{23} + k_{22}k_{33}s_{23} + k_{13}k_{23}s_{33} + k_{23}k_{33}s_{33} - k_{12} - k_{23}, \\
&k_{11}k_{13}s_{11} + k_{12}k_{13}s_{12} + k_{11}k_{23}s_{12} + k_{13}^2 s_{13} + k_{11}k_{33}s_{13} + k_{12}k_{23}s_{22} + k_{13}k_{23}s_{23} + k_{12}k_{33}s_{23} + k_{13}k_{33}s_{33} - k_{13}.
\end{aligned}$$

The zero set of these 15 polynomials in 12 unknowns consists of three points $(\widehat{\Sigma}, \widehat{K})$. We present a concrete instance with multiple local solutions:

$$S = \begin{bmatrix} 4/5 & -9/5 & -1/25 \\ -9/5 & 79/16 & 25/24 \\ -1/25 & 25/24 & 17/16 \end{bmatrix} \approx \begin{bmatrix} 0.8000 & -1.8000 & -0.0400 \\ -1.8000 & 4.9375 & 1.0417 \\ -0.0400 & 1.0417 & 1.0625 \end{bmatrix}. \tag{5}$$

For this sample covariance matrix all three critical points are real and positive definite. The three Toeplitz matrices that solve the score equations for this S are:

$\hat{\gamma}_0$	$\hat{\gamma}_1$	$\hat{\gamma}_2$	log-likelihood value	
2.52783	−0.215929	−1.45229	−5.35	global maximum
2.39038	−0.286009	0.949965	−5.41	local maximum
2.28596	−0.256394	0.422321	−5.42	saddle point

So, even in this tiny example, our optimization problem has multiple local maxima in the cone $\mathbb{S}^3_+$. A numerical study of this phenomenon will be presented in Section 6.

In this paper we also consider the *dual* maximum likelihood estimator as a more computationally efficient alternative. Dual estimation is based on the maximization of a dual likelihood function. In the Gaussian case this is motivated by interchanging the role of the parameter matrix Σ and the empirical covariance matrix S. The *Kullback–Leibler divergence* of two Gaussian distributions $N(0, \Sigma_0)$ and $N(0, \Sigma_1)$ on $\mathbb{R}^n$ is equal to

$$\mathrm{KL}(\Sigma_0, \Sigma_1) = \frac{1}{2}\left(\mathrm{tr}(\Sigma_1^{-1}\Sigma_0) - n + \log\frac{\det\Sigma_1}{\det\Sigma_0}\right).$$

Computing the MLE is equivalent to minimizing $\mathrm{KL}(\Sigma_0, \Sigma_1)$ with respect to Σ_1 with $\Sigma_0 = S$. On the other hand, the dual MLE is obtained by minimizing $\mathrm{KL}(\Sigma_0, \Sigma_1)$ with respect to Σ_0 with $\Sigma_1 = S$. Equivalently, we set $W = S^{-1}$ and maximize

$$\ell^\vee(\Sigma) = \log\det\Sigma - \mathrm{tr}(W\Sigma).$$

The idea of utilizing the "wrong" Kullback–Leibler distance is ubiquitous in variational inference and is central for mean field approximation and related methods. The idea of using this estimation method for Gaussian linear covariance models is very natural. It results in a unique maximum, since $\Sigma \mapsto \ell^\vee(\Sigma)$ is a convex function on the positive definite cone $\mathbb{S}^n_+$. See [6] and also [5, §3.2; 18, §4].

The following algebraic formulation is the analogue to Proposition 3.1.

Proposition 3.3. *Finding all the critical points of the dual log-likelihood function $\ell^\vee$ amounts to solving the following system of equations in $2\cdot\binom{n+1}{2}$ unknowns*:

$$\Sigma \in \mathscr{L}, \qquad K\Sigma = I_n, \qquad K - W \in \mathscr{L}^\perp. \tag{6}$$

Proof. After switching the roles of K and Σ, and of W and S, our problem becomes MLE for linear concentration models. Equation (6) is found in [28, (10)]. □

The next result lists properties of the dual MLE that are important for statistics.

Proposition 3.4. *The dual maximum likelihood estimator of a Gaussian linear covariance model is consistent, asymptotically normal, and first-order efficient.*

Proof. See Theorems 3.1 and 3.2 in [6]. □

First-order efficiency means that the asymptotic variance of the properly normalized dual MLE is optimal, that is, it equals the asymptotic variance of the MLE.

In this paper, we focus on algebraic structures, and we note the following important distinction between our two estimators. The MLE requires the quadratic equations $KSK - K \in \mathscr{L}^\perp$ in (4), whereas the dual MLE requires the linear equations $K - W \in \mathscr{L}^\perp$ in (6). The latter are easier to solve than the former, and they give far fewer solutions. This is quantified by the tables for the ML degrees in the next sections.

We are particularly interested in models whose dual ML estimator $(\check{\Sigma}, \check{K})$ can be written as an explicit expression in the sample covariance matrix S. We identify such scenarios in Sections 6 and 7. Here is a first example to illustrate this point.

Example 3.5. We revisit the Toeplitz model in Example 3.2. For the dual MLE, the three quadratic polynomials in K are now replaced by three linear polynomials

$$k_{11}+k_{22}+k_{33}-w_{11}-w_{22}-w_{33}, \qquad k_{12}+k_{23}-w_{12}-w_{23}, \qquad k_{13}-w_{13}.$$

The w_{ij} are the entries of the inverse sample covariance matrix $W = S^{-1}$. The new system has two solutions, and we can write the $\check{\sigma}_{ij}$ and $\check{k}_{ij}$ in terms of the w_{ij} (or the s_{ij}) using the familiar formula for solving quadratic equations in one variable. Specifically, for the covariance matrix S in (5) we find that the dual MLE is given by

$$\begin{aligned}[\check{\gamma}_0, \check{\gamma}_1, \check{\gamma}_2] &= [0.203557267562, -0.189349961613, 0.1963649733282] \\ &= \Big[\tfrac{1284368265268038839512}{12363704694314904961417} + \tfrac{52\sqrt{561647777654592987689702150027364667081}}{12363704694314904961417}, \\ &\quad -\tfrac{5817390611804320873051}{6181852347157452480 7085} - \tfrac{655679934637\sqrt{561647777654592987689702150027364667081}}{16314690552471559970524472 9886305}, \\ &\quad \tfrac{1990451408446510673691859}{22254668449766828930550600} + \tfrac{264990063915733\sqrt{561647777654592987689702150027364667081}}{58732885988897615893888102759069800}\Big].\end{aligned}$$

Needless to say, nonlinear algebra goes much beyond the quadratic formula. In what follows we shall employ state-of-the-art methods for solving polynomial equations.

4. General linear constraints

The *maximum likelihood degree* of a linear covariance model $\mathscr{L}$ is, by definition, the number of complex solutions to the likelihood equations (4) for generic data S. This is abbreviated *ML degree* [30, §7.1]. To compute the ML degree, take S to be a random symmetric $n \times n$ matrix and count all complex critical points of the likelihood function $\ell(\Sigma)$ for $\Sigma \in \mathscr{L}$. Equivalently, the ML degree of the model $\mathscr{L}$ is the number of complex solutions (Σ, K) to the polynomial equations in (4).

We also consider the complex critical points of the dual likelihood function $\ell^\vee(\Sigma)$. Their number, for a generic matrix $S \in \mathbb{S}^n$, is the *dual ML degree* of $\mathscr{L}$. It coincides with the number of complex solutions (Σ, K) to the polynomial equations in (6).

Our ML degrees can be computed symbolically in a computer algebra system that rests on Gröbner bases. However, this approach is limited to small instances. To get further, we use the methods from numerical nonlinear algebra described in Section 5.

m \ n	2	3	4	5	6
2	1	3	5	7	9
3	1	7	19	37	61
4		7	45	135	299
5		3	71	361	1121
6		1	81	753	3395
7			63	1245	8513
8			29	1625	17867
9			7	1661	31601
10			1	1323	47343
11				801	60177
12				347	64731
13				97	58561
14				15	44131
15				1	27329
16					13627
17					5341
18					1511
19					289
20					31
21					1

m \ n	2	3	4	5	6
2	1	2	3	4	5
3	1	4	9	16	25
4		4	17	44	90
5		2	21	86	240
6		1	21	137	528
7			17	188	1016
8			9	212	1696
9			3	188	2396
10			1	137	2886
11				86	3054
12				44	2886
13				16	2396
14				4	1696
15				1	1016
16					528
17					240
18					90
19					25
20					5
21					1

Table 1. ML degrees and dual ML degrees for generic models.

Here we focus on a generic m-dimensional linear subspace $\mathscr{L}$ of $\mathbb{S}^n$. In practice this means that a basis for $\mathscr{L}$ is chosen by sampling m matrices at random from $\mathbb{S}^n$.

Proposition 4.1. *The ML degree and the dual ML degree of a generic subspace $\mathscr{L}$ of dimension m in $\mathbb{S}^n$ depends only on m and n. It is independent of the particular choice of $\mathscr{L}$. For small parameter values, these ML degrees are listed in Table 1.*

Proof. The independence rests on general results in algebraic geometry [27, Corollary A.14.2], to the effect that the system (4) (resp. (6)) can be considered as a system parametrized by the coordinates of $\mathscr{L}$ and S (resp. W). The ML degree will be the same for all specializations to $\mathbb{R}$ that remain outside a certain discriminant hypersurface. Table 1 and further values are computed rapidly using the software described in Section 5. □

The dual ML degree was already studied by 28, §2]. Our table on the right is in fact found in their paper. The symmetry along its columns is proved in [28, Theorem 2.3]. It states that the dual ML degree for dimension m coincides with the dual ML degree for codimension $m-1$. This is derived from the equations (6) by an appropriate homogenization. Namely, the middle equation is clearly symmetric under switching the roles of K and Σ, and the linear equations on the left and on the right in (6) can also be interchanged under this switch.

It was conjectured in [28, §2] that, for fixed m, the dual ML degree is a polynomial of degree $m-1$ in the matrix size n. This is easy to see for $m \leq 3$. The polynomials for $m = 4$ and $m = 5$ were also derived in [28, §2].

The situation is similar but more complicated for the ML degree. First of all, the symmetry along columns no longer holds as seen on the left in Table 1. This is explained by the fact that the linear equation $K - W \in \mathscr{L}^{\perp}$ is now replaced by the quadratic equation $KSK - K \in \mathscr{L}^{\perp}$. However, the polynomiality along the rows of Table 1 seems to persist. For $m = 2$ the ML degree equals $2n-3$, as shown recently by Coons, Marigliano, and Ruddy [7]. For $m \geq 3$ we propose the following conjecture.

Conjecture 4.2. The ML degree of a linear covariance model of dimension m is a polynomial of degree $m-1$ in the ambient dimension n. For $m = 3$ this ML degree equals $3n^2 - 9n + 7$, and for $m = 4$ it equals $\frac{11}{3}n^3 - 18n^2 + \frac{85}{3}n - 15$.

We now come to *diagonal linear covariance models*. For these models, $\mathscr{L}$ is a linear subspace of dimension m inside the space $\mathbb{R}^n$ of diagonal $n \times n$ matrices. We wish to determine the ML degree and dual ML degree when $\mathscr{L}$ is generic in $\mathbb{R}^n$.

In the diagonal case, the score equations simplify as follows. Both the covariance matrix and the concentration matrix are diagonal. We eliminate the entries of Σ by setting $K = \operatorname{diag}(k_1, \ldots, k_n)$ and $\Sigma = \operatorname{diag}(k_1^{-1}, \ldots, k_n^{-1})$. We also write $s_1, \ldots, s_n$ for the diagonal entries of the sample covariance matrix S and $w_i = s_i^{-1}$ for their reciprocals. Finally, let $\mathscr{L}^{-1}$ denote the *reciprocal linear space* of $\mathscr{L}$, i.e., the variety obtained as the closure of the set of coordinatewise reciprocals of vectors in $\mathscr{L} \cap (\mathbb{R}^*)^n$.

Proposition 4.3. *Let $\mathscr{L} \subset \mathbb{R}^n$ be a linear space, viewed as a Gaussian covariance model of diagonal matrices. The score equations for the likelihood in (4) and the dual likelihood in (6) can be written as systems of n equations in n unknowns*:

$$(k_1, \ldots, k_n) \in \mathscr{L}^{-1} \quad \text{and} \quad (s_1k_1^2 - k_1, s_2k_2^2 - k_2, \ldots, s_nk_n^2 - k_n) \in \mathscr{L}^{\perp}, \tag{4'}$$

$$(k_1, \ldots, k_n) \in \mathscr{L}^{-1} \quad \text{and} \quad (k_1 - w_1, k_2 - w_2, \ldots, k_n - w_n) \in \mathscr{L}^{\perp}. \tag{6'}$$

The number of complex solutions to (6′) for generic $\mathscr{L}$ of dimension m equals $\binom{n-1}{m-1}$.

Proof. The translation of (4) and (6) to (4′) and (6′) is straightforward. The equations (6′) represent a general linear section of the reciprocal linear space $\mathscr{L}^{-1}$. Proudfoot and Speyer showed that the degree of $\mathscr{L}^{-1}$ equals the Möbius invariant of the underlying matroid. We refer to [19] for a recent study. This Möbius invariant equals $\binom{n-1}{m-1}$ in the generic case, when the matroid is uniform. □

It would be desirable to express the number of complex solutions to (4′) as a matroid invariant, and thereby explain the entries on the left side of Table 2. As before, the m-th row gives the values of a polynomial of degree $m-1$. For instance, for $m = 3$ we find $2n^2 - 8n + 7$, and for $m = 4$ we find $\frac{4}{3}n^3 - 10n^2 + \frac{68}{3}n - 15$.

m	n				
	3	4	5	6	7
2	3	5	7	9	11
3	1	7	17	31	49
4		1	15	49	111
5			1	31	129
6				1	63
7					1

m	n				
	3	4	5	6	7
2	2	3	4	5	6
3	1	3	6	10	15
4		1	4	10	21
5			1	5	21
6				1	15
7					1

Table 2. ML degrees and dual ML degrees for generic diagonal models.

5. Numerical nonlinear algebra

Linear algebra is the foundation of scientific computing and applied mathematics. *Nonlinear algebra* [23] is a generalization where linear systems are replaced by nonlinear equations and inequalities. At the heart of this lies algebraic geometry, but there are links to many other branches, such as combinatorics, algebraic topology, commutative algebra, convex and discrete geometry, tensors and multilinear algebra, number theory, and representation theory. Nonlinear algebra is not simply a rebranding of algebraic geometry. It highlights that the focus is on computation and applications, and the theoretical needs that this requires results in a new perspective.

We refer to *numerical nonlinear algebra* as the branch of nonlinear algebra which is concerned with the efficient numerical solution of polynomial equations and inequalities. In the existing literature, this is referred to as numerical algebraic geometry. In the following we discuss the numerical solution of polynomial equations, and we describe the techniques used for deriving the computational results in this paper.

One of our main contributions is the Julia package `LinearCovarianceModels.jl` for estimating linear covariance models [31]. Given $\mathscr{L}$, our package computes the ML degree and the dual ML degree. For any S, it finds all critical points and selects those that are local maxima. The following example explains how this is done.

Example 5.1. We use the package to verify Example 3.2:

```
julia> using LinearCovarianceModels
julia> Σ = toeplitz(3)
3-dimensional LCModel:
  θ₁ θ₂ θ₃
  θ₂ θ₁ θ₂
  θ₃ θ₂ θ₁
```

We compute the ML degree of the family Σ by computing all solutions for a generic instance. The pair of solutions and generic instance is called an *ML degree witness*:

```
julia> W = ml_degree_witness(Σ)
MLDegreeWitness:
```

```
  ∘ ML degree → 3
  ∘ model dimension → 3
  ∘ dual → false
```

By default, the computation of the ML degree witness relies on a heuristic stopping criterion. We can numerically verify the correctness by using a trace test [21]:

```
julia> verify(W)
Compute additional witnesses for completeness...
Found 10 additional witnesses
Compute trace...
Norm of trace:  2.6521474798326718e-12
true
```

We now input the specific sample covariance matrix in (5), and we compute all critical points of this MLE problem using the ML degree witness from the previous step:

```
julia> S = [4/5 -9/5 -1/25; -9/5 79/16 25/24; -1/25 25/24 17/16];
julia> critical_points(W, S)
3-element Array{Tuple{Array{Float64,1},Float64,Symbol},1}:
  ([2.39038, -0.286009, 0.949965], -5.421751313919751, :local_maximum)
  ([2.52783, -0.215929, -1.45229], -5.346601549034418, :global_maximum)
  ([2.28596, -0.256394, 0.422321], -5.424161999175718, :saddle_point)
```

If only the global maximum is of interest then this can also be computed directly:

```
julia> mle(W, S)
3-element Array{Float64,1}:
  2.527832268219689
  -0.21592947057775033
  -1.4522862659134732
```

By default only positive definite solutions are reported. To list all critical points we run the command with an additional option:

```
julia> critical_points(W, S, only_positive_definite=false)
```

In this case, since the ML degree is 3, we are not getting more solutions.

In the rest of this section we explain the mathematics behind our software, and how it applies to our MLE problems. A textbook introduction to the numerical solution of polynomial systems by homotopy continuation methods is [27].

Suppose we are given m polynomials $f_1, \ldots, f_m$ in n unknowns $x_1, \ldots, x_n$ with complex coefficients, where $m \geq n$. We are interested in computing all isolated complex solutions of the system $f_1(x) = \cdots = f_m(x) = 0$. These solutions comprise the zero-dimensional components of the variety $V(F)$ where $F = (f_1, \ldots, f_m)$.

The general idea of homotopy continuation is as follows. Assume we have another system $G = (g_1, \dots, g_m)$ of polynomials for which we know some or all of its solutions. Suppose there is a *homotopy* $H(x, t)$ with $H(x, 0) = G(x)$ and $H(x, 1) = F(x)$ with the property that, for every $x^* \in V(G)$, there exists a smooth path $x : [0, 1) \to \mathbb{C}^n$ with $x(0) = x^*$ and $H(x(t), t) = 0$ for all $t \in [0, 1)$. Then we can track each point in $V(G)$ to a point in $V(F)$. This is done by solving the *Davidenko differential equation*

$$\frac{\partial H}{\partial x}(x(t), t) \cdot \dot{x}(t) + \frac{\partial H}{\partial t}(x(t), t) = 0$$

with initial condition $x(0) = x^*$. Using a *predictor-corrector* scheme for numerical path tracking, both the local and global error can be controlled. Methods called *endgames* are used to handle divergent paths and singular solutions [27, Chapter 10].

Here is a general framework for start systems and homotopies. Embed F in a family of polynomial systems $\mathcal{F}_Q$, continuously parametrized by a convex open set $Q \subset \mathbb{C}^k$. We have $F = F_q \in \mathcal{F}_Q$ for some $q \in Q$. Outside a Zariski closed set $\Delta \subset Q$, every system in $\mathcal{F}_Q$ has the same number of solutions. If $p \in Q \setminus \Delta$, then F_p is such a *generic instance* of the family $\mathcal{F}_Q$, and the following is a suitable homotopy [25]:

$$H(x, t) = F_{(1-t)p+tq}(x). \tag{7}$$

Now, to compute $V(F_q)$, it suffices to find all solutions of a generic instance F_p and then track these along the homotopy (7). Obtaining all solutions of a generic instance can be a challenge, but this has to be done *only once*! That is the *offline phase*. Tracking from a generic to a specific instance of interest is the *online phase*.

A key point in applying this method is the choice of the family $\mathcal{F}_Q$. For MLE problems in statistics, it is natural to choose Q as the space of data or instances. In our scenario, Q is $\mathbb{S}^n$, or a complex version thereof. We shall discuss this below.

First, we explain the *monodromy method* for an arbitrary family $\mathcal{F}_Q$. Suppose the general instance has d solutions, and that we are given one *start pair* (x_0, p_0). This means that x_0 is a solution to the instance F_{p_0}. Consider the incidence variety

$$Y := \{(x, p) \in \mathbb{C}^n \times Q \mid F_p(x) = 0\}.$$

Let π be the projection from $\mathbb{C}^n \times Q$ onto the second factor. For $q \in Q \setminus \Delta$, the fiber $\pi^{-1}(q)$ has exactly d points. A loop in $Q \setminus \Delta$ based at q has d lifts to Y. Associating a point in the fiber to the endpoint of the corresponding lift gives a permutation in S_d. This defines an action of the fundamental group of $Q \setminus \Delta$ on the fiber $\pi^{-1}(q)$. The *monodromy group* of our family is the image of the fundamental group in S_d.

The monodromy method fills the fiber $\pi^{-1}(p_0)$ by exploiting the monodromy group. For this, the start solution x_0 is numerically tracked along a loop in $Q \setminus \Delta$, yielding a solution x_1 at the end. If $x_1 \neq x_0$, then x_1 is also tracked along the *same* loop, possibly again yielding a new solution. This is done until no more solutions are found. Then, all solutions are tracked along a new loop, where the process is repeated. This process is stopped by use of a *trace test*. For a detailed description of the monodromy method and the trace test, see [4; 21]. To get this off the ground, one needs a start pair (x_0, p_0). This can often be

found by inverting the problem. Instead of finding a solution x_0 to a given p_0, we start with x_0 and look for p_0 such that $F_{p_0}(x_0) = 0$.

We now explain how this works for the score equations (4) of our MLE problem. First pick a random matrix Σ_0 in the subspace $\mathscr{L}$. We next compute K_0 by inverting Σ_0. Finally we need to find a symmetric matrix S_0 such that $K_0 S_0 K_0 - K_0 \in \mathscr{L}^\perp$. Note that this is a linear system of equations and hence directly solvable. In this manner, we easily find a start pair (x_0, p_0) by setting $p_0 = S_0$ and $x_0 = (\Sigma_0, K_0)$.

The number d of solutions to a generic instance is the ML degree of our model. A priori knowledge of d is useful because it serves as a stopping criterion in the monodromy method. This is one reason for focusing on the ML degree in this paper.

6. Local maxima versus rational MLE

The theme of this paper is maximum likelihood inference for linear covariance models. We developed some numerical nonlinear algebra for this problem, and we offer a software package [31]. From the applications perspective, this is motivated by the fact that the likelihood function is nonconvex. It can have multiple local maxima. A concrete instance for 3×3 Toeplitz matrices was shown in Example 3.2.

In this section we undertake a more systematic experimental study of local maxima. Our aim is to answer the following question: there is the theoretical possibility that $\ell(\Sigma)$ has many local maxima, but can we also observe this in practice?

To address this question, we explored a range of linear covariance models $\mathscr{L}$. For each model, we conducted the following experiment. We repeatedly generated sample covariance matrices $S \in \mathbb{S}^n_+$. This was done as follows. We first sample a matrix $X \in \mathbb{R}^{n \times n}$ by picking each entry independently from a normal distribution with mean zero and variance one. And then we set $S := XX^T/n$. This is equivalent to sampling $nS \in \mathbb{S}^n_+$ from the standard Wishart distribution with n degrees of freedom.

For each of the generated sample covariance matrices S, we computed the real solutions of the likelihood equations (4). From these, we identified the set of all local maxima in $\mathbb{S}^n$, and we extracted its subset of local maxima in the positive definite cone $\mathbb{S}^n_+$. We recorded the numbers of these local maxima. Moreover, we kept track of the fraction of instances S for which there were multiple (positive define) local maxima. In Table 3 we present our results for $n = 5$ and generic linear subspaces $\mathscr{L}$.

	m												
	2	3	4	5	6	7	8	9	10	11	12	13	14
ML degree	7	37	135	361	753	1245	1625	1661	1323	801	347	97	15
max	2	3	3	5	5	5	5	6	7	5	4	2	1
max pd	1	2	3	3	4	4	4	4	5	5	4	2	1
multiple	0.4%	5.8	13.8	31.2	37.2	39.0	40.6	37.4	32.0	20.4	13.8	3.0	0.0
multiple pd	0.0%	4.6	11.2	22.4	25.2	31.6	33.0	34.8	29.6	19.4	13.0	3.0	0.0

Table 3. Experiments for generic m-dimensional linear subspaces of $\mathbb{S}^5$.

	tree number										
	1	2	3	4	5	6	7	8	9	10	11
ML degree	37	37	81	31	27	31	31	27	13	17	17
max	3	3	4	3	3	3	4	3	3	3	3
max pd	3	2	3	3	3	3	2	2	3	3	3
multiple	21.2%	22.8	24.2	15.6	23.0	21.2	21.2	15.4	13.8	16.2	12.4
multiple pd	8.2%	9.4	14.0	10.0	15.8	13.0	12.2	8.8	13.8	16.2	12.4

Table 4. Experiments for eleven Brownian motion tree models with 5 leaves.

For each m between 2 and 14, we selected five generic linear subspaces $\mathscr{L}$ in the 15-dimensional space $\mathbb{S}^5$. Each linear subspace $\mathscr{L}$ was constructed by choosing a basis of positive definite matrices. The basis elements were constructed with the same sampling method as the sample covariance matrices. The ML degree of this linear covariance model is the corresponding entry in the $n = 5$ column on the left in Table 1. These degrees are repeated in the row named *ML degree* in Table 3.

For each model $\mathscr{L}$, we generated 100 sample covariance matrices S, and we solved the likelihood equations (4) using our software `LinearCovarianceModels.jl`. The row *max* denotes the largest number of local maxima that was observed in these 100 experiments. The row *multiple* gives the fraction of instances which resulted in two or more local maxima. These two numbers pertain to local maxima in $\mathbb{S}^5$. The rows *max pd* and *multiple pd* are the analogues restricted to the positive definite cone $\mathbb{S}^5_+$.

For an illustration, let us discuss the models of dimension $m = 7$. These equations (4) have 1245 complex solutions, but the number of real solutions is much smaller. Nevertheless, in two fifths of the instances (39.0%) there were two or more local maxima in $\mathbb{S}^5$. In one third of the instances (31.6%) the same happened $\mathbb{S}^5_+$. The latter is the case of interest in statistics. One instance had four local maxima in $\mathbb{S}^5_+$.

The second experiment we report concerns a combinatorially defined class of linear covariance models, namely the Brownian motion tree models in (1). We consider eleven combinatorial types of trees with 5 leaves. For each model we perform the experiment described above, but we now used 500 sample covariance matrices per model. Our results are presented in Table 4, in the same format as in Table 3.

The eleven trees are numbered by the order in which they appear in Table 5. For instance, tree 1 gives the 7-dimensional model in $\mathbb{S}^5_+$ whose covariance matrices are

$$\Sigma = \begin{bmatrix} \gamma_1 & \gamma_6 & \gamma_6 & \gamma_6 & \gamma_7 \\ \gamma_6 & \gamma_2 & \gamma_6 & \gamma_6 & \gamma_7 \\ \gamma_6 & \gamma_6 & \gamma_3 & \gamma_6 & \gamma_7 \\ \gamma_6 & \gamma_6 & \gamma_6 & \gamma_4 & \gamma_7 \\ \gamma_7 & \gamma_7 & \gamma_7 & \gamma_7 & \gamma_5 \end{bmatrix}.$$

This model has ML degree 37. Around eight percent of the instances led to multiple maxima among positive definite matrices. Up to three such maxima were observed.

The results reported in Tables 3 and 4 show that the maximal number of local maxima increases with the ML degree. But they do not increase as fast as one would expect from the growth of the ML degree. On the other hand, the frequency of observing multiple local maxima seems to be roughly related to the ML degree.

n	clades	ML degree	dual ML degree
5	{1, 2, 3, 4}	37	11
5	{1, 2}	37	11
5	{1, 2, 3}	81	16
5	{1, 2}, {3, 4, 5}	31	4
5	{1, 2}, {3, 4}	27	4
5	{1, 2, 3}, {1, 2, 3, 4}	31	4
5	{1, 2}, {1, 2, 3}	31	4
5	{1, 2}, {1, 2, 3, 4}	27	4
5	**{1, 2}, {3, 4}, {1, 2, 3, 4}**	13	1
5	**{1, 2}, {3, 4}, {1, 2, 5}**	17	1
5	**{1, 2}, {1, 2, 3}, {1, 2, 3, 4}**	17	1
6	{1, 2, 3, 4, 5}	95	26
6	{1, 2}	95	26
6	{1, 2, 3, 4}	259	44
6	{1, 2, 3}	259	44
6	{1, 2, 3}, {4, 5, 6}	221	16
6	{1, 2}, {3, 4, 5, 6}	101	11
6	{1, 2, 3, 4}, {1, 2, 3, 4, 5}	101	11
6	{1, 2}, {3, 4}	81	11
6	{1, 2}, {1, 2, 3}	101	11
6	{1, 2}, {3, 4, 5}	181	16
6	{1, 2}, {1, 2, 3, 4, 5}	81	11
6	{1, 2, 3}, {1, 2, 3, 4}	221	16
6	{1, 2, 3}, {1, 2, 3, 4, 5}	181	16
6	{1, 2}, {1, 2, 3, 4}	181	16
6	{1, 2}, {3, 4}, {5, 6}	63	4
6	{1, 2}, {3, 4}, {1, 2, 3, 4}	99	4
6	{1, 2}, {1, 2, 3}, {4, 5, 6}	115	4
6	{1, 2}, {3, 4, 5}, {3, 4, 5, 6}	115	4
6	{1, 2}, {3, 4, 5}, {1, 2, 3, 4, 5}	99	4
6	{1, 2}, {3, 4}, {1, 2, 5, 6}	83	4
6	{1, 2}, {3, 4}, {1, 2, 3, 4, 5}	63	4
6	{1, 2, 3}, {1, 2, 3, 4}, {1, 2, 3, 4, 5}	115	4
6	{1, 2}, {3, 4}, {1, 2, 5}	83	4
6	{1, 2}, {1, 2, 3}, {1, 2, 3, 4}	115	4
6	{1, 2}, {1, 2, 3}, {1, 2, 3, 4, 5}	83	4
6	{1, 2}, {1, 2, 3, 4}, {1, 2, 3, 4, 5}	83	4
6	**{1, 2}, {3, 4}, {5, 6}, {1, 2, 3, 4}**	53	1
6	**{1, 2}, {3, 4}, {1, 2, 5}, {3, 4, 6}**	61	1
6	**{1, 2}, {3, 4}, {1, 2, 3, 4}, {1, 2, 3, 4, 5}**	53	1
6	**{1, 2}, {3, 4}, {1, 2, 5}, {1, 2, 5, 6}**	61	1
6	**{1, 2}, {3, 4}, {1, 2, 5}, {1, 2, 3, 4, 5}**	53	1
6	**{1, 2}, {1, 2, 3}, {1, 2, 3, 4}, {1, 2, 3, 4, 5}**	61	1

Table 5. ML degrees and dual ML degrees for Brownian motion tree models with five and six leaves. Binary trees are in bold.

Here is an interesting observation to be made in Table 4. The last three trees, labeled 9, 10, and 11, are the binary trees. These have the maximum dimension $2n-2$. For these models, every local maximum in $\mathbb{S}^n$ is also in the positive definite cone $\mathbb{S}^n_+$. We also verified this for all binary trees with $n=6$ leaves. This is interesting since the positive-definiteness constraint is the hardest to respect in an optimization routine. It is tempting to conjecture that this persists for all binary trees with $n \geq 7$.

There is another striking observation in Table 5. The dual ML degree for binary trees is always equal to one. We shall prove in Theorem 7.3 that this holds for any n. This means that the dual MLE can be expressed as a rational function in the data S. Hence there is only one local maximum, which is therefore the global maximum.

We close this section with a few remarks on the important special case when the ML degree or the dual ML degree is equal to one. This holds if and only if each entry of the estimated matrix $\widehat{\Sigma}$ or $\check{\Sigma}$ is a rational function in the $\binom{n+1}{2}$ quantities s_{ij}.

Rationality of the MLE has received a lot of attention in the case of *discrete random variables*. See [30, §7.1] for a textbook reference. If the MLE of a discrete model is rational, then its coordinates are alternating products of linear forms in the data [30, Theorem 7.3.4]. This result due to Huh was refined in [11, Theorem 1]. At present we have no idea what the analogue in the Gaussian case might look like.

Problem 6.1. Characterize all Gaussian models whose MLE is a rational function.

In addition to the binary trees in Theorem 7.3, statisticians are familiar with a number of situations when the dual MLE is rational. The dual MLE is the MLE of a linear concentration model with the sample covariance matrix S replaced by its inverse W. This is studied in [28] and in many other sources on Gaussian graphical models and exponential families. The following result paraphrases [28, Theorem 4.3].

Proposition 6.2. *If a linear covariance model $\mathcal{L}$ is given by zero restrictions on Σ, then the dual ML degree is equal to one if and only if the associated graph is chordal.*

It would be interesting to extend this result to other combinatorial families, such as colored covariance graph models [17], including structured Toeplitz matrices.

The following example illustrates Problem 6.1 and raises some further questions.

Example 6.3. We present a linear covariance model such that both the MLE and the dual MLE are rational functions. Fix $n \geq 2$ and let $\mathcal{L}$ be the hyperplane with equation $\sigma_{12}=0$. By Proposition 6.2, the dual ML degree of $\mathcal{L}$ is one. The model is dual to the decomposable undirected graphical model with missing edge $\{1,2\}$.

Following [20; 28], we obtain the rational formula for its dual MLE:

$$\check{k}_{12} = W_{1,R} W_{R,R}^{-1} W_{R,2} \quad \text{and} \quad \check{k}_{ij} = w_{ij} \quad \text{for } (i,j) \neq (1,2). \tag{8}$$

Here $R = \{3,\ldots,n\}$ and $W_{\cdot,\cdot}$ is our notation for submatrices of $W = (w_{ij}) = S^{-1}$.

The ML degree of the model $\mathcal{L}$ is also one. To see this, we note that $\mathcal{L}$ is the DAG model with edges $i \to j$ whenever $i < j$ unless $(i,j) = (1,2)$. By [20, §5.4.1], the MLE of any Gaussian DAG model is

rational. In our case, we find $\widehat{K} = W + A$, where A is the $n \times n$ matrix which is zero apart from the upper left 2×2 block

$$A_{12,12} = \begin{bmatrix} s_{11}^{-1} & 0 \\ 0 & s_{22}^{-1} \end{bmatrix} - \frac{1}{s_{11}s_{22} - s_{12}^2} \begin{bmatrix} s_{22} & -s_{12} \\ -s_{12} & s_{11} \end{bmatrix}.$$

The entries in $\breve{\Sigma} = (\breve{K})^{-1}$ and $\widehat{\Sigma} = (\widehat{K})^{-1}$ are rational functions in the data s_{ij}. But, unlike in the discrete case of [11], here the rational functions are not products of linear forms. Problem 6.1 asks for an understanding of its irreducible factors.

Example 6.3 raises many questions. First of all, can we characterize all linear spaces $\mathcal{L}$ with rational formulas for their MLE, or their dual MLE, or both of them? Second, it would be interesting to study arbitrary models $\mathcal{L}$ that are hyperplanes. Consider the entries for $m = \binom{n+1}{2} - 1$ in Tables 1 and 3. We know from [28, §2.2] that the dual ML degree equals $n - 1$. The ML degree seems to be $2^{n-1} - 1$. In all cases there seems to be only one local (and hence global) maximum. How could one prove these observations? Finally, it is worthwhile to study the MLE when $\mathcal{L}^\perp$ is a generic symmetric matrix of rank r. What is the ML degree in terms of r and n?

7. Brownian motion tree models

We now study the linear space $\mathcal{L}_T$ associated with a rooted tree T with n leaves. The equations of $\mathcal{L}_T$ are $\sigma_{ij} = \sigma_{kl}$ whenever $\mathrm{lca}(i, j) = \mathrm{lca}(k, l)$. In the literature [13; 29] one assumes that the parameters θ_A in (1) are nonnegative. Here, we relax this hypothesis: we allow all covariance matrices in the spectrahedron $\mathcal{L}_T \cap \mathbb{S}_+^n$.

The ML degree and its dual do not depend on how the leaves of a tree are labeled but only on the tree topology. For fixed n each tree topology is uniquely identified by the set of clades. Since the root clade $\{1, \ldots, n\}$ and the leaf-clades $\{1\}, \ldots, \{n\}$ are part of every tree, they are omitted in our notation. For example, if $n = 5$, then the tree $\{\{1, 2\}, \{3, 4\}, \{3, 4, 5\}\}$ is the binary tree with four inner vertices corresponding to the three nontrivial clades mentioned explicitly. This tree is depicted in Figure 1.

We computed the ML degree and the dual ML degree of $\mathcal{L}_T$ for many trees T. In Table 5 we report results for five and six leaves. We notice that the dual ML degree is exactly one for all binary trees. This suggests that the dual MLE is a rational function. Our main result in this section (Theorem 7.3) says that this is indeed true.

The equations (6) for the dual ML degree can be written as $e_A^T(K - W)e_A = 0$ for all clades A. Here $W = (w_{ij})$ is given and $K^{-1} \in \mathcal{L}_T$ is unknown. We abbreviate

$$w_{A,B} = \sum_{i \in A} \sum_{j \in B} w_{ij} = e_A^T W e_B. \tag{9}$$

The same notation is used for general matrices. We present two examples with $n = 4$.

Example 7.1. Consider the tree with clades $\{1, 2\}, \{3, 4\}$, shown in [29, Figure 1]. The dual MLE $\breve{K}$ satisfies $\breve{k}_{ii} = w_{ii}$ for $i = 1, 2, 3, 4$, and $\breve{k}_{12} = w_{12}$, $\breve{k}_{34} = w_{34}$, and

$$\breve{k}_{ij} = w_{12,34} \frac{w_{i,12} w_{j,34}}{w_{12,12} w_{34,34}} \quad \text{for } i \in \{1, 2\} \text{ and } j \in \{3, 4\}.$$

Example 7.2. The tree with clades $\{1,2\}, \{1,2,3\}$ has $\check{k}_{ii} = w_{ii}$, $\check{k}_{12} = w_{12}$, and

$$\check{k}_{13} = w_{12,3}\frac{w_{1,12}}{w_{12,12}}, \qquad \check{k}_{14} = w_{123,4}\frac{w_{1,12}w_{12,123}}{w_{12,12}w_{123,123}}, \qquad \check{k}_{23} = w_{12,3}\frac{w_{2,12}}{w_{12,12}},$$
$$\check{k}_{24} = w_{123,4}\frac{w_{2,12}w_{12,123}}{w_{12,12}w_{123,123}}, \qquad \check{k}_{34} = w_{123,4}\frac{w_{123,3}}{w_{123,123}}.$$

Both examples were computed in Mathematica using the description of the Brownian motion tree model in terms of the inverse covariance matrix given in [29].

Recall that for $v \in V$ we write $\mathrm{de}(v)$ for the set of leaves of T that are descendants of v. The following theorem generalizes formulas in the above two examples. It is our main result in Section 7.

Theorem 7.3. *Consider the model $\mathcal{L}_T$ given by a rooted binary tree T with n leaves. The dual MLE $\check{K} = (\check{k}_{ij})$ satisfies $\check{k}_{ii} = w_{i,i}$ for all i, and its off-diagonal entries are*

$$\check{k}_{ij} = w_{A,B}\prod_{u\to v}\frac{w_{\mathrm{de}(v),\mathrm{de}(u)}}{w_{\mathrm{de}(u),\mathrm{de}(u)}} \quad \text{for } 1 \le i < j \le n. \tag{10}$$

Here A, B are the clades of the two children of $\mathrm{lca}(i,j)$. The product is over all edges $u \to v$ of T, except for the two edges with $u = \mathrm{lca}(i,j)$, on the path from i to j in T.

Proof. Define $p_{ij} = -k_{ij}$ for $1 \le i < j \le n$ and $p_{0i} = \sum_{j=1}^n k_{ij}$ for $1 \le i \le n$. By [29, Theorem 1.2], in the new coordinates $\mathcal{L}_T$ is a toric variety with a monomial parametrization $p_{0i} = 1/t_i$ and $p_{ij} = t_{\mathrm{lca}(i,j)}/(t_i t_j)$. The condition $\Sigma \in \mathcal{L}_T$, $K\Sigma = I_n$ in (6) is therefore equivalent to requiring that the coordinates p_{ij}, p_{0i} admit such a monomial parametrization.

The last condition, $K - W \in \mathcal{L}_T{}^{\perp}$, in (6) means that $k_{A,A} = w_{A,A}$ for every clade A of T. This can be rewritten in the new coordinates as

(i) $\sum_{j\neq i} p_{ij} = w_{i,i}$ for all $1 \le i \le n$, and

(ii) $p_{A,B} = -w_{A,B}$ for all inner vertices u of T, where $A \mid B$ is the partition of $\mathrm{clade}(u) = A \cup B$ given by the two children of u.

Now fix u with clade partition $A \mid B$ as above, so $u = \mathrm{lca}(i,j)$ for all $i \in A$ and $j \in B$. The parametrization $p_{0i} = 1/t_i$ and $p_{ij} = t_{\mathrm{lca}(i,j)}/(t_i t_j)$ yields

$$p_{A,B} = t_u \cdot \sum_{i\in A}\frac{1}{t_i}\cdot\sum_{j\in B}\frac{1}{t_j} = t_u \cdot p_{0,A}\cdot p_{0,B}.$$

Using the equations in (ii), we obtain

$$t_u = -\frac{w_{A,B}}{p_{0,A}p_{0,B}} \quad \text{and hence} \quad p_{ij} = -w_{A,B}\frac{p_{0i}}{p_{0,A}}\frac{p_{0j}}{p_{0,B}}. \tag{11}$$

We claim that the following identity holds for any clade $A \subseteq [n]$:

$$p_{0,A} = w_{[n],[n]}\prod_{u\to v}\frac{w_{\mathrm{de}(v),\mathrm{de}(u)}}{w_{\mathrm{de}(u),\mathrm{de}(u)}}, \tag{12}$$

where the product is over all edges $u \to v$ of T in the path from the root to the node with clade A. Note that (11) and (12) imply (10) and so the theorem.

We now prove (12). Since $p_{0,[n]} = w_{[n],[n]}$, the claim holds for $A = [n]$. Fix a clade $A \subset [n]$ and assume (12) for all clades $A_1 \subset \cdots \subset A_k \subset A_{k+1} = [n]$ strictly containing $A_0 = A$. For each $i = 0, \ldots, k$ denote

$$\alpha_i := w_{A_{k+1},A_{k+1}} \frac{w_{A_k,A_{k+1}}}{w_{A_{k+1},A_{k+1}}} \cdots \frac{w_{A_i,A_{i+1}}}{w_{A_{i+1},A_{i+1}}}.$$

By the induction hypothesis $p_{0,A_i} = \alpha_i$ for all $i = 1, \ldots, k$. Our goal is to prove that $p_{0,A} = \alpha_0$. The clades $A_1 \setminus A, \ldots, A_{k+1} \setminus A_k$ form a partition of $\bar{A} = [n] \setminus A$. We have

$$\begin{aligned} p_{0,A} &= w_{A,A} + k_{A,\bar{A}} = w_{AA} - p_{A,A_1\setminus A} - p_{A,A_2\setminus A_1} - \cdots - p_{A,A_{k+1}\setminus A_k} \\ &= w_{A,A} + w_{A,A_1\setminus A} + \sum_{i=1}^{k} w_{A_i,A_{i+1}\setminus A_i} \frac{p_{0A}}{p_{0,A_i}}. \end{aligned} \tag{13}$$

Here the last equality follows because for every $i \in A$ and every $j \in A_{l+1} \setminus A_l$ the vertex $u = \mathrm{lca}(i, j)$ is the same. The clades of the children of u are A_l and $A_{l+1} \setminus A_l$. Therefore, using (11), we get $p_{A,A_{l+1}\setminus A_l} = w_{A_l,A_{l+1}\setminus A_l} p_{0,A}/p_{0,A_l}$. We rewrite (13) as

$$w_{A,A_1} = p_{0,A}\left(1 - \sum_{i=1}^{k} \frac{w_{A_i,A_{i+1}} - w_{A_i,A_i}}{\alpha_i}\right). \tag{14}$$

To simplify the bracketed expression, note that $w_{A_i,A_{i+1}}/\alpha_i = w_{A_{i+1},A_{i+1}}/\alpha_{i+1}$, and so

$$1 - \sum_{i=1}^{k} \frac{w_{A_i,A_{i+1}} - w_{A_i,A_i}}{\alpha_i} = 1 + \frac{w_{A_1,A_1}}{\alpha_1} - \frac{w_{A_k,A_{k+1}}}{\alpha_k} = \frac{w_{A_1,A_1}}{\alpha_1}.$$

Plugging this back into (14) gives

$$p_{0,A} = \alpha_1 \frac{w_{A,A_1}}{w_{A_1,A_1}} = \alpha_0.$$

This proves the correctness of (12).

We have shown that (6) implies the rational formula (10) for $\check{K}$ in terms of W. To argue that this is the MLE, we need that $W \in \mathbb{S}^n_+$ implies $\check{K} \in \mathbb{S}^n_+$. For this, we use an analytic argument. Since W is positive definite, the dual likelihood function has a unique maximum $K = W$ over the whole cone $\mathbb{S}^n_+$. The model $\mathscr{L}_T \cap \mathbb{S}^n_+$ is a relatively closed subset of $\mathbb{S}^n_+$ and so the dual likelihood restricted to this set attains its maximum. The ML degree is equal to one and so there is at most one optimum in $\mathscr{L}_T$. We conclude there is exactly one optimum in $\mathscr{L}_T \cap \mathbb{S}^n_+$ and it is equal to $\check{K}$. □

In our concluding example we compare the MLE and its dual in a special case.

Example 7.4. Fix the five-leaf tree in Figure 1, with clades $\{1, 2\}, \{3, 4\}, \{3, 4, 5\}$. For simplicity assume that the data-generating distribution has all parameters θ_A in (1) equal to one. For each sample size $n = 50, 200, 500, 5000$, we run 1000 iterations to obtain a simple Monte Carlo estimator of the mean squared errors as measured by $\mathbb{E}\|\widehat{\Sigma} - \Sigma^*\|^2$ and $\mathbb{E}\|\check{\Sigma} - \Sigma^*\|^2$, where Σ^* is the true covariance matrix and

$\|\cdot\|$ is a given matrix norm. To have a direct comparison between both estimators we also approximate $\mathbb{E}\|\widehat{\Sigma} - \breve{\Sigma}\|^2$. We obtain the following numbers for the operator norm:

	50	200	500	5000
approx. $\mathbb{E}\|\widehat{\Sigma} - \Sigma^*\|^2$	5.44	1.30	0.55	0.05
approx. $\mathbb{E}\|\breve{\Sigma} - \Sigma^*\|^2$	5.28	1.31	0.55	0.05
approx. $\mathbb{E}\|\widehat{\Sigma} - \breve{\Sigma}\|^2$	0.38	0.02	0.00	0.00

We see that the two estimators have essentially the same statistical performance. On average, they lie very close to each other. The dual MLE, which is available in closed form, thus offers a very attractive alternative to the MLE. Similar results were obtained for the Frobenius norm and the ℓ_∞-norm but they are not reported here.

The estimates in the previous example were computed by evaluating the function in Theorem 7.3. We stress that nonlinear algebra and our software [31] played an essential role in getting to this point. Namely, with computations as described in Section 5, we created Table 5. After seeing that table, we conjectured that the dual MLE for binary trees is one. This led us to find the rational formula. The expression (10) is an alternating product of linear forms, reminiscent of [11, Theorem 1]. However, this structure does not generalize, by Example 6.3, thus underscoring Problem 6.1.

Acknowledgements

We thank Steffen Lauritzen for useful discussions and for providing references. We also thank Jane Coons, Orlando Marigliano, and Michael Ruddy for their comments. Zwiernik was supported by Spanish government grants RYC-2017-22544, PGC2018-101643-B-I00, and SEV-2015-0563 and by the Ayudas Fundación BBVA a Equipos de Investigación Cientifica 2017. Timme was supported by the Deutsche Forschungsgemeinschaft Graduiertenkolleg *Facets of Complexity* (GRK 2434).

References

[1] T. W. Anderson, "Estimation of covariance matrices which are linear combinations or whose inverses are linear combinations of given matrices", pp. 1–24 in *Essays in probability and statistics*, edited by R. C. Bose et al., University of North Carolina, Chapel Hill, NC, 1970.

[2] P. Breiding and S. Timme, "HomotopyContinuation.jl: a package for homotopy continuation in Julia", pp. 458–465 in *Mathematical software — ICMS 2018* (South Bend, IN, 2018), edited by J. H. Davenport et al., Lecture Notes in Computer Science **10931**, Springer, 2018.

[3] J. P. Burg, D. G. Luenberger, and D. L. Wenger, "Estimation of structured covariance matrices", *Proceedings of the IEEE* **70**:9 (1982), 963–974.

[4] A. Martín del Campo and J. I. Rodriguez, "Critical points via monodromy and local methods", *J. Symbolic Comput.* **79**:3 (2017), 559–574.

[5] S. Chaudhuri, M. Drton, and T. S. Richardson, "Estimation of a covariance matrix with zeros", *Biometrika* **94**:1 (2007), 199–216.

[6] E. S. Christensen, "Statistical properties of I-projections within exponential families", *Scand. J. Statist.* **16**:4 (1989), 307–318.

[7] J. I. Coons, O. Marigliano, and M. Ruddy, "Maximum likelihood degree of the two-dimensional linear Gaussian covariance model", *Alg. Stat.* **11**:2 (2020). Preprint available at arXiv 1909.04553v2.

[8] D. R. Cox and N. Wermuth, "Linear dependencies represented by chain graphs", *Statist. Sci.* **8**:3 (1993), 204–218.

[9] C. Dellacherie, S. Martinez, and J. San Martin, *Inverse M-matrices and ultrametric matrices*, Lecture Notes in Mathematics **2118**, Springer, 2014.

[10] M. Drton and T. Richardson, "A new algorithm for maximum likelihood estimation in Gaussian graphical models for marginal independence", pp. 184–191 in *Proceedings of the Nineteenth Annual Conference on Uncertainty in Artificial Intelligence*, Morgan Kaufmann, San Francisco, 2003.

[11] E. Duarte, O. Marigliano, and B. Sturmfels, "Discrete statistical models with rational maximum likelihood estimator", preprint, 2019. arXiv 1903.06110v1

[12] J. Fan, Y. Liao, and H. Liu, "An overview of the estimation of large covariance and precision matrices", *Econom. J.* **19**:1 (2016), C1–C32.

[13] J. Felsenstein, "Maximum-likelihood estimation of evolutionary trees from continuous characters", *Am. J. Hum. Genet.* **25**:5 (1973), 471–492.

[14] J. D. Hamilton, *Time series analysis*, Princeton University, 1994.

[15] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer, 2001.

[16] L. Ho and C. Ané, "A linear-time algorithm for Gaussian and non-Gaussian trait evolution models", *Syst. Biol.* **63**:3 (2014), 397–408.

[17] S. Højsgaard and S. L. Lauritzen, "Graphical Gaussian models with edge and vertex symmetries", *J. R. Stat. Soc. Ser. B Stat. Methodol.* **70**:5 (2008), 1005–1027.

[18] G. Kauermann, "On a dualization of graphical Gaussian models", *Scand. J. Statist.* **23**:1 (1996), 105–116.

[19] M. Kummer and C. Vinzant, "The Chow form of a reciprocal linear space", *Michigan Math. J.* **68**:4 (2019), 831–858.

[20] S. L. Lauritzen, *Graphical models*, Oxford Statistical Science Series **17**, Clarendon, New York, 1996.

[21] A. Leykin, J. I. Rodriguez, and F. Sottile, "Trace test", *Arnold Math. J.* **4**:1 (2018), 113–125.

[22] G. M. Marchetti, "Independencies induced from a graphical Markov model after marginalization and conditioning: the R package **ggm**", *J. Stat. Softw.* **15**:6 (2006).

[23] M. Michałek and B. Sturmfels, "Invitation to nonlinear algebra", book in progress, https://personal-homepages.mis.mpg.de/michalek/NonLinearAlgebra.pdf.

[24] M. I. Miller and D. L. Snyder, "The role of likelihood and entropy in incomplete-data problems: applications to estimating point-process intensities and Toeplitz constrained covariances", *Proceedings of the IEEE* **75**:7 (1987), 892–907.

[25] A. P. Morgan and A. J. Sommese, "Coefficient-parameter polynomial continuation", *Appl. Math. Comput.* **29**:2 (1989), 123–160.

[26] M. Pourahmadi, "Joint mean-covariance models with applications to longitudinal data: unconstrained parameterisation", *Biometrika* **86**:3 (1999), 677–690.

[27] A. J. Sommese and C. W. Wampler, II, *The numerical solution of systems of polynomials arising in engineering and science*, World Scientific, Hackensack, NJ, 2005.

[28] B. Sturmfels and C. Uhler, "Multivariate Gaussian, semidefinite matrix completion, and convex algebraic geometry", *Ann. Inst. Statist. Math.* **62**:4 (2010), 603–638.

[29] B. Sturmfels, C. Uhler, and P. Zwiernik, "Brownian motion tree models are toric", preprint, 2019. arXiv 1902.09905v1

[30] S. Sullivant, *Algebraic statistics*, Graduate Studies in Mathematics **194**, American Mathematical Society, Providence, RI, 2018.

[31] S. Timme, `LinearCovarianceModels.jl`, 2019, https://github.com/saschatimme/LinearCovarianceModels.jl. Julia package, version 0.1.2.

[32] P. Zwiernik, C. Uhler, and D. Richards, "Maximum likelihood estimation for linear Gaussian covariance models", *J. R. Stat. Soc. Ser. B. Stat. Methodol.* **79**:4 (2017), 1269–1292.

Received 2019-09-04. Revised 2020-01-23. Accepted 2020-04-30.

BERND STURMFELS: bernd@mis.mpg.de
Max Planck-Institute for Mathematics in the Sciences, Leipzig, Germany

and

Department of Mathematics, University of California, Berkeley, Berkeley, CA, United States

SASCHA TIMME: timme@math.tu-berlin.de
Institut für Mathematik, Technische Universität Berlin, Berlin, Germany

PIOTR ZWIERNIK: piotr.zwiernik@upf.edu
Department of Economics and Business, Universitat Pompeu Fabra, Barcelona, Spain

Algebraic Statistics
Vol. 11, No. 1, 2020
https://doi.org/10.2140/astat.2020.11.53

EXPECTED VALUE OF THE ONE-DIMENSIONAL EARTH MOVER'S DISTANCE

REBECCA BOURN AND JEB F. WILLENBRING

From a combinatorial point of view, we consider the earth mover's distance (EMD) associated with a metric measure space. The specific case considered is deceptively simple: Let the finite set of integers $[n] = \{1, \dots, n\}$ be regarded as a metric space by restricting the usual Euclidean distance on the real numbers. The EMD is defined on ordered pairs of probability distributions on $[n]$. We provide an easy method to compute a generating function encoding the values of EMD in its coefficients, which is related to the Segre embedding from projective algebraic geometry. As an application we use the generating function to compute the expected value of EMD in this one-dimensional case. The EMD is then used in clustering analysis for a specific data set.

1. Introduction

Fix a positive integer n. We will denote the finite set of integers $\{1, \dots, n\}$ by $[n]$. By a *probability measure* on $[n]$ we mean, as usual, a nonnegative real-valued function f on the set $[n]$ such that $f(1) + \cdots + f(n) = 1$. By the *probability simplex* on $[n]$ we mean the set of all probability measures on $[n]$, denoted $\mathcal{P}_n$. We view $\mathcal{P}_n$ as embedded in $\mathbb{R}^n$. Given $\mu, \nu \in \mathcal{P}_n$ define the set of joint distribution

$$\mathcal{J}_{\mu\nu} = \left\{ J \in \mathbb{R}^{n\times n} : \begin{array}{l} J \text{ is a nonnegative real number } n \text{ by } n \text{ matrix such that} \\ \sum_{i=1}^n J_{ij} = \mu_j \text{ for all } j \text{ and } \sum_{j=1}^n J_{ij} = \nu_i \text{ for all } i \end{array} \right\}.$$

For results concerning the geometry of $\mathcal{J}_{\mu\nu}$ see [3] and [14] where they are referred to as *transportation polytopes* and *discrete copulas* respectively.

The *earth mover's distance* is defined as

$$\mathbb{EMD}(\mu, \nu) = \inf_{J\in\mathcal{J}_{\mu\nu}} \sum_{i,j=1}^{n} |i - j| J_{ij}.$$

We remark that the set $\mathcal{P}_n \times \mathcal{P}_n$ is a compact subset of $\mathbb{R}^{2n}$ and so by continuity the infinum is actually a minimum value. Also, the EMD is sometimes referred to by other names, for example, in a two-dimensional setting it is called the *image distance*. More generally it is called the Wasserstein metric (see [11]).

We recall that the set of all finite distributions, $\mathcal{P}_n$, embeds as a compact polyhedron on a hyperplane in $\mathbb{R}^n$ and inherits Lebesgue measure and has finite volume. We normalize this measure so that the total mass of $\mathcal{P}_n$ is one. We then obtain a probability measure on $\mathcal{P}_n$, which is uniform. Similarly, $\mathcal{P}_n \times \mathcal{P}_n$ may be embedded in $\mathbb{R}^n \times \mathbb{R}^n$ and can be given the (uniform) product probability measure.

2010: primary 05E40; secondary 62H30.
Keywords: earth mover's distance, generating function, Segre embedding, spectral graph theory, clustering.

From its definition, the function $\mathbb{EMD}$ is a metric on $\mathcal{P}_n$. The subject of this paper concerns the expected value of $\mathbb{EMD}$ with respect to the uniform probability measure. In this light, we define a function $\mathcal{M}$ on ordered pairs of nonnegative integers, (p, q), as

$$\mathcal{M}_{p,q} = \frac{(p-1)\mathcal{M}_{p-1,q} + (q-1)\mathcal{M}_{p,q-1} + |p-q|}{p+q-1} \tag{1-1}$$

with $\mathcal{M}_{p,q} = 0$ if either p or q is not positive. Let $\mathcal{M}_n = \mathcal{M}_{n,n}$ for any nonnegative integer n.

We will prove the following theorem in Section 5.

Theorem 1. *Fix a positive integer n. Let $\mathcal{P}_n \times \mathcal{P}_n$ be given the uniform probability measure defined by Lebesgue measure from the embedding into $\mathbb{R}^{2n}$. The expected value of $\mathbb{EMD}$ on $\mathcal{P}_n \times \mathcal{P}_n$ is $\mathcal{M}_n$.*

From a theoretical point of view, this paper concerns the expected value of $\mathbb{EMD}$. Additionally, we consider a discrete version of the EMD and compute the mean. In turn, we discuss the relationship to cluster analysis. Then, we end with a comparison of the theoretical results to grade distributions where we have noticed persistent clustering.

The above theorem is obtained as a limit of a discrete version of $\mathbb{EMD}$ (denoted by EMD_s, for nonnegative integer s) which is described using a generating function. The generating function is a deformation of the Hilbert series of the Segre embedding. Some standard tools from algebraic combinatorics show up in a new way in the proofs.

From a practical point of view, we will also consider a "real world" data set with a finite number of joint probability measures derived from letter grade distributions. Specifically, we consider a network of grade distributions from the University of Wisconsin - Milwaukee campus, where two nodes are joined when the EMD_s between them falls below a prespecified distance threshold. Determining this threshold so that data features are revealed is a subject of research. The expected value of the EMD in the uniformly random situation helps guide this choice.

The family of networks obtained by varying the distance threshold will be used for metric hierarchical clustering. When the threshold is set so that the network is connected, the spectrum of the corresponding Laplacian matrix (see [1]) will be computed, as it also relates to clustering. Our use of the EMD in this context should be viewed as an attempt at exploratory data analysis to identify the "communities" in this network rather than rigorous hypothesis testing.

2. Nontechnical preliminaries

In this section we consider some specific examples of finite distributions. A motivating situation comes from grade distributions, which in the United States are often considered with five outcomes: A, B, C, D, and F. The standard grade point average (GPA) assigns 4.0 to A, 3.0 to B, 2.0 to C, 1.0 to D, and 0.0 to F. The relative distances between these five grades is computed by the absolute value of the difference of point values. That is, a B grade is three units away from an F, while one unit away from an A.

Suppose we are given three distributions in the five grade setting for classes with $s = 30$ students, e.g.,

	A	*B*	*C*	*D*	*F*
X	0	19	8	2	1
Y	12	2	5	11	0
Z	2	20	2	3	3

To compare distribution X to distribution Y, one notices that if the 12 A grades in Y were moved down to B, 5 C grades moved up to B, 8 D grades moved up to C, and one D grade moved down to F, then the distributions would be identical. The matrix

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 12 & 2 & 5 & 0 & 0 \\ 0 & 0 & 0 & 8 & 0 \\ 0 & 0 & 0 & 2 & 0 \\ 0 & 0 & 0 & 1 & 0 \end{bmatrix}$$

encodes the "conversion". That is, if the rows and columns correspond to the grades (A,B,C,D,F) then the entry in row i and column j records how many grades to move from position i in Y to position j in X. The entries on the diagonal reflect no "earth" movement, while the entries in the first sub- and super-diagonals reflect one unit of movement. The row sums return the X distribution, while the column sums return the Y distribution. In total, the value of EMD_s is 26.

The Y and Z distributions compare as follows: move 10 B's up one unit to a grade of A, to reflect the fact that Y had 12 grades of A. We move 5 grades down from B to C, and 3 grades from B to D. This latter change is noted as a jump across two positions which will "cost" 2 units in EMD, and since there are 3 grades to move, this makes an overall contribution of 6. Finally, 2 C grades are moved down to D, and the 3 F grades in the Z distribution would be moved to D in the distribution Y.

The joint distribution matrix for Y and Z is

$$\begin{bmatrix} 2 & 10 & 0 & 0 & 0 \\ 0 & 2 & 0 & 0 & 0 \\ 0 & 5 & 0 & 0 & 0 \\ 0 & 3 & 2 & 3 & 3 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix}.$$

Interestingly, the total EMD is again 26.

Finally, the X and Z distributions are compared. The joint matrix is

$$\begin{bmatrix} 0 & 0 & 0 & 0 & 0 \\ 2 & 17 & 0 & 0 & 0 \\ 0 & 3 & 2 & 3 & 0 \\ 0 & 0 & 0 & 0 & 2 \\ 0 & 0 & 0 & 0 & 1 \end{bmatrix}.$$

The total EMD is 10.

The three distributions above have the same GPA of 2.5. We note that this is a feature, which we point out to give indication that the EMD will clearly distinguish between distributions even if the GPA is constant.

To help the reader gain some intuitive feel for the EMD we augment the three distributions by:

	A	B	C	D	F
U	13	13	0	0	4
V	9	1	13	2	5
W	9	7	8	6	0

As an exercise, one can compute the 36 pairwise distances between each of the six provided distributions. We computed them with Mathematica as shown below:

EMD	U	V	W	X	Y	Z
U	0	24	20	24	24	18
V	24	0	12	26	16	22
W	20	12	0	16	10	16
X	24	26	16	0	26	10
Y	24	16	10	26	0	26
Z	18	22	16	10	26	0

One can sample distributions of five grades with 30 students in many distinct ways. For each sampling method one can ask how the EMD is distributed. The sampling method could be chosen to accurately simulate synthetic data to match previously observed samples from a particular subject at a particular institution. Or, a prior distribution on grade distributions could be assumed, such as a discretization of the multivariate normal distribution.

Upon exploration of observed data one notices clear clustering of the distributions relative to the EMD. Indeed, if some distributions are encountered more frequently than others in a particular model then clustering should be expected. With this fact in mind one is led to question of sampling distributions at flat random. That is, sampling independently with each distribution being equally likely. The theoretical behavior of the uniform model can then be compared to observed data. Clustering in the uniform model can be considered "random", while additional observed clustering in a specific data set is likely related to a causal feature. Statistics describing clustering should be understood for the uniform distribution as it has maximal entropy.

For any given distribution, one seeks a theoretical understanding of any given descriptive statistic. The present article restricts the focus to the mean of EMD. Other statistics will be considered in future work. Moreover, we focus on the uniform distribution only over the space of finite probability distributions.

Finally, we note that the results of this article imply that the mean discrete EMD on 30 student, five grade distributions is slightly larger than 26. The maximum EMD is 120 reflecting the fact that the distance between all 30 students with A grades is 120 units away from the distribution with all 30 students with F grade. Such large values of EMD are unlikely. The distribution of actual grade data, as we expect, is skewed to the right (i.e., the mean is larger than the median).

3. Technical preliminaries

We now recall basic notation from combinatorics and linear algebra that is used throughout the paper.

3.1. ***Notation from linear algebra.*** Let $\mathbb{M}_{n,m}$ be the vector space of real matrices with n rows and m columns. Throughout, we assume that the field of scalars is the real numbers, $\mathbb{R}$. For i and j with $1 \le i \le n$, $1 \le j \le m$ we let $e_{i,j}$ denote the n by m matrix with 1 in the i-th row, j-th column, and 0 elsewhere. A matrix, $M \in \mathbb{M}_{n,m}$ is written as $M = (M_{i,j})$ where $M_{i,j}$ is the entry in the i-th row and j-th column. So, $M = \sum M_{i,j}\, e_{i,j}$. We assume standard notation for the algebra of matrices. For example, the standard inner product of $X, Y \in \mathbb{M}_{n,m}$ is

$$\langle X, Y\rangle = \mathrm{Trace}(X^T Y).$$

In the case that $m = 1$ we write $e_i = e_{i,1}$ for $1 \le i \le n$. As usual, Let $\mathbb{R}^n$ denote the n-dimensional real vector space consisting of column vectors of length n. The set $\{e_1, \ldots, e_n\}$ is a basis for $\mathbb{R}^n$. For our purposes a very useful alternative basis is given by

$$\omega_j = \sum_{i=1}^{j} e_i$$

where $1 \le j \le n$. We call the set of ω_j the *fundamental basis for* $\mathbb{R}^n$. The terminology here comes from the the root system of type A in Lie theory (see [7]).

Let the orthogonal complement, $\omega_n^{\perp}$, to ω_n be denoted by

$$\mathbb{R}_0^n = \left\{ v \in \mathbb{R}^n \mid \langle v, \omega_n\rangle = 0 \right\}.$$

Column vectors in $\mathbb{R}_0^n$ have coordinates that sum to zero. We let

$$\pi_0 : \mathbb{R}^n \to \mathbb{R}_0^n, \quad \pi_0(v) = v - \frac{\langle v, \omega_n\rangle}{n}\omega_n,$$

denote the orthogonal projection from $\mathbb{R}^n$ onto $\mathbb{R}_0^n$. The image of π_0 is $\mathbb{R}_0^n$, and the kernel contains ω_n. For $1 \le j \le n-1$, let $\tilde{\omega}_j = \pi_0(\omega_j)$. Observe that $\tilde{\omega}_1, \cdots, \tilde{\omega}_{n-1}$ span $\mathbb{R}_0^n$, and by considering dimension, form a basis for $\mathbb{R}_0^n$. We will call this the *fundamental basis for* $\mathbb{R}_0^n$.

The subspace $\mathbb{R}_0^n$ has another basis that is of importance to us:

$$\Pi = \{\alpha_1, \ldots, \alpha_{n-1}\},$$

where $\alpha_j = e_j - e_{j+1}$. We refer to Π as the *simple basis for* $\mathbb{R}_0^n$. An essential point is that Π is dual to the fundamental basis. That is, $\langle \alpha_i, \tilde{\omega}_j\rangle$ vanishes if $i \ne j$ and equals 1 if $i = j$.

Let $\mathcal{E} : \mathbb{R}^n \to \mathbb{R}$ be defined as

$$\mathcal{E}(v) = |v_1| + |v_1 + v_2| + |v_1 + v_2 + v_3| + \cdots + |v_1 + \cdots + v_n| \qquad \text{if } v = \sum_{j=1}^{n} v_j e_j \in \mathbb{R}^n.$$

The restriction of $\mathcal{E}$ to $\mathbb{R}_0^n \subset \mathbb{R}^n$, denoted by the same symbol, satisfies

$$\mathcal{E}(v) = \sum_{i=1}^{n-1} |c_i| \qquad \text{if } v = \sum_{i=1}^{n-1} c_i \alpha_i \in \mathbb{R}_0^n.$$

This is easily seen since the fundamental basis is dual to the simple basis relative to the standard inner product, and

$$\langle v, \omega_j \rangle = v_1 + \cdots + v_j$$

for $1 \leq j \leq n$.

In Section 5 we will prove this equality:

Theorem 2. *For all* $\mu, \nu \in \mathcal{P}_n$,

$$\mathbb{EMD}(\mu, \nu) = \mathcal{E}(\mu - \nu).$$

This allows for a much more explicit combinatorial analysis of EMD. For situations in which the metric space is not a subset of the real line, the analysis is more difficult. Indeed, $\mathbb{EMD}$ is often computed as an optimization problem that minimizes the cost under the constraints imposed by the marginal distribution. Consequently, the computational complexity is the same as for linear programming.

3.2. ***Compositions and related combinatorics.*** Let $\mathbb{N}$ be the set of nonnegative integers. Given $s \in \mathbb{N}$, and a positive integer n, define

$$\mathcal{C}(s, n) = \{(a_1, a_2, \ldots, a_n) \in \mathbb{N}^n : a_1 + \cdots + a_n = s\}.$$

An element of the set $\mathcal{C}(s, n)$ will be referred to as a *composition of s into n parts*. (We note that in some places of the literature these are referred to as *weak compositions*, since we allow zero. However, the distinction is not needed for us.)

It is an elementary fact that there are $\binom{s+n-1}{n-1}$ compositions, and therefore for fixed n, the number of compositions of s grows as a polynomial function of s with degree $n-1$. An essential fact for this paper is the asymptotic approximation

$$\binom{s+n-1}{n-1} \sim \frac{s^{n-1}}{(n-1)!}.$$

As in the introduction, given compositions μ and ν of s we let $\mathcal{J}_{\mu\nu}$ denote the set of n by n matrices with row sums μ and column sums ν. There is a slight difference here in that we are not requiring μ and ν to be normalized to sum to one. In the same light, EMD can be extended as a metric on $\mathcal{C}(s, n)$.

We fix an n by n matrix C with i-th row and j-th column entry to be $|i - j|$. That is,

$$C = \begin{bmatrix} 0 & 1 & 2 & \cdots & n-1 \\ 1 & 0 & 1 & \cdots & n-2 \\ 2 & 1 & 0 & \cdots & n-3 \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ n-1 & n-2 & n-3 & \cdots & 0 \end{bmatrix}.$$

So, for $\mu, \nu \in \mathcal{C}(s, n)$ and regarding the set $\mathcal{J}_{\mu\nu}$ as nonnegative integer matrices with prescribed row and column sums, we arrive at

$$\mathrm{EMD}_s(\mu, \nu) = \min_{J \in \mathcal{J}_{\mu\nu}} \langle J, C \rangle,$$

which is a discrete version of EMD. When we take $s \to \infty$ we recover the value referred to in the introduction.

The function EMD may be further generalized to the case where C has p rows and q columns, with i, j entry $|i-j|$. In this case μ has p components and ν has q components (each a composition of s). This generalization will be needed in an induction argument in Section 5. However, applications need only consider the $p = q$ case.

A further generalization beyond the scope of this paper is to consider more general cost matrices than C. This is equivalent to a variation of the metric.

4. Generating functions

In algebraic combinatorics it is often useful to record discrete data in a formal (multivariate) power series – sometimes called a *generating function*. By "formal" we mean that the variables are indeterminates rather than numbers. In fact, from this point of view one can consider formal power series that only converge at zero, yet encode combinatorial data in their coefficients. Consequently, convergence is not an issue. Nonetheless, our series are all geometric series expansions of rational functions, and so will be convergent if, say, all complex variables have modulus less than 1.

Starting from the viewpoint of algebraic combinatorics we define

$$H_n(z,t) := \sum_{s=0}^{\infty} \left(\sum_{(\mu,\nu)\in\mathcal{C}(s,n)\times\mathcal{C}(s,n)} z^{\mathrm{EMD}_s(\mu,\nu)} \right) t^s,$$

where t and z are indeterminates. We see that the coefficient of t^s in $H_n(z,t)$ is a polynomial in z whose coefficients record the distribution of the values of EMD_s.

It is useful to see the first few values of H, which we compute using Mathematica and Theorem 3:

$$H_1(z,t) = \frac{1}{1-t}, \quad H_2(z,t) = \frac{tz+1}{(1-t)^2(1-tz)}, \quad H_3(z,t) = \frac{-t^3z^4-t^2(2z+1)z^2+t(z+2)z+1}{(1-t)^3(1-tz)^2(1-tz^2)}.$$

As before, when considering p by q matrices we can analogously define $H_{p,q}(z,t)$. We also extend the definition so that $H_{p,q} = 0$ if either of p or q is not positive. We obtain a similar series, namely

$$H_{p,q}(z,t) := \sum_{s=0}^{\infty} \left(\sum_{\substack{\mu\in\mathcal{C}(s,p)\\ \nu\in\mathcal{C}(s,q)}} z^{\mathrm{EMD}_s(\mu,\nu)} \right) t^s.$$

The function EMD_s is defined since for $p < q$ we can regard p-tuples as q-tuples, by appending zeros.

Theorem 3. *For positive integers p and q,*

$$H_{p,q}(z,t) = \frac{H_{p-1,q}(z,t) + H_{p,q-1}(z,t) - H_{p-1,q-1}(z,t)}{1 - z^{|p-q|}t}$$

if $(p,q) \neq (1,1)$ and $H_{1,1} = 1/(1-t)$.

This proof will also be given in Section 5, after we have developed some of the consequences in the remainder of this section.

4.1. ***The partially ordered set*** $[p] \times [q]$**.** Recall that a *partially ordered set* is a set S together with a relation, $\preceq$, which is required to be reflexive, antisymmetric and transitive. In particular, given positive integers p and q, we define $S = [p] \times [q]$, and

$$(i, j) \preceq (i', j') \Longleftrightarrow i' - i \in \mathbb{N} \text{ and } j' - j \in \mathbb{N}$$

for $1 \leq i, i' \leq p$ and $1 \leq j, j' \leq q$, which is a partially ordered set.

It is important to note that not all elements are comparable with respect to this order. For example, if $p = q = 2$, clearly $(1, 2) \not\preceq (2, 1)$ and $(2, 1) \not\preceq (1, 2)$. We say that $(1, 2)$ and $(2, 1)$ are incomparable. A subset of S in which all pairs are comparable is called a *chain*.

Given a p by q matrix, J, we define the *support* as

$$\text{support}(J) := \{(i, j) : J_{ij} > 0\}.$$

Proposition 4. *Let p and q be positive integers, and $s \in \mathbb{N}$. For $\mu \in \mathcal{C}(s, p)$, $\nu \in \mathcal{C}(s, q)$ and $J \in \mathcal{J}_{\mu\nu}$, there exists a $J' \in \mathcal{J}_{\mu\nu}$ such that the support of J' is a chain in $[p] \times [q]$, and $\langle J', C\rangle \leq \langle J, C\rangle$.*

Proof. Suppose there exist incomparable elements (i, j), (i', j') such that $J_{ij}, J_{i'j'} > 0$. Without loss of generality assume $i' < i$, $j < j'$, and $0 < J_{ij} \leq J_{i'j'}$. We construct J' as follows:

$$\begin{bmatrix} & & \\ & & J_{i'j'} \\ & & \\ & J_{ij} & \end{bmatrix}$$

For $(k, l) \notin \{(i, j), (i', j'), (i', j), (i, j')\}$ let $J'_{kl} = J_{kl}$. Next, let $J'_{ij} = 0$ and

$$J'_{i'j'} = J_{i'j'} - J_{ij},$$

which is nonnegative. Next,

$$J'_{i'j} = J_{i'j} + J_{ij} \quad \text{and} \quad J'_{ij'} = J_{ij'} + J_{ij}.$$

The result follows. □

The point here is that we will only need to consider matrices J with support on a chain.

4.2. ***A special case of the Robinson-Schensted-Knuth correspondence.*** In this subsection we recall, in detail, a special case of the Robinson–Schensted–Knuth correspondence (RSK), see [6]. For those familiar with RSK, we consider the case where the Young diagrams have only one row; however, the exposition here does not require any knowledge of RSK.

Given $\mu \in \mathcal{C}(s, n)$ with $\mu = (\mu_1, \ldots, \mu_n)$ we define the *word* of μ to be a finite weakly increasing sequence of positive integers, $w = w(\mu) = w_1 w_2 w_3 \cdots$ where the number of times k occurs in w is equal to μ_k. For example, for $\mu = (3, 0, 2, 1, 0)$ we have $w = 111334$.

Note that the length of the word is equal to $s = \sum \mu_i$, and all components of w are at most n.

Next, for $s \in \mathbb{N}$ and positive integers p and q, we define

$$R(p, q; s) := \left\{ J \in \mathbb{M}_{p,q} : (\forall i, j),\, J_{ij} \in \mathbb{N},\, \sum_{i,j} J_{ij} = s \text{ and support}(J) \text{ is a chain} \right\}.$$

Proposition 5. *For a given $s \in \mathbb{N}$ and positive integers p and q, we have a bijection, Φ, between $\mathcal{C}(s, p) \times \mathcal{C}(s, q)$ and $R(p, q; s)$.*

Proof. Given $(\mu, \nu) \in \mathcal{C}(s, p) \times \mathcal{C}(s, q)$, let the words of μ and ν be, respectively,

$$u = u_1 u_2 u_3 \cdots u_s \qquad \text{with } 1 \le u_i \le p,$$
$$v = v_1 v_2 v_3 \cdots v_s \qquad \text{with } 1 \le v_j \le q.$$

Define a p by q matrix by

$$J_{ij} = |\{k : (u_k, v_k) = (i, j)\}|$$

for $1 \le i \le p$ and $1 \le j \le q$. Note that the support of J is a chain. We define $\Phi(\mu, \nu) = J$. Given J, we can recover μ and ν as the row and column sums of J. □

4.3. ***Rank one matrices and the Segre embedding.*** We let $\mathbb{D}^{\le k}(p, q)$ denote the set of p by q matrices with rank at most k, which is a closed affine algebraic set, called a *determinantal variety*. For a relatively recent expository article about the role these varieties play in algebraic geometry and representation theory see [4].

In this section we consider the $k = 1$ case, in our context. Define $P : \mathbb{R}^p \times \mathbb{R}^q \to \mathbb{M}_{p,q}$ by

$$P(v, w) = vw^T$$

for $v \in \mathbb{R}^p$ and $w \in \mathbb{R}^q$. Note that if $P(v, w) \neq 0$ then the rank is 1. In fact, the image of P consists of those matrices with rank at most 1. So if $p, q > 1$ then P is not surjective.

Injectivity of P fails as well since for nonzero $c \in \mathbb{R}$, v, w we have $P(v, w) = P(cv, \frac{1}{c}w)$. However, if we pass to projective space we recover an injective map.

In this light, let $\mathbb{RP}^n$ be an n-dimensional real projective space, that is:

$$\mathbb{RP}^n = \{\mathbb{R}v : 0 \neq v \in \mathbb{R}^{n+1}\}$$

where $\mathbb{R}v$ denotes the 1-dimensional subspace spanned by nonzero v. We will also write $\mathbb{RP}^n := \mathbb{P}(\mathbb{R}^{n+1})$.

The *Segre* embedding,

$$\mathbb{P}(\mathbb{R}^p) \times \mathbb{P}(\mathbb{R}^q) \to \mathbb{P}(\mathbb{R}^{pq})$$

is defined as follows: first, we note that we can identify $\mathbb{R}^{pq}$ with the p by q matrices by choosing bases. Next, given an ordered pair of projective points (i.e., one-dimensional subspaces) we can choose nonzero vectors v and w respectively. The value of the Segre embedding is the one-dimensional subspace in $\mathbb{M}_{p,q}$ spanned by the matrix $P(v, w)$. It is easily checked that this map is well-defined and injective.

The image of the Segre embedding gives rise to a projective variety structure on the set-cartesian product of the two projective varieties. The projective coordinate algebra of the Segre embedding is intimately related to $H_{p,q}(z, t)$, which we will see next.

Let m_{ij} be a choice of (algebraically independent) indeterminates. We consider the polynomial algebra

$$\mathcal{A}_{p,q} = \mathbb{R}[m_{ij} : 1 \leq i \leq p, 1 \leq j \leq q].$$

Then for $1 \leq i < i' \leq p$, and $1 \leq j < j' \leq q$ define

$$\Delta(i, i'; j, j') = \begin{vmatrix} m_{ij} & m_{ij'} \\ m_{i'j} & m_{i'j'} \end{vmatrix}.$$

The ideal, $\mathcal{I}$, generated by the $\Delta(i, i'; j, j')$ vanishes exactly on the matrices of rank 1. Conversely, any polynomial function that vanishes on the rank at most 1 matrices is in $\mathcal{I}$. The algebra of coordinate functions on the rank at most 1 matrices is then isomorphic to the quotient of $\mathcal{A}_{p,q}$ by $\mathcal{I}$. Define $\mathcal{R}(p, q) := \mathcal{A}_{p,q}/\mathcal{I}$.

The point here is that monomials involving variables which have indices that are not comparable with respect to $\preceq$ may be replaced (modulo $\mathcal{I}$) with comparable indices. That is, $m_{ij}m_{i'j'}$ can be replaced with $m_{i'j}m_{ij'}$. This process may be thought of as "straightening" and is related to the nonnegative integer matrices J with support in a chain. The matrix J may be thought of as the exponents in a monomial.

More generally, the situation may be put into the context of Gröbner bases. The cost matrix C used here assigns a number to each pair of indices. This number can be used to scale the degree of m_{ij}. Using this new notion of degree, we can set up a partial order of the monomials, which can then be extended (say, lexicographically) to a well ordering of the monomials that is compatible with multiplication. That is, we can create a term order (see [2]). The minors generating the ideal $\mathcal{I}$ are indeed a Gröbner basis. The complement of the ideal of leading terms is then a vector space basis for the quotient by $\mathcal{I}$.

For $s \in \mathbb{N}$, let $\mathcal{A}^s_{p,q}$ denote the subspace of homogeneous degree s polynomials, and set $\mathcal{R}^s_{p,q} = \mathcal{A}^s_{p,q}/(\mathcal{A}^s_{p,q} \cap \mathcal{I})$. Since $\mathcal{I}$ is generated by homogeneous polynomials, we have

$$\mathcal{R}_{p,q} = \bigoplus_{s=0}^{\infty} \mathcal{R}^s_{p,q}.$$

That is, we have an algebra gradation by polynomial degree.

The polynomial functions on $\mathbb{R}^p$ (resp. $\mathbb{R}^q$) will be denoted $\mathcal{A}_p$ (resp. $\mathcal{A}_q$). Given vectors $v \in \mathbb{R}^p$ and $w \in \mathbb{R}^q$ an element of the tensor product $\mathcal{A}_p \otimes \mathcal{A}_q$ defines a function on $\mathbb{R}^p \times \mathbb{R}^q$ with value $f(v)g(w)$. Given an element $(v, w) \in \mathbb{R}^p \times \mathbb{R}^q$, and $f \otimes g \in \mathcal{A}_p \otimes \mathcal{A}_q$, the value of $f \otimes g$ on (v, w) is given by $f(v)g(w)$. Extending by linearity we obtain an algebra isomorphism from the polynomials on $\mathbb{R}^p \times \mathbb{R}^q$ to the tensor product algebra $\mathcal{A}_p \otimes \mathcal{A}_q$.

The quadratic map P, defined above, gives rise to an algebra homomorphism,

$$P^* : \mathcal{R}_{p,q} \to \mathcal{A}_p \otimes \mathcal{A}_q,$$

defined such that $P^*(F)$ is a function on $\mathbb{R}^p \times \mathbb{R}^q$ from a function F on $\mathbb{D}^{\leq 1}_{p,q}$. This is done via the usual adjoint map given by $[P^*(F)](v, w) = F(P(v, w))$.

The image of P^* is given as the "diagonal" subalgebra:

$$\bigoplus_{s=0}^{\infty} \mathcal{A}_p^s \otimes \mathcal{A}_q^s.$$

From our point of view, the significance of this structure is as follows:

- The dimension is
$$\dim(\mathcal{A}_p^s \otimes \mathcal{A}_q^s) = \binom{s+p-1}{p-1}\binom{s+q-1}{q-1},$$
which is equal to the cardinality of $\mathcal{C}(s, p) \times \mathcal{C}(s, q)$. That is, a basis may be parameterized by a pair of compositions.
- The (finite dimensional) vector space $\mathcal{R}_{p,q}^s$ will have a dimension also equal to the above since P^* is an isomorphism.
- The bijection Φ from Proposition 5 establishes that the above dimension is given by the cardinality of $R(p, q; s)$.
- A basis for $\mathcal{R}_{p,q}^s$ may be given by the monomials of the form $\prod m_{ij}^{P_{ij}}$ where $P \in R(p, q; s)$.
- These monomials correspond to the monomials in $\mathcal{A}_p^s \otimes \mathcal{A}_q^s$ with exponents $(\mu, \nu) \in \mathcal{C}(s, p) \times \mathcal{C}(s, q)$.

4.4. ***The specialization and a derivative.*** From the definition it is relatively easy to see that

$$H_{p,q}(0, t) = \frac{1}{(1-t)^{\min(p,q)}}.$$

We next turn to the specialization $H_{p,q}(1, t)$, which turns out to be the Hilbert series of the rank at most 1 matrices. That is to say

$$H_{p,q}(1, t) = \sum_{s=0}^{\infty} (\dim \mathcal{R}_{p,q}^s) t^s.$$

In [5, Equation (6.4)] this series was computed as

$$H_{p,q}(1, t) = \frac{\sum_{i=0}^{\min(p-1,q-1)} \binom{p-1}{i}\binom{q-1}{i} t^i}{(1-t)^{p+q-1}}.$$

In this sense, $H(z, t)$ interpolates between the generating function for compositions and the Hilbert series of the determinantal varieties (at least in the rank one case). Moreover, for generic z we have a relationship to the EMD.

Our goal is to compute the expected value of EMD_s. Therefore, it is natural to compute the partial derivative of $H_{p,q}(z, t)$ with respect to z, and then set $z = 1$. From the definition of $H_{p,q}$,

$$\left.\frac{\partial H_{p,q}(z, t)}{\partial z}\right|_{z=1} = \sum_{s=0}^{\infty} \Bigg(\sum_{\substack{\mu \in \mathcal{C}(s,p) \\ \nu \in \mathcal{C}(s,q)}} \mathrm{EMD}_s(\mu, \nu) \Bigg) t^s.$$

To expand this we start with

$$H_{p,q}(z,t)=\frac{H_{p-1,q}+H_{p,q-1}-H_{p-1,q-1}}{1-z^{|p-q|}t},$$

the recursive relationship from Theorem 3. Then let the partial derivative of $H_{p,q}$ with respect to z be denoted $H'_{p,q}$. We find the derivative using the "quotient rule"

$$H'_{p,q}(z,t)=\frac{\partial}{\partial z}H_{p,q}(z,t)=$$
$$\frac{(H'_{p-1,q}+H'_{p,q-1}-H'_{p-1,q-1})(1-z^{|p-q|}t)+|p-q|z^{|p-q|-1}t(H_{p-1,q}+H_{p,q-1}-H_{p-1,q-1})}{(1-z^{|p-q|}t)^2}. \quad (4\text{-}1)$$

When $z=1$ this becomes

$$H'_{p,q}(1,t)=\frac{1}{(1-t)^2}\big((H'_{p-1,q}(1,t)+H'_{p,q-1}(1,t)-H'_{p-1,q-1}(1,t))(1-t)$$
$$+|p-q|t(H_{p-1,q}(1,t)+H_{p,q-1}(1,t)-H_{p-1,q-1}(1,t))\big). \quad (4\text{-}2)$$

Before proceeding it is useful to see some initial values:

$$H'_{1,1}=0, \qquad H'_{1,2}=\frac{t}{(1-t)^3}, \qquad H'_{1,3}=\frac{3t}{(1-t)^4},$$
$$H'_{2,1}=\frac{t}{(1-t)^3}, \quad H'_{2,2}=\frac{2t}{(1-t)^4}, \quad H'_{2,3}=\frac{t(3t+5)}{(1-t)^5},$$
$$H'_{3,1}=\frac{3t}{(1-t)^4}, \quad H'_{3,2}=\frac{t(3t+5)}{(1-t)^5}, \quad H'_{3,3}=\frac{8t(t+1)}{(1-t)^6}.$$

Both $H_{p,q}(1,t)$ and $H'_{p,q}(1,t)$ are rational functions. We anticipate that their numerators are

$$W_{p,q}(t):=(1-t)^{p+q-1}H_{p,q}(1,t) \quad \text{and} \quad N_{p,q}(t):=(1-t)^{p+q}H'_{p,q}(1,t).$$

Thus, multiplying by $(1-t)^{p+q}$ on both sides of (4-2) gives

$$N_{p,q}=\frac{1}{(1-t)^2}\big(\big((1-t)(N_{p-1,q}+N_{p,q-1})-(1-t)^2N_{p-1,q-1}\big)(1-t)$$
$$+|p-q|t\big((1-t)^2(W_{p-1,q}+W_{p,q-1})-(1-t)^3W_{p-1,q-1}\big)\big)$$

or

$$N_{p,q}=N_{p-1,q}+N_{p,q-1}-(1-t)N_{p-1,q-1}+|p-q|\,t\,(W_{p-1,q}+W_{p,q-1}-(1-t)W_{p-1,q-1})$$

If we also note that

$$W_{p,q}=W_{p-1,q}+W_{p,q-1}-(1-t)W_{p-1,q-1},$$

we ultimately obtain

$$N_{p,q}=N_{p-1,q}+N_{p,q-1}-(1-t)N_{p-1,q-1}+|p-q|\,t\,W_{p,q}. \quad (4\text{-}3)$$

An easy induction shows that both $W_{p,q}(t)$ and $N_{p,q}(t)$ are polynomials in t.

Before proceeding it is instructive to recall our goal of finding the expected value of EMD_s. In this light, define

$$\mathcal{N}(p, q; s) := \sum_{(\mu,\nu)\in\mathcal{C}(s,p)\times\mathcal{C}(s,q)} \mathrm{EMD}_s(\mu, \nu)$$

for $s \in \mathbb{N}$ and positive integers p and q. The expected value of EMD_s will be then obtained from

$$\lim_{s\to\infty} \frac{1}{s} \frac{\mathcal{N}(p, q; s)}{\binom{s+p-1}{p-1}\binom{s+q-1}{q-1}},$$

which we will show in the proof of Theorem 1 to be $\mathcal{M}_{p,q}$. In the next subsection we will use Proposition 6 to find the asymptotic value as $s \to \infty$ for fixed values of p and q. First we need more information about $\mathcal{N}(p, q; s)$.

Proposition 6. *Given positive integers p and q,*

$$\frac{N_{p,q}(t)}{(1-t)^{p+q}} = \sum_{s=0}^{\infty} \mathcal{N}(p, q; s)t^s$$

Proof. We have seen that

$$\left.\frac{\partial H}{\partial z}\right|_{z=1} = \frac{N_{p,q}(t)}{(1-t)^{p+q}}.$$

Differentiating the definition $H_{p,q}(z, t)$ term by term and then setting $z = 1$ gives the result. □

Using (4-3) and the formula for $W_{p,q}(t)$ one can efficiently compute $N_{p,q}(t)$ for specific values of p and q. Following the method for generating functions, we multiply $N_{p,q}(t)$ and the series expansion of $\frac{1}{(1-t)^{p+q}}$. Because the series expansion involves only binomial coefficients we are led to an efficient method for finding the expected value of EMD_s on $\mathcal{C}(s, p) \times \mathcal{C}(s, q)$ for any given values of p, q and s. Consequently we determine the values of $\mathcal{N}(p, q; s)$.

Some initial data for $N_{n,n}(t)$ for $n = 1, \ldots, 12$ are:

$$0$$
$$2t$$
$$8t(t+1)$$
$$4t(5t^2+14t+5)$$
$$8t(5t^3+27t^2+27t+5)$$
$$2t(35t^4+308t^3+594t^2+308t+35)$$
$$16t(7t^5+91t^4+286t^3+286t^2+91t+7)$$
$$8t(21t^6+378t^5+1755t^4+2860t^3+1755t^2+378t+21)$$
$$16t(15t^7+357t^6+2295t^5+5525t^4+5525t^3+2295t^2+357t+15)$$
$$2t(165t^8+5016t^7+42636t^6+142120t^5+209950t^4+142120t^3+42636t^2+5016t+165)$$
$$8t(55t^9+2079t^8+22572t^7+99484t^6+203490t^5+203490t^4+99484t^3+22572t^2+2079t+55)$$
$$4t(143t^{10}+6578t^9+88803t^8+499928t^7+1352078t^6+1872108t^5 + 1352078t^4+499928t^3+88803t^2+6578t+143)$$

The coefficients of the above are nonnegative integers, which we prove inductively. Furthermore, they apparently are *palindromic* – that is the coefficient of t^i matches the coefficient of t^{d-i} where d is the polynomial degree. Lastly we note that the coefficients rise in value until the middle and then decrease – that is to say they are *unimodal*. Polynomials with these properties are often of interest.

Conjecture 1. *The coefficients of the polynomials $N_{n,n}(t)$ are unimodal and palindromic.*

In Theorem 1, the $\mathbb{EMD}$ has been normalized so that μ and ν are probability distributions. In terms of pairs of compositions of s, this amounts to multiplying by $\frac{1}{s}$. Note that since the order of the pole at $t = 1$ in $H_{p,q}(1, t)$ is one less than the order of the pole at $t = 1$ in $H'(1, t)$ we see that the (normalized) EMD is approaching a constant as $s \to \infty$. Alternatively, we could choose not to normalize and then obtain a linear growth of $s\,\mathcal{M}_{p,q}$.

The following is a table with approximate values of $\mathcal{M}_{p,q}$ for $1 \leq p, q, \leq 5$,

p	$q=1$	2	3	4	5
1	0.000	0.500	1.000	1.500	2.000
2	0.500	0.333	0.667	1.100	1.567
3	1.000	0.667	0.533	0.800	1.190
4	1.500	1.100	0.800	0.686	0.914
5	2.000	1.567	1.190	0.914	0.813

We plot $\mathcal{M}_{p,q}$ for $1 \leq p, q \leq 12$ in Figure 1.

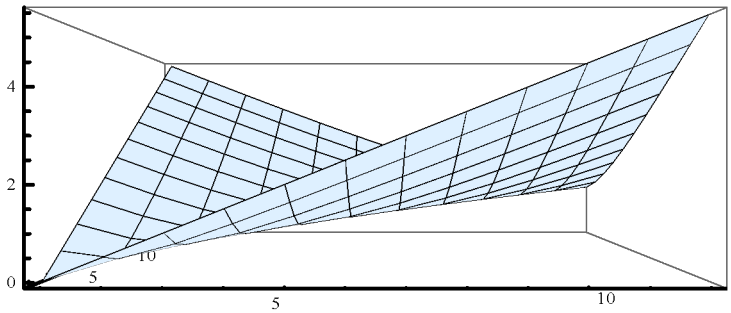


Figure 1. Plot of $\mathcal{M}_{p,q}$ for $1 \le p, q \le 12$.

5. Further calculations and proofs of the theorems

This section includes the main technical points of the paper, including the proofs of the main theorems. It is more convenient to prove Theorem 1 after Theorems 2 and 3.

First, however, we make explicit two useful scaled versions of the EMD.

5.1. ***Unit normalized earth mover's distance.*** Averaging $\mathcal{E}(\mu - \nu)/s$ over $(\mu, \nu) \in \mathcal{C}(s, n) \times \mathcal{C}(s, n)$ gives rise to the expected value of the discrete EMD. Taking the limit as $s \to \infty$ gives the expected value of the *normalized EMD* on $\mathcal{P}_n$. Observe that the maximum value of the normalized EMD on $\mathcal{P}_n$ is $n - 1$. The *unit normalized EMD* will be defined as the normalized EMD scaled by $\frac{1}{n-1}$. This scaling makes $\mathcal{P}_n$ into a metric space with diameter 1. When working with real data in the last section of this paper, we will always use the unit normalized earth mover's distance.

As an example, we consider the discrete case where $\mathcal{P}_n$ is replaced by $\mathcal{C}(s, n)$. In particular, choose $s = 30$ and $n = 5$ and calculate the exact histogram for the unit normalized distance. The mean of the distribution is obtained by expanding

$$\frac{8t\left(5t^3 + 27t^2 + 27t + 5\right)}{(1-t)^{10}}$$

as a series around $t = 0$, then taking the coefficient of t^{30} and dividing by $\binom{30+5-1}{5-1}^2$ (the number of ordered pairs of distributions). The approximate value is 26.2938.

For the unit normalized distance we divide by $s(n-1) = 30(5-1) = 120$. That is, we divide by s to obtain a probability distribution, and then divide by $n - 1$ to scale the diameter to 1. The unit normalized mean is approximately 0.219115 as shown in Figure 2.

In the limiting case as $s \to \infty$, the mean decreases slightly from the $s = 30$ case. Again, the scaling sets the diameter of the metric space $\mathcal{P}_n$ to 1. Thus, the expected value of the unit normalized EMD is

$$\widetilde{\mathcal{M}}_n := \frac{\mathcal{M}_n}{n-1}.$$

Recall the notation that $\mathcal{M}_n = \mathcal{M}_{n,n}$, for positive integer n.

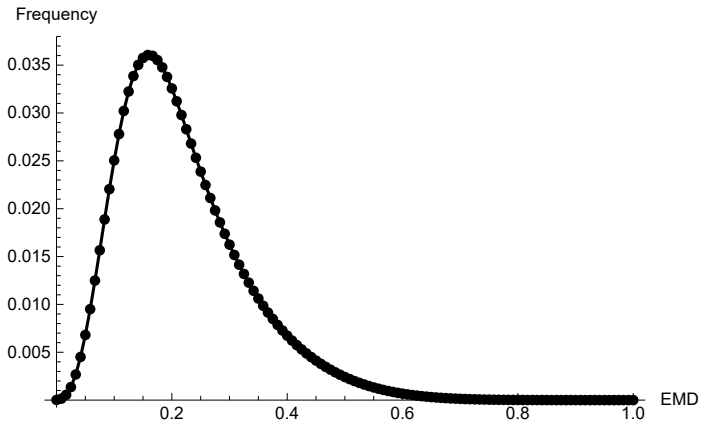


Figure 2. Histogram of the unit normalized EMD for $s = 30$, $n = 5$.

It should be noted that all values of $\widetilde{\mathcal{M}}_n$ are rational numbers. We present the following approximate values of $\widetilde{\mathcal{M}}_n$ for $n = 2, \ldots, 12$,

n	2	3	4	5	6	7	8	9	10	11	12
$\widetilde{\mathcal{M}}_n$	0.3333	0.2667	0.2286	0.2032	0.1847	0.1705	0.1591	0.1498	0.1419	0.1351	0.1293

which are the limiting values as $s \to \infty$. For finite choices of s and n we can compute the exact mean of the unit normalized EMD by expanding

$$\frac{N_{p,q}(t)}{(1-t)^{p+q}}$$

in the case when $p = q = n$, and then dividing by $s(n-1)\binom{s+n-1}{n-1}^2$. We show the approximate values in Table 1.

5.2. ***Proof of Theorem 2.*** By continuity it will suffice to show that this is true on a dense subset of $\mathcal{P}_n$. Specifically, we will consider the special case that if μ (resp. ν) is of the form

$$\mu = \left(\frac{a_1}{s}, \ldots, \frac{a_n}{s}\right)$$

for some positive integer s and $(a_1, \ldots, a_n) \in \mathcal{C}(s, n)$. As $s \to \infty$, such points are dense in $\mathcal{P}_n$

For $p \leq n$ (resp. $q \leq n$) we can regard $\mathcal{C}(s, p)$ (resp. $\mathcal{C}(s, q)$) as being embedded in $\mathcal{C}(s, n)$ by appending zeros onto the right.

By induction on $p + q$ we will show that for any nonnegative integer s,

$$\mathrm{EMD}_s(\mu, \nu) = \mathcal{E}(\mu - \nu)$$

for $\mu \in \mathcal{C}(s, p)$ and $\nu \in \mathcal{C}(s, q)$.

If $p = q = 1$ the result is trivial since there is only one composition of s. Consider $p + q \geq 2$. Without loss of generality assume $p \leq q$.

p	$n=2$	3	4	5	12
1	0.5000	0.4444	0.4167	0.4000	0.3611
2	0.4444	0.3889	0.3600	0.3422	0.2991
3	0.4167	0.3600	0.3300	0.3113	0.2649
4	0.4000	0.3422	0.3113	0.2918	0.2428
5	0.3889	0.3302	0.2985	0.2784	0.2272
10	0.3636	0.3020	0.2681	0.2462	0.1881
15	0.3542	0.2912	0.2561	0.2333	0.1716
20	0.3492	0.2854	0.2497	0.2264	0.1624
30	0.3441	0.2794	0.2430	0.2191	0.1524
60	0.3388	0.2732	0.2360	0.2114	0.1415
120	0.3361	0.2700	0.2323	0.2073	0.1355
180	0.3352	0.2689	0.2311	0.2060	0.1335
360	0.3343	0.2678	0.2298	0.2046	0.1314
500	0.3340	0.2675	0.2295	0.2042	0.1308
750	0.3338	0.2672	0.2292	0.2039	0.1303
1000	0.3337	0.2671	0.2290	0.2037	0.1300
1250	0.3336	0.2670	0.2289	0.2036	0.1299
1500	0.3336	0.2669	0.2289	0.2035	0.1298
2000	0.3335	0.2669	0.2288	0.2034	0.1296
10000	0.3334	0.2667	0.2286	0.2032	0.1293

Table 1. Mean of the unit normalized EMD.

We proceed by induction on s (inside the induction on $p+q$). If $s=0$ the statement is vacuous, and so the base case is clear.

For positive integer s let J be a p-by-q nonnegative integer matrix such that J has row and column sums μ and ν respectively and $\langle J, C\rangle$ is minimal. We shall show that $\langle J, C\rangle = \mathcal{E}(\mu-\nu)$.

If the first row (resp. column) of J is zero we can delete it and reduce to the inductive hypothesis on $p+q$. Therefore, we assume that there is a positive entry in the first row (resp. column) of J. If $J_{11}>0$ then we can subtract J_{11} from s and reduce to the inductive hypothesis on s.

We are therefore left with $J_{11}=0$ and the existence of $i>1$, $j>1$ with $J_{1j}>0$ and $J_{i1}>0$. However, $(1, j)$ and $(i, 1)$ are incomparable in the poset $[p]\times[q]$. However, by Proposition 4 we can assume that the support of J is a chain.

The cost for the joint distribution, J, is minimized when the support is a chain. Upon inspection, one sees that the value of $\mathcal{E}$ agrees with the cost in the case that the support of J is on a chain. The value of $J_{i,j}$ is multiply counted $|i-j|$ times – once for each of the contributing terms of $\mathcal{E}$. In the noncontributing terms there is a telescoping inside the absolute value.

Theorem 2 follows for chains, to which we have inductively reduced the problem. □

5.3. ***Proof of Theorem 3.*** The vector space of degree s homogeneous polynomial functions on the rank at most one p-by-q matrices is denoted $\mathcal{R}^s_{p,q}$. By Proposition 4, we obtain a basis for this space by

considering the monomials

$$\prod_{i=1}^{p}\prod_{j=1}^{q} x_{ij}^{J_{ij}}$$

where J is a nonnegative integer matrix with support on a chain. The row and column sums of J are a pair of compositions of s with p and q parts respectively. We denote these by μ and ν. Note that by Proposition 5, μ and ν determine J.

If we assign each of these monomials the formal expression $z^{\mathrm{EMD}_s(\mu,\nu)}t^s$ and sum them as formal series, we obtain the Hilbert series of $\mathcal{R}_{p,q}$,

$$\sum_{s=0}^{\infty}\left(\sum_{(u,v)\in\mathcal{C}(s,p)\times\mathcal{C}(s,q)} z^{\mathrm{EMD}_s(u,v)}\right)t^s$$

which we then recognize as the definition of $H_{p,q}(z,t)$.

Each monomial has a nonnegative integer matrix J as its exponents, with support on a chain. This chain terminates at or before $x_{p,q}^{J_{p,q}}N$. From Theorem 2 the variable $x_{p,q}$ is multiplied by $z^{|p-q|}t$, and contributes

$$\sum_{J_{p,q}=0}^{\infty}\left(z^{|p-q|}t\right)^{J_{p,q}}$$

to all monomials. The geometric series sums to $\dfrac{1}{1-z^{|p-q|}t}$.

The preceding variables in the monomial may contain $x_{p,j}$ for some $1\le j\le q$, or $x_{i,q}$ for some $1\le i\le p$, but not both – since the exponent matrix has support in a chain. In the former case these monomials are in the sum $H_{p,q-1}$, while in the latter are counted in $H_{p,q-1}$.

The sum $H_{p-1,q}+H_{p,q-1}$ over counts monomials. That is to say, if a monomial has an exponent with support involving variables $x_{i,j}$ with $i<p$ and $j<q$ then it is counted once in $H_{p-1,q}$ and once in $H_{p,q-1}$. It also appears once in $H_{p-1,q-1}$. We therefore observe that such monomials are counted exactly once in the expression

$$H_{p-1,q}+H_{p,q-1}-H_{p-1,q-1}.$$

Finally, we see that all such monomials are counted exactly once in the product

$$\frac{H_{p-1,q}+H_{p,q-1}-H_{p-1,q-1}}{1-z^{|p-q|}t}$$

if $(p,q)\neq(1,1)$ and $H_{1,1}=\frac{1}{1-t}$. □

5.4. ***Proof of Theorem 1.*** Fix positive integers p and q. The coefficient of t^s in $\frac{1}{(1-t)^{p+q}}$ is

$$\binom{s+p+q-1}{p+q-1}=\frac{s^{p+q-1}}{(p+q-1)!}+\text{ lower order terms in } s.$$

And, we have

$$N_{p,q}(t)=c_0+c_1t+c_2t^2+\cdots+c_kt^k$$

for some nonnegative integers $c_0, \ldots, c_k$. Thus the coefficient of t^s in

$$\frac{N_{p,q}(t)}{(1-t)^{p+q}}$$

is therefore asymptotic to

$$N_{p,q}(1)\frac{s^{p+q-1}}{(p+q-1)!}$$

We will next find an inhomogeneous three term recursive formula for $N_{p,q}(1)$ from (4-3). First we observe that $W_{p,q}(1) = \binom{p+q-2}{p-1}$. Therefore, we have

$$N_{p,q}(1) = N_{p-1,q} + N_{p,q-1} + |p-q|\binom{p+q-2}{p-1}.$$

Then we divide by $(p+q-1)!$ to obtain the asymptotic. However, our goal is to obtain the expected value of EMD_s. So, in light of Proposition 6, we will need to divide by

$$\binom{s+p-1}{p-1}\binom{s+q-1}{q-1} \sim \frac{s^{p+q-2}}{(p-1)!(q-1)!}.$$

Thus, we find that the expected value is

$$\mathcal{M}_{p,q} = \frac{(p-1)!(q-1)!}{(p+q-1)!}N_{p,q}(1),$$

which we can rewrite as

$$\mathcal{M}_{p,q} = \frac{(p-1)\mathcal{M}_{p-1,q} + (q-1)\mathcal{M}_{p,q-1} + |p-q|}{p+q-1}.$$ □

6. Relation to spectral graph theory

We next turn our attention to some results from spectral graph theory and describe a connection with the expected value of the EMD.

The concept of a graph (or network) is likely familiar to the reader. We recall the terminology briefly. By a *graph* we mean an ordered pair, (V, E), where V is a finite set whose elements are called *vertices* and E is a finite set whose elements are called *edges*, together with an injective mapping from E to unordered pairs of distinct vertices. The elements of E are said to *join* the corresponding pair of vertices. The number of vertices joined to a given vertex, $v \in V$, is called the degree of v, denoted $deg(v)$.

A sequence of distinct vertices, $v_1, v_2, \ldots, v_t$ with v_i joined to v_{i+1} for each i is called a *path*. If a path exists between all pairs of vertices then we say that the graph is *connected*.

Given vertices v and w in a connected graph, the *distance* between vertex v and vertex w is the length of the shortest path starting with v and ending with w, and will be denoted $\rho(v, w)$. The function ρ is a metric on V. For an integer r, the *ball* of radius r centered at $v \in V$ will be defined as

$$B_r(v) := \{w \in V : \rho(v, w) \leq r\}.$$

Furthermore, let $n_r(v) = |B_r(v)| - |B_{r-1}(v)|$, and set

$$S(v) := \sum_{r \geq 0} r n_r(v),$$

which is a finite sum giving the expected distance a vertex is from v.

As we shall see, this metric is related to the topic of this article. Specifically, the *mean distance* in a graph is defined to be:

$$\overline{\rho}(G) := \frac{1}{m(m-1)} \sum_{v \in V} S(v).$$

where $m = |V|$.

The above notation is from [12], Section 3. Observe that the definition is equivalent to averaging the distance between all two-element subsets of V. We also point out that if we consider all ordered pairs of vertices we have:

$$\left(1 - \frac{1}{m}\right) \overline{\rho}(G) = \frac{1}{m^2} \sum_{(v,w) \in V \times V} \rho(v, w).$$

From our point of view, we consider the graph to be on the vertex set $\mathcal{C}(s, n)$ where two compositions are joined when the (unnormalized) earth mover's distance is exactly 1. We call this graph the *earth mover's graph*, denoted $G(s, n)$. In this case the length of the shortest path between two vertices is the earth mover's distance.

The average distance between ordered pairs of vertices in $G(s, n)$ is the subject of this article. More generally, one can consider a graph G whose vertices are a subset of $\mathcal{C}(s, n)$ and two vertices v and w are joined when $\mathrm{EMD}_s(v, w) \leq t$ for some fixed constant $t \geq 0$. We will call t the *threshold*. In practice, determining values of the threshold that uncover features in the data is an important research topic. Understanding the expected EMD in the uniform case is only one line of research.

The connected components of G are of interest in cluster analysis. For example, the connected components of G may be interpreted as clusters. If $t = 0$ there are no edges in G, thus no connected components. For sufficiently large t, every pair of vertices is joined and there is only one component. As t decreases the graph disconnects. Hierarchical connection of components defined by t gives rise to many different types of clustering. An example of this type of analysis will be given in the next section.

There are several "off the shelf" methods for clustering analysis. See [10] Chapter 20 for some commentary, especially about the popular "k-means" algorithm. Here we present only *metric hierarchical clustering* and *spectral clustering* as they relate to Theorem 1. However, the data that we consider in this article can and should be looked at from several points of view. In particular, unsupervised machine learning techniques are merited. See [8] as a general reference.

The average distance of a graph is also related to other invariants; we recommend the survey [13]. First we recall some additional terminology. Given a graph G with vertices $\{v_1, \ldots, v_m\}$ and k edges, one can form the Laplacian matrix $L_G = D_G - A_G$ where D_G is the diagonal matrix with the degree of vertex v_i in the i-th row and i-th column, while $A(G)$ is the adjacency matrix in which the entry in row i and column j is a one if v_i and v_j are joined by an edge, and zero otherwise.

The spectrum of L_G is of interest. To begin, L_G is a positive semidefinite matrix. The multiplicity of the 0-eigenspace is equal to the number of connected components of G. If the spectrum of L_G is denoted by $0 = \lambda_1 \leq \lambda_2 \leq \cdots \leq \lambda_m$, the *algebraic connectivity* is given by λ_2. Intuitively, we expect "clustering" when λ_2 is small relative to the rest of the spectrum.

Related to the algebraic connectivity are inequalities proved in [12]. We recall them here because they partially describe the structure of the graph based on λ_2. In fact we can use λ_2 to calculate bounds on the average distance between vertices in a graph and another invariant to be defined next.

The discrete Cheeger inequality asserts that the *isoperimetric number*, $i(G)$, is closely related to the spectrum of a graph. This number is defined as

$$i(G) := \min\left\{\frac{|\delta X|}{|X|} : X \subseteq V(G) \text{s.t.} 0 < |X| < \frac{1}{2}|V(G)|\right\}$$

where $\delta(X)$ is defined to be the boundary of a set of vertices X (that is $v \in X$ iff v is in X but is joined to a vertex not in X). One result from [12] is

$$\frac{\lambda_2}{2} \leq i(G) \leq \sqrt{\lambda_2(2d_{max} - \lambda_2)} \tag{6-1}$$

where d_{max} is the maximum degree of a vertex in G. These results have their underpinnings in geometry and topology, see [11], for example. Intuitively, the point here is that if G has two large subgraphs that are joined only by a small set of edges then $i(G)$ is small. Unfortunately, computing $i(G)$ exactly is difficult. However, the spectrum of G can be computed more easily, providing the stated bounds for $i(G)$.

A third invariant for G is $\overline{\rho}(G)$, the mean distance. This value can also be bounded. The inequality presented in [12] is

$$\frac{1}{m-1}\left(\frac{2}{\lambda_2} + \frac{m-2}{2}\right) \leq \overline{\rho}(G) \leq \frac{m}{m-1}\left[\frac{d_{max} - \lambda_2}{4\lambda_2}\ln(m-1)\right] \tag{6-2}$$

where m is the number of vertices in G.

7. Real world data

In this section we consider a real world data set coming from the Section Attrition and Grade Report published by the Office of Assessment and Institutional Research at the University of Wisconsin - Milwaukee for Fall semesters of academic years 2013-2017. We selected data for courses with enrollments greater than 1,000. Analysis is done for the 12 grades A through F (with plus/minus grading). "W" grades were not reported by the University for the entire period and are not included.

Most universities collect and analyze similar types of retention and attrition data. Analysis of grade distribution data may be an as-yet untapped portal for self-examination that reveals patterns in large, multi-section courses, or allows for insight into cross-divisional course boundaries. Cluster analysis provides a visual reinforcement of the calculated EMD.

The data for 2013–2017, shown in Table 2, comprise a total of twenty-one courses. In Fall 2013 there are five courses: English 101 and 102, Math 095 and 105, and Psychology 101. Math 095 was redesigned

d	division	course	year	enrollment	id	division	course	year	enrollment
1	English	101	2013	1800	11	Math	095	2013	1166
2	English	102	2013	1224	12	Math	105	2013	1555
3	English	101	2014	1762	13	Math	105	2014	1701
4	English	102	2014	1299	14	Math	105	2015	1466
5	English	101	2015	1742	15	Math	105	2016	1604
6	English	102	2015	1525	16	Math	105	2017	1732
7	English	101	2016	1693	17	Psychology	101	2013	1507
8	English	102	2016	1410	18	Psychology	101	2014	1443
9	English	101	2017	1569	19	Psychology	101	2015	1337
10	English	102	2017	1142	20	Psychology	101	2016	1192
					21	Psychology	101	2017	1333

Table 2. UWM grade data.

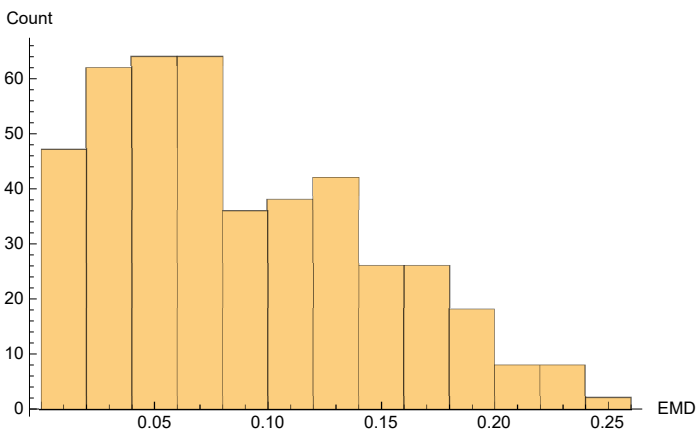


Figure 3. Histogram of EMD for UWM grade data.

after Fall 2013 and enrollment dropped below 1,000; the other four courses were offered every year. Input data are sorted by division and year.

7.1. ***Histogram of EMD sample.*** We form the earth mover's graph by computing the unit normalized EMD for each pair of the 21 courses. A histogram of the results is presented in Figure 3. The rough structure reflects some aspects of the distribution of the theoretical case for 30 students and 5 grades depicted in Figure 2. For example, it is almost unimodal and skewed to the right. It is also interesting to note that the histogram in Figure 3 shows maximal EMD between courses at around 0.25. Additionally, for each course one can compute the average EMD to all others. This course pairwise average has a minimum EMD of 0.067, mean 0.086, and maximum 0.129.

To better understand the relative sizes of these numbers, we can compare them with the mean in the uniform model with $n = 12$, which is approximately 0.1300 when $s = 1000$ and then drops to 0.1293 when $s \to \infty$. For actual grade distribution data, not all grade distributions are equally likely. Nonetheless,

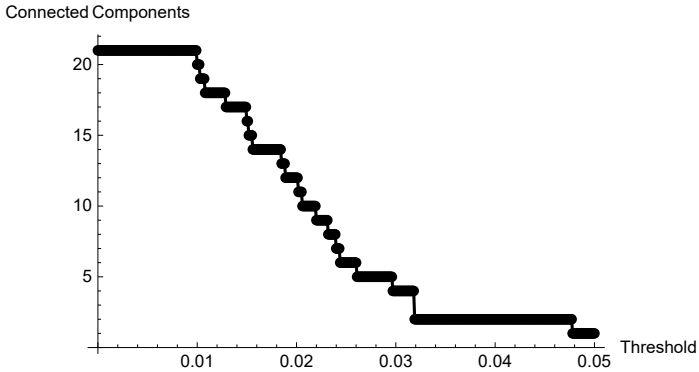


Figure 4. Number of connected components with change in threshold.

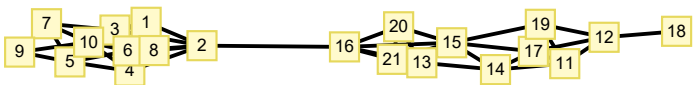


Figure 5. Connected components.

for this data set the maximum mean EMD corresponds to Psychology 101, Fall 2014 and is surprisingly close to the theoretical value. No doubt this is an anomaly, but we were surprised by how large the unit normalized EMD was between courses when compared to uniform sampling.

7.2. ***Hierarchical Clustering.*** Using the unit normalized EMD, we next determine a threshold, t, such that two courses are joined by an edge when the EMD falls below t. Thus, we obtain a family of graphs parameterized by t. We consider the connected component structure for each value of t. In practice, one sees a single giant component when the threshold is "large". As a heuristic, we consider large to mean greater than the expected distance in the uniform model.

In Figure 4 we step through various distance threshold values t and count the number of connected components. When t is very small, the graph has 21 components. As t increases, the components of the graph continue to connect, with the largest range of persistence for $t \in [0.0319, 0.0477]$ with two components. It is useful to compare these values to the expected EMD in the uniform model. The endpoints of this interval are 25% and 37% of the uniform mean, respectively.

Set $t = 0.0478$. We form a graph on the vertex set of these 21 courses. Two courses are joined by an edge when the unit normalized distance between them falls below t. Since t is larger than 0.0477, there is only one connected component as depicted in Figure 5.

7.3. ***Spectral Analysis.*** A key component of spectral clustering analysis is the construction of the Laplacian Matrix L_G as defined in Section 6. The second smallest eigenvalue of L_G gives the algebraic connectivity. Figure 6 presents a plot of the full spectrum for $t = 0.0478$.

For the 21 vertex graph in Figure 5, the algebraic connectivity is $\lambda_2 \cong 0.1213$, and the maximum degree is $d_{max} = 8$. It appears as if elements 2 and 16 create a tenuous bridge between two "clusters".

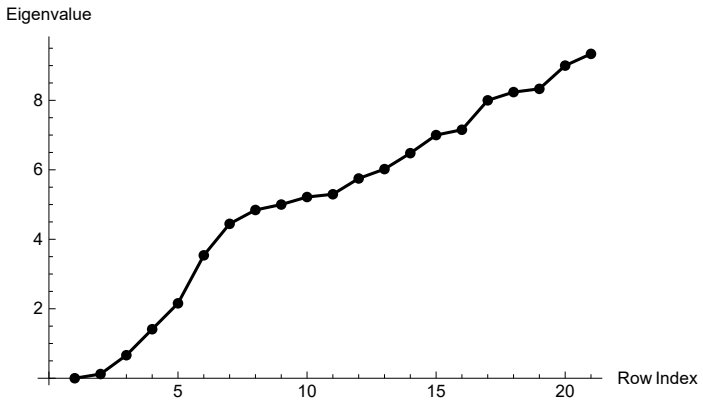


Figure 6. Eigenvalues for $t = 0.0478$.

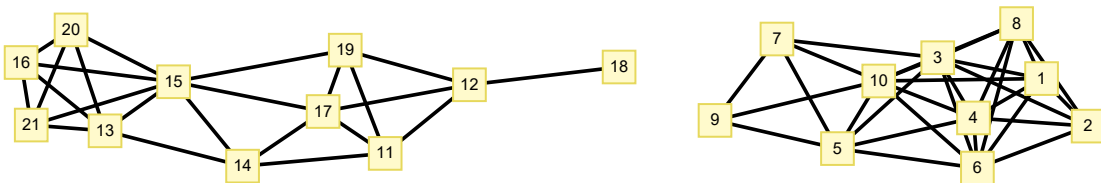


Figure 7. Partitioned components.

The theoretical bounds on the isoperimetric number from (6-1) are

$$\frac{0.1213}{2} \leq i(G) \leq \sqrt{0.1213(2 \times 8 - 0.1213)}$$

$$0.06065 \leq i(G) \leq 1.3878$$

with a computed value of $i(G) \cong 0.1$ for this data set. The relatively small isoperimetric number is consistent with the single edge of connection between elements 2 and 16 in Figure 5.

The theoretical bounds on the mean distance $\overline{\rho}(G)$ from (6-2) are

$$\frac{1}{21-1}\left(\frac{2}{0.1213} + \frac{21-2}{2}\right) \leq \overline{\rho}(G) \leq \frac{21}{21-1}\left[\frac{8-0.1213}{4 \times 0.1213}\ln(21-1)\right]$$

$$1.2995 \leq \overline{\rho}(G) \leq 51.08$$

and one can compute $\overline{\rho}(G) \cong 2.910$ for this data set.

The most curious part of this analysis appears in the plot of the components. At threshold $t = 0.0477$ the algebraic connectivity λ_2 goes to zero, and the EMD separates all of the English courses from the cluster of Math and Psychology courses. More specifically, EMD splits English (Course ID's 1 through 10) off from Math (ID's 11 - 16) and Psychology (ID's 17-21). Or equivalently, grade distributions in English courses are most similar to grade distributions in other English courses, and least similar to grade distributions in both Math and Psychology courses.

Although we see this clear partition of the data in Figure 7, one needs to be cautious about clustering algorithms. For example, clustering identified in one algorithm may not be the same as another. See the paper [9] for a careful treatment of this topic in general. In our more specific setting the clustering is determined by the threshold t. Thus, the question of where to set this value is delicate: too small and we see too many components, while too large we see very little clustering at all. Furthermore, our intention is only for exploratory data analysis and not to suggest policy regarding instructional assessment.

Acknowledgments

Willenbring thanks Anthony Gamst for conversations concerning cluster analysis (and especially the reference [9]), and Ryan (Skip) Garibaldi for pointing out the relationship between the earth mover's distance and comparing grade distributions.

Both authors thank the referees for making substantial improvements to the mathematical content and exposition of this article.

References

[1] F. R. K. Chung, *Spectral graph theory*, CBMS Regional Conference Series in Mathematics **92**, Amer. Math. Soc., 1997.

[2] D. Cox, J. Little, and D. O'Shea, *Using algebraic geometry*, Graduate Texts in Mathematics **185**, Springer, 1998.

[3] J. A. De Loera and E. D. Kim, "Combinatorics and geometry of transportation polytopes: an update", pp. 37–76 in *Discrete geometry and algebraic combinatorics*, Contemp. Math. **625**, Amer. Math. Soc., 2014.

[4] T. J. Enright and W. A. Hunziker, Markus Pruett, "Diagrams of Hermitian type, highest weight modules, and syzygies of determinantal varieties", pp. 121–184 in *Symmetry: representation theory and its applications*, Progr. Math. **257**, Birkhäuser, 2014.

[5] T. J. Enright and J. F. Willenbring, "Hilbert series, Howe duality, and branching rules", *Proc. Natl. Acad. Sci. USA* **100**:2 (2003), 434–437.

[6] W. Fulton, *Young tableaux, with applications to representation theory and geometry*, London Mathematical Society Student Texts **35**, Cambridge University Press, 1997.

[7] R. Goodman and N. R. Wallach, *Symmetry, representations, and invariants*, Graduate Texts in Mathematics **255**, Springer, 2009.

[8] T. Hastie, R. Tibshirani, and J. Friedman, *The elements of statistical learning: data mining, inference, and prediction*, Springer, 2001.

[9] J. M. Kleinberg, "An impossibility theorem for clustering", pp. 463–470 in *Advances in neural information processing systems*, edited by S. Becker et al., NIPS **15**, MIT Press, Cambridge, MA, 2003.

[10] D. J. C. MacKay, *Information theory, inference and learning algorithms*, Cambridge University Press, 2003.

[11] F. Mémoli, "A spectral notion of Gromov–Wasserstein distance and related methods", *Appl. Comput. Harmon. Anal.* **30**:3 (2011), 363–401.

[12] B. Mohar, "Eigenvalues, diameter, and mean distance in graphs", *Graphs Combin.* **7**:1 (1991), 53–64.

[13] B. Mohar, "The Laplacian spectrum of graphs", pp. 871–898 in *Graph theory, combinatorics, and applications* (Kalamazoo, MI, 1988), vol. 2, Wiley, 1991.

[14] E. Perrone, L. Solus, and C. Uhler, "Geometry of discrete copulas", *J. Multivariate Anal.* **172** (2019), 162–179.

Received 2019-05-03. Revised 2019-10-17. Accepted 2019-11-05.

REBECCA BOURN: bourn@uwm.edu
Department of Mathematical Sciences, University of Wisconsin - Milwaukee, 3200 N. Cramer St., Milwaukee, WI 53211, United States

JEB F. WILLENBRING: jw@uwm.edu
Department of Mathematical Sciences, University of Wisconsin - Milwaukee, 3200 North Cramer Street, Milwaukee, WI 53211, United States

Algebraic Statistics
Vol. 11, No. 1, 2020
https://doi.org/10.2140/astat.2020.11.79

INFERRING PROPERTIES OF PROBABILITY KERNELS FROM THE PAIRS OF VARIABLES THEY INVOLVE

LUIGI BURIGANA AND MICHELE VICOVARO

A probabilistic model may involve families of probability functions such that the functions in a family act on a definite (possibly multiple) variable and are indexed by the values of some other (possibly multiple) variable. "Probability kernel" is the term here adopted for referring to any one such family. This study highlights general properties of probability kernels that may be inferred from set-theoretic characteristics of the pairs of variables on which the kernels are defined. In particular, it is shown that any complete set of such pairs of variables has the algebraic form of a lattice, which is then inherited by any complete set of compatible kernels defined on those pairs; that on pairs of variables a criterion may be applied for testing whether corresponding probability kernels are compatible with one another and may thus be the building blocks of a consistent probabilistic model; and that the order between pairs of variables within their lattice provides a general diagnostic about deducibility relations between probability kernels. These results especially relate to models that involve a number of random variables and several interrelated conditional distributions acting on them; for example, hierarchical Bayesian models and graphical models in statistics, Bayesian networks and Markov fields, and Bayesian models in the experimental sciences.

1. Introduction

A general way of expressing the (possible) probabilistic dependence of a random variable Y on another random variable X is in terms of a family of probability distributions for Y that are conditional on the distinct possible values of X. If we denote by X° and Y° the spaces of the two variables — or, more concretely, the sets of their possible values — then such a family may formally be indicated as $(p(Y|x) : x \in X^\circ)$, where $p(Y|x)$ (for any $x \in X^\circ$) is the probability function on the domain Y° that would rule the variable Y under the condition $X = x$. If there are differences within the family — that is, if $p(Y|x) \neq p(Y|x')$ for some $x \neq x'$ belonging to X° — then there is some form of stochastic dependence of Y on X. Otherwise, the two variables are stochastically independent of each other. In this article, we refer to such a family of conditional probability functions by the name *probability kernel* and by the symbol $p(Y|X)$ — so that $p(Y|X)$ is the same as $(p(Y|x) : x \in X^\circ)$ by definition. The terms X and Y may be *multiple* variables and are here conceived as *disjoint* subsets of a *full* variable T, this being the collection of all elementary random variables involved in a probabilistic model. Typically, the probability functions constituting one kernel $p(Y|X)$ are of the same measure-theoretic type; for example, they may

Burigana is the corresponding author.
Keywords: probability kernel, conditional probability, compatibility, lattice, Bayesian model.

be either all mass functions ($p(Y|X)$ would be a kernel of discrete type) or all density functions ($p(Y|X)$ would be a kernel of continuous type)[1].

A probabilistic model with a suitably large set T of elementary variables may involve several such probability kernels $p(Y|X)$, $p(V|U)$, $p(Z|W)$, ... that concern (elementary or multiple) variables included in T. The two variables in any single kernel are assumed disjoint, but variables involved in different kernels may have non-empty intersections, which themselves count as variables included in T. The importance of any single kernel may be due either to its being a postulate in a given model, that is, a basic assumption characterizing the probabilistic dependence between relevant variables (or specifying the kind of distributions of those variables), or to its being a logical consequence of the postulates, which may be crucial for the interpretation of the model and its fitting to empirical data. Models involving several interrelated kernels may be generally referred to as *conditionally specified* probabilistic structures, for marking the central role played by assumptions of conditional distributions in the definition of such models (Arnold, Castillo, & Sarabia, 1999). Examples can be found in various areas of applied probability, such as graphical models in statistics (Koller & Friedman, 2009), hierarchical Bayesian modeling (Gelman et al., 2014, chapter 5), Bayesian networks in artificial intelligence (Darwiche, 2009), applications of Markov random fields (Blake, Kohli, & Rother, 2011), and Bayesian modeling in experimental sciences (Kersten, Mamassian, & Yuille, 2004; Rouder, Morey, & Pratte, 2017).

Any probability kernel $p(Y|X)$ has a definite *variable pair* $(Y|X)$ as its field of action, that is, an ordered pair formed of a conditioned variable Y (on the left of the bar) and a conditioning variable X (on the right of the bar) which are, in general, disjoint sub-variables of a full variable T. Variable pairs underlying distinct probability kernels may be subjected to comparisons, combinations, or transformations in mere set-theoretic terms, that is, as pairs of sets of elementary variables, irrespectively of the individual properties of the elementary variables they collect. Such set-theoretic manipulations may imply algebraic regularities worthy of note, and one may conjecture that these regularities concerning the variable pairs have meaningful consequences regarding the probability kernels acting on the variable pairs themselves. The aim of this article is precisely that of highlighting some general properties of probability kernels that may be inferred from the set-theoretic configuration of the variable pairs on which the kernels are defined.

In order to illustrate the association between actions on probability kernels and actions on the underlying variable pairs, let us refer to Figure 1, which represents a basic case in the theory of conditional probabilities. For simplicity, we suppose that X, W, and Z are elementary variables of discrete type, but the example is easily generalizable to multiple and/or continuous variables. Three basic operations on kernels are illustrated by the figure.

The first is *projection*. For example, one may pass from $p(W, Z|X)$ to $p(Z|X)$ by setting $p(z|x) = \sum_{w \in W^\circ} p(w, z|x)$ for all $(x, z) \in (X, Z)^\circ$ (this means: $x \in X^\circ$ and $z \in Z^\circ$).

[1] Our use of the word "kernel" is consistent with the meaning taken by this word in the theories of Markov processes and other probabilistic structures extensively involving conditional probability distributions (Meyn & Tweedie, 1993, p. 65; Lauritzen, 1996, p. 46). Indeed, the concept of a probability kernel, in such theories, generalizes the concept of a transition matrix in finite-state Markov chains, which amounts to an indexed set of conditional mass functions.

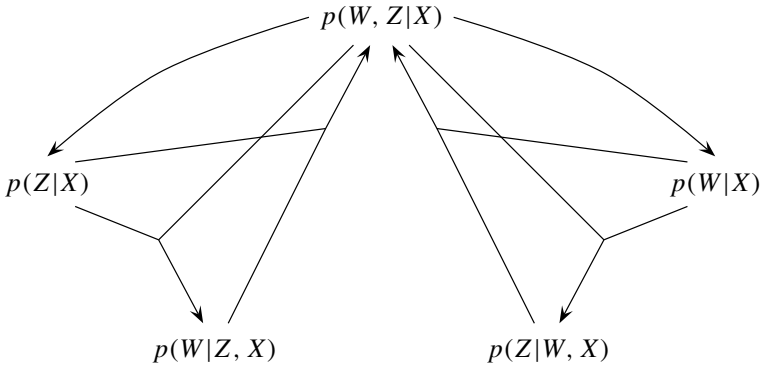


Figure 1. Probability kernels derivable from a top kernel $p(W, Z|X)$ through projection or conditioning.

The second is *conditioning*. For example, one may pass from $p(W, Z|X)$ to $p(W|Z, X)$ by setting $p(w|z, x) = p(w, z|x)/p(z|x)$ for all $(x, w, z) \in (X, W, Z)^\circ$, on presuming $p(z|x) > 0$ for all (x, z) in $(X, Z)^\circ$.

The third operation is *promotion*. For example, one may return from $p(W|Z, X)$ and $p(Z|X)$ to $p(W, Z|X)$ by setting $p(w, z|x) = p(w|z, x) \cdot p(z|x)$ for all $(x, w, z) \in (X, W, Z)^\circ$.

In this paper, the letters J, C, and M are used to denote proJection, Conditioning, and proMotion, respectively, so that the three moves just described may be symbolized as follows:

$$\begin{aligned} p(Z|X) &= J[p(W, Z|X), W] \\ p(W|Z, X) &= C[p(W, Z|X), Z] \\ p(W, Z|X) &= M[p(W|Z, X), p(Z|X)]. \end{aligned} \tag{1}$$

Of this example, what mostly matters for the aims of our study are the effects of the three operations on the variable pairs involved: projection J implies canceling a targeted component W from the left field (to the left of the bar), conditioning C implies moving a component Z from the left to the right field, and promotion M implies moving a component Z from the right to the left field of $p(W|Z, X)$ with the aid of a "promoter" $p(Z|X)$. We shall see that, based on these simple moves affecting the assignment of the variables to the left or the right fields in the probability kernels, algebraic constructs can be elaborated that have meaningful implications for the kernels at hand.

Our paper is formed of three main sections. Section 2 focuses on variable pairs and set-theoretic operations on them, and defines a binary relation that organizes any complete collection of such pairs as a lattice. In Section 3 the results obtained for variable pairs are transferred to probability kernels, and conditions are discussed that make it possible to ascend from kernels of low rank to kernels of higher rank in a lattice, which is a typical move in the construction of probabilistic models. In Section 4 the key binary relation between variable pairs will be shown to possess a general diagnostic ability about the deducibility relation between probability kernels.

2. Lattice of variable pairs

Let $T = \{T_1, \ldots, T_n\}$ be the complete set of elementary random quantities (observables, parameters, hyper-parameters, etc.) involved in a probabilistic model. By a *variable* we mean any subset of T. Thus, T itself is a variable, referred to as the *full variable* in the assumed model. Each singleton $\{T_i\}$ in T is an *elementary variable*. The symbol $\varnothing$ denotes the *empty variable*, which is the empty subset of T. Because variables are here understood as sets, statements such as "X is a sub-variable of Y" and "X and Y are disjoint variables", and formulas such as $X \subseteq Y$ and $X \cap Y = \varnothing$, are legitimate and meaningful in the language adopted in this paper.

A *variable pair* is any ordered pair $(Y|X)$ such that $X \cup Y \subseteq T$, $X \cap Y = \varnothing$, and $Y \neq \varnothing$, and the symbol $O(T)$ here denotes the complete collection of such pairs. Thus, if n is the cardinality of T, then $3^n - 2^n$ is the cardinality of $O(T)$. Besides, the symbol $\perp$ and the name *null variable pair* are here used for referring to any pair $(\varnothing|X)$ with $X \subseteq T$, and the symbol $\widetilde{O}(T)$ is a substitute for $O(T) \cup \{\perp\}$.

Drawing on equations (1), we define *projection* J, *conditioning* C, and *promotion* M on variable pairs by setting, for all $(Y|X) \in O(T)$:

$$J[(Y|X), W] = (Y \setminus W|X) \text{ for all } W \subset Y \tag{2}$$

$$C[(Y|X), W] = (Y \setminus W|W \cup X) \text{ for all } W \subset Y \tag{3}$$

$$M[(Y|X), (V|U)] = (Y \cup V|U) \text{ for all } (V|U) \in O(T) \text{ such that } V \cup U = X. \tag{4}$$

These definitions are designed so that reference to the null term $\perp$ is avoided. This limitation, however, can consistently be overcome by setting

$$\begin{aligned} &J[(Y|X), Y] = \perp, \quad C[(Y|X), Y] = \perp, \\ &M[\perp, (V|U)] = (V|U), \quad M[(Y|X), \perp] = (Y|X). \end{aligned} \tag{5}$$

Note, in particular, that the two additional equations concerning the M operation turn out to be consistent with rule (4) if $\perp$ becomes replaced by $(\varnothing|V \cup U)$ in the one equation and by $(\varnothing|X)$ in the other.

The equations in the next composite statement are self-evident:

$$\text{for all } (Y|X) \in O(T) \text{ and all } W, Z \subseteq Y \text{ such that } W \cap Z = \varnothing,$$

$$J[J[(Y|X), W], Z] = (Y \setminus (W \cup Z)|X) = J[J[(Y|X), Z], W] \tag{6}$$

$$C[C[(Y|X), W], Z] = (Y \setminus (W \cup Z)|W \cup Z \cup X) = C[C[(Y|X), Z], W] \tag{7}$$

$$J[C[(Y|X), W], Z] = (Y \setminus (W \cup Z)|W \cup X) = C[J[(Y|X), Z], W]. \tag{8}$$

They express invariance to change in order (commutativity) for combined J and C operations. In describing the result of the M operation, the rule is followed of writing the "to be promoted" variable pair $(Y|X)$ on the left, and its chosen "promoter" $(V|U)$ on the right, so that $X = V \cup U$. Therefore, if $M[(Y|X), (V|U)]$ is a syntactically correct formula, then $M[(V|U), (Y|X)]$ cannot be syntactically correct, because $U \neq Y \cup X$. For this simple reason, the M operation is not commutative. It is, however,

associative, because

$$\begin{aligned}&\text{for all } (Y|X), (V|U), (Z|W) \in O(T) \\ &\text{if } X = V \cup U \text{ and } U = Z \cup W, \\ &\text{then } M[M[(Y|X), (V|U)], (Z|W)] = (Y \cup V \cup Z|W) = M[(Y|X), M[(V|U), (Z|W)]].\end{aligned} \tag{9}$$

Furthermore, the following equations are easily proved:

$$C[M[(Y|X), (V|U)], V] = (Y|X) \tag{10}$$

$$J[M[(Y|X), (V|U)], Y] = (V|U). \tag{11}$$

These show how the operands of the M operation may be recovered from its result by suitably applying the C and J operations.

By combining the J and C operations, a binary relation between variable pairs is now defined, which is a key component of the theory in this study.

Definition 1. Let $(V|U)$ and $(Y|X)$ be variable pairs in $O(T)$. The former is *JC-derivable* from the latter (notation $(V|U) \preceq (Y|X)$) if $(V|U) = J[C[(Y|X), W], Z]$ for some W and Z disjoint sub-variables of Y. Furthermore, $\perp \preceq (Y|X)$ for all $(Y|X)$ in $O(T)$.

In other words, a variable pair is said to be JC-derivable from another variable pair if the former may be obtained from the latter through a conditioning followed by a projection — or equivalently, on account of (8), through a projection followed by a conditioning. Note that if $(V|U) \preceq (Y|X)$, then the variables W and Z satisfying the equation in Definition 1 are uniquely determined by

$$W = U \setminus X \text{ and } Z = Y \setminus (V \cup U).$$

Also note this bi-conditional

$$(V|U) \preceq (Y|X) \text{ if and only if } V \cup U \subseteq Y \cup X \text{ and } U \supseteq X \tag{12}$$

which is readily proved and offers a useful characterization of the relation just defined. A still simpler characterization is expressed by the following formula, which makes use of the set-theoretic labels in Figure 2, with $V = B \cup F \cup G$, $U = D \cup E \cup H$, $Y = A \cup E \cup F$, and $X = C \cup G \cup H$:

$$(V|U) \preceq (Y|X) \text{ if and only if } B \cup C \cup D \cup G = \varnothing. \tag{13}$$

The relation introduced with Definition 1 endows its domain with a regular algebraic organization.

Proposition 1. *The relation $\preceq$ is a partial order over the set $\widetilde{O}(T) = O(T) \cup \{\perp\}$. It organizes this set as a lattice, whose supremum and infimum are the pair $(T|\varnothing)$ and the term $\perp$, respectively, and whose join*

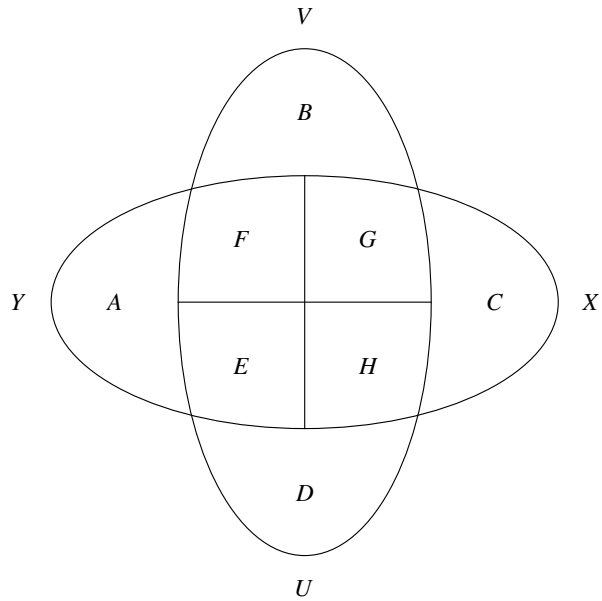


Figure 2. Quatrefoil representing a crossing between two generic variable pairs $(V|U)$ and $(Y|X)$. Some of the eight parts could be empty.

and meet operations are determined by the following equations, for all $(V|U)$ *and* $(Y|X)$ *in* $O(T)$:

$$(V|U) \vee (Y|X) = (V \cup Y \cup (U + X)|U \cap X) \tag{14}$$

with $U + X = (U \setminus X) \cup (X \setminus U)$;

$$(V|U) \wedge (Y|X) = (V \cap Y|U \cup X) \text{ or } = \bot \text{ depending on whether} \tag{15}$$

the conditions $V \cap Y \neq \varnothing$, $U \subseteq Y \cup X$, *and* $X \subseteq V \cup U$ *are or are not jointly true.*

Proof. In the light of characterization (12), it directly appears that reflexivity, transitivity, and antisymmetry of the relation $\preceq$ follow from the homonymous properties of the set-theoretic inclusion $\subseteq$. Thus, $(\widetilde{O}(T), \preceq)$ is a poset (partially ordered set) having $(T|\varnothing)$ as its supremum and $\bot$ as its infimum (concluding statement in Definition 1). Let us consider any two members $(V|U)$ and $(Y|X)$ of the set $O(T)$. We develop our argument concerning their join and meet in three stages. *First* we note that $(V \cup Y \cup (U + X)|U \cap X)$ is itself a member of $O(T)$ (the intersection between left variable and right variable in the pair is empty) and both $(V|U)$ and $(Y|X)$ are JC-derivable from it (according to (12)). Furthermore, if $(Z|W)$ is any member of $O(T)$ such that $(V|U) \preceq (Z|W)$ and $(Y|X) \preceq (Z|W)$, then ($V \cup U \subseteq Z \cup W$ and $U \supseteq W$) and ($Y \cup X \subseteq Z \cup W$ and $X \supseteq W$) (again because of (12)), so that ($V \cup Y \cup (U + X) \subseteq Z \cup W$ and $U \cap X \supseteq W$), which implies $(Y \cup V \cup (U + X)|U \cap X) \preceq (Z|W)$. Thus, $(V \cup Y \cup (U + X)|U \cap X)$ is the least upper bound of $(V|U)$ and $(Y|X)$ in the poset, which proves the equation concerning the join. At a *second stage*, let us suppose that $(V|U)$ and $(Y|X)$ satisfy the three conditions

$$V \cap Y \neq \varnothing, \quad U \subseteq Y \cup X, \quad X \subseteq V \cup U \tag{16}$$

and consider the variable pair $(V \cap Y|U \cup X)$. It is seen that this pair is a member of $O(T)$ (in particular, $V \cap Y \neq \varnothing$ is the first hypothesis in (16)), and is JC-derivable both from $(V|U)$ (in particular, $(V \cap Y) \cup U \cup X \subseteq V \cup U$ is ensured by the third hypothesis in (16)) and from $(Y|X)$ (for similar reasons). Furthermore, if $(Z|W)$ is any member of $O(T)$ such that $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, then ($Z \cup W \subseteq V \cup U$ and $W \supseteq U$) and ($Z \cup W \subseteq Y \cup X$ and $W \supseteq X$), so that $Z \cup W \subseteq (V \cup U) \cap (Y \cup X)$ and $W \supseteq U \cup X$. But the second and third hypotheses in (16) imply $(V \cup U) \cap (Y \cup X) = (V \cap Y) \cup (U \cap Y) \cup (V \cap X) \cup (U \cap X) = (V \cap Y) \cup U \cup X$. Therefore, $Z \cup W \subseteq (V \cap Y) \cup U \cup X$ and $W \supseteq U \cup X$, which means $(Z|W) \preceq (V \cap Y|U \cup X)$. Because of the genericity of $(Z|W)$, this proves that $(V \cap Y|U \cup X)$ is the greatest lower bound of $(V|U)$ and $(Y|X)$ in the poset, and confirms the stated formula for the meet. At a *third stage*, we refer to any two members $(V|U)$ and $(Y|X)$ of $O(T)$ that falsify some of the conditions in (16) and prove that there cannot exist any member $(Z|W)$ of $O(T)$ such that both $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, so that $\perp$ is the only common lower bound of $(V|U)$ and $(Y|X)$ in the poset, which means $(V|U) \wedge (Y|X) = \perp$. Suppose the first condition in (16) is false, that is $V \cap Y = \varnothing$ is true. Should a pair $(Z|W)$ exist in $O(T)$ such that $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, then (12) combined with $V \cap U = \varnothing = Y \cap X$ would imply $V \cap Y \supseteq Z$, so that $Z = \varnothing$, which contradicts the assumption $(Z|W) \in O(T)$. Next, suppose the second condition in (16) is false, that is $U \setminus (Y \cup X) \neq \varnothing$ is true, so that there is some non-empty variable $S \subseteq U \setminus (Y \cup X)$. Should a pair $(Z|W)$ exist in $O(T)$ such that $(Z|W) \preceq (V|U)$ and $(Z|W) \preceq (Y|X)$, then we would have $S \subseteq W$ (because $W \supseteq U$) and not($S \subseteq W$) (because $W \subseteq Y \cup X$), which of course is contradictory. The same argument may be applied when the third condition in (16) is false. □

Using the labels in Figure 2, the equations that specify joins and meets in the lattice $\widetilde{O}(T)$ can be written as

$$(V|U) \vee (Y|X) = (A \cup B \cup C \cup D \cup E \cup F \cup G|H) \tag{17}$$
$$(V|U) \wedge (Y|X) = (F|E \cup G \cup H) \text{ if } F \neq \varnothing \text{ and } C \cup D = \varnothing.$$

The atoms in the lattice are the pairs $(V|U)$ such that $|V| = 1$, that is, the left-hand component is an elementary variable. Thus, if $n = |T|$, then there are $n2^{n-1}$ atoms. It is readily proved that any member of $O(T)$ is expressible as the join of suitably chosen atoms and that the lattice is rankable, the rank of any pair $(V|U)$ being the cardinality $|V|$ of its left-hand component. Figure 3 illustrates the concept by showing three-dimensional Hasse diagrams of three lattices of variable pairs. In these diagrams, each backward (resp., downward) line represents a projection operation $J[(Y|X), W]$ (resp., a conditioning operation $C[(Y|X), W]$) with $|W| = 1$.

The relation $\preceq$, which organizes the set $\widetilde{O}(T)$ as a lattice, has been defined in terms of projection J and conditioning C as specified by (2) and (3). Some additional comments are in order concerning promotion M as specified by (4). First, if $(V|U)$ is a promoter of $(Y|X)$ (i.e., $V \cup U = X$), then the two variable pairs are $\preceq$-incomparable (because $Y \cap V = \varnothing$) and the result of their promotion equals their join, that is

$$M[(Y|X), (V|U)] = (Y \cup V|U) = (Y|X) \vee (V|U).$$

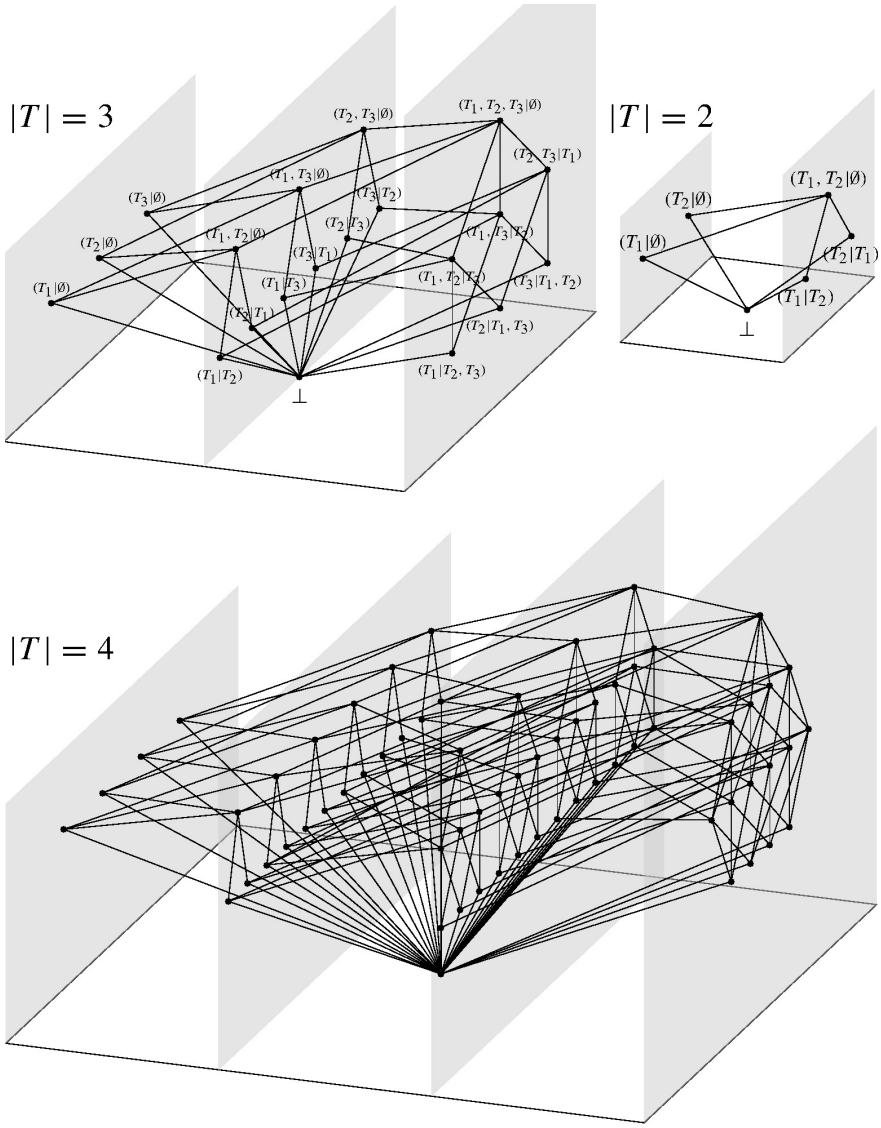


Figure 3. Three-dimensional Hasse diagrams of the lattices $\widetilde{O}(T)$ for $|T| = 2$ (top right), $|T| = 3$ (top left), and (bottom) $|T| = 4$.

Thus, promotion M within a lattice $\widetilde{O}(T)$ is tantamount to a part of the join operation. Second, if $(V|U) \prec (Y|X)$, then one or two promotions are enough for ascending from $(V|U)$ to $(Y|X)$ within the lattice. Specifically, in a writing justified by (9) (associativity of M), the following equation holds true:

$$(Y|X) = (Y \setminus (V \cup U)|V \cup U)\ M\ (V|U)\ M\ (Y \cap U|X). \tag{18}$$

Indeed, if $(V|U) \preceq (Y|X)$ but not $(Y|X) \preceq (V|U)$ (which is the meaning of the hypothesis $(V|U) \prec (Y|X)$), then using the labels in Figure 2 we obtain $B \cup C \cup D \cup G = \varnothing$ but $A \cup E \neq \varnothing$ (because of (13) and

its rewrite characterizing $(Y|X) \preceq (V|U)$), so that $V = F$, $U = E \cup H$, $Y = A \cup E \cup F$, and $X = H$, and (18) may be rewritten as $(A \cup E \cup F|H)$=$(A|E \cup F \cup H)\ M\ (F|E \cup H)\ M\ (E|H)$, which is true according to the definition of the M operation. Note that if either $A = \varnothing$ or $E = \varnothing$, then either the first or the third operand in the right hand side of (18) would be the infimum $\perp$ in the lattice and could be ignored (consistently with (5)), so that one single application of M would be enough for ascending from $(V|U)$ to $(Y|X)$. Third, as a special case of (18) we note the following equation:

$$(V|U) \vee (Y|X) = ((Y \cup X) \setminus (V \cup U)|V \cup U)\ M\ (V|U)\ M\ (U \setminus X|U \cap X).$$

It can be directly proved by noting that the indicated double promotion gives the result $((Y \cup X) \setminus (V \cup U) \cup V \cup (U \setminus X)|U \cap X)$, that is $(A \cup B \cup C \cup D \cup E \cup F \cup G|H)$ by the labeling in Figure 2, and this variable pair is precisely the join $(V|U) \vee (Y|X)$ according to (17). The equation thus proved supplements the first comment in this paragraph, by showing that single and double promotions are enough to simulate the *whole* of the join operation within any lattice of variable pairs.

3. Lattice of probability kernels

In this section we return to probability kernels mentioned in the Introduction, in order to show how the results obtained in discussing variable pairs in the preceding section may conveniently be applied to them.

In the first step, we present the following formulas, which define projection J, conditioning C, and promotion M as operations on probability kernels:

$$J[p(W \cup Z|X), W] = p(Z|X) \tag{19}$$
$$\text{in which } p(z|x) = \sum_{w \in W^\circ} p(w, z|x) \text{ for all } (x, z) \in (X, Z)^\circ$$

$$C[p(W \cup Z|X), Z] = p(W|Z \cup X) \tag{20}$$
$$\text{in which } p(w|z, x) = \frac{p(w, z|x)}{\sum_{w' \in W^\circ} p(w', z|x)} \text{ for all } (x, w, z) \in (X, W, Z)^\circ$$

$$M[p(Y|V \cup U), p(V|U)] = p(Y \cup V|U) \tag{21}$$
$$\text{in which } p(y, v|u) = p(y|v, u) \cdot p(v|u) \text{ for all } (u, v, y) \in (U, V, Y)^\circ.$$

These formulas generalize the rules mentioned in the Introduction and correspond to operations ordinarily performed on conditional probabilities in Bayesian computations (Bernardo & Smith, 2000, pp. 127–130; Koski & Noble, 2009, pp. 53–57). Here we assume that X, W, and Z — as well as U, V, and Y — are (possibly multiple) variables that are disjoint from one another, and that $p(W \cup Z|X)$, $p(Y|V \cup U)$, and $p(V|U)$ are probability kernels acting on them. Formulas (19) and (20), in the given writing, apply when the kernel $p(W \cup Z|X)$ is of discrete type; similar formulas, with Σ replaced by $\int$, are suitable for kernels of continuous type. Of course, the specification of the term $p(w|z, x)$ in formula (20) is acceptable only for any point (x, w, z) such that the denominator in the fraction is non-null, that is, the value $p(z|x)$ resulting from projection is positive. Furthermore, the kernels $p(Y|V \cup U)$ and $p(V|U)$ in formula (21) are here assumed to be of the same measure-theoretic type, that is, either both of them

are families of mass functions (on domains Y° and V°, respectively), or both are families of density functions[2]. Lastly, it is readily seen that there is correspondence between the stated operations on kernels and the operations on variable pairs defined by (2)–(4). For example, if $p(Z|X) = J[p(Y|X), W]$ then $(Z|X) = (Y \setminus W|X) = J[(Y|X), W]$, and similarly for the conditioning and promotion operations. Also it is easily proved that the (6)–(11) still hold true when they are rewritten in terms of probability kernels and operations on these.

The second step in this section is about the relation $\preceq$ specified by Definition 1, which may be recast in terms of probability kernels as follows:

$$\begin{aligned} &p(V|U) \text{ is } JC\text{-derivable from } p(Y|X) \text{ (notation } p(V|U) \preceq p(Y|X)) \\ &\text{if } p(V|U) = J[C[p(Y|X), W], Z] \text{ for some } W, Z \subseteq Y \text{ such that } W \cap Z = \varnothing. \end{aligned} \tag{22}$$

Just as with the relation $\preceq$ between variable pairs, this relation between kernels is a partial order. Indeed, it is reflexive, as W and Z in (22) could be the empty variable. It is transitive, by virtue of properties (6)–(8) as referred to the J and C operations on kernels. It is antisymmetric, because if $p(V|U) \preceq p(Y|X)$ and $p(Y|X) \preceq p(V|U)$, then also $(V|U) \preceq (Y|X)$ and $(Y|X) \preceq (V|U)$, so that $(V|U) = (Y|X)$, which combined with $p(V|U) \preceq p(Y|X)$ implies $p(V|U) = p(Y|X)$. Note that, in comparing any two kernels $p(V|U)$ and $p(V|W \cup U)$, it could turn out that

$$p(v|u) = p(v|w, u) \text{ for all } (u, v, w) \in (U, V, W)^\circ \tag{23}$$

which would mean that V and W are "conditionally independent" given U (Dawid, 1979, p. 3). In that event, $p(V|W \cup U)$ would amount to a replica of $p(V|U)$, and we would accept $p(V|U) \preceq p(V|W \cup U)$ as a true sentence, for a reason similar to accepting $p(V|U) \preceq p(V|U)$ as a true sentence.

In order to relate Proposition 1 to probability kernels, we need to refer to a complete collection of mutually consistent kernels. Let a full kernel $p(T) = p(T|\varnothing)$ be given, that is, one single probability function over the range T° of the assumed full variable T. Then, in correspondence to any variable pair $(Y|X) \in O(T)$, we may JC-derive a kernel $p(Y|X)$ from $p(T|\varnothing)$ by applying the conditioning operation (20) relative to the variable X (i.e., variable X is transferred from the left to the right side of the bar) and then applying the projection operation (19) relative to the variable $T \setminus (Y \cup X)$ (i.e., variable $T \setminus (Y \cup X)$ is canceled from $T \setminus X$, so that precisely Y is what remains on the left side of the bar). By doing so for each of the variable pairs in $O(T)$, a complete collection of kernels is obtained, here denoted by $P(T)$ and thus formally defined:

$$P(T) = \big\{p(Y|X) : p(Y|X) = J[C[p(T|\varnothing), X], T \setminus (Y \cup X)] \text{ for } (Y|X) \in O(T)\big\}.$$

[2]This is a limitation of the M operation as understood in this paper, which may be overcome by setting the concept of probability kernel in measure-theoretic terms. Kernels $p(Y|V \cup U) = (p(Y|v, u) : v \in V^\circ, u \in U^\circ)$ and $p(V|U) = (p(V|u) : u \in U^\circ)$ may generally be families of Radon-Nikodym derivatives (of probability distributions) with respect to reference measures (possibly of different kinds) μ and ν on the spaces Y° and V°, respectively (Billingsley, 1995, pp. 439–440; Pollard, 2002, pp. 84, 119). Hence, for each $u \in U^\circ$ the product function $p(Y \cup V|u) = (p(y|v, u) \cdot p(v|u) : y \in Y^\circ, v \in V^\circ)$ in turn is a Radon-Nikodym derivative (of a probability distribution) with respect to the product measure $\mu \times \nu$ on the product space $Y^\circ \times V^\circ$, and the promoted kernel $p(Y \cup V|U) = (p(Y \cup V|u) : u \in U^\circ)$ is the whole collection of these derivatives.

The kernels in the collection $P(T)$ are mutually consistent as they originate from the same "parent" $p(T)$ — indeed, $P(T)$ is the collection of all kernels that are $\preceq$-dominated by the assumed full distribution $p(T)$. In addition, let $\widetilde{P}(T)$ stand for $P(T) \cup \{\sharp\}$, where the term $\sharp$ — here called the *null kernel* — is assumed to be lower in the order $\preceq$ than all members of $P(T)$ and represents "fictitious kernels" $p(\varnothing|X)$ for $X \subseteq T$[3]. The one-to-one correspondence between variable pairs in $O(T)$ and kernels in $P(T)$ — and between the terms $\perp$ and $\sharp$ — is an isomorphism between the ordered sets $(\widetilde{O}(T), \preceq)$ and $(\widetilde{P}(T), \preceq)$. Proposition 1 shows that the former set is a lattice, so that also the latter is a lattice. The assumed full kernel $p(T)$ and the null kernel $\sharp$ are the supremum and the infimum in the lattice $\widetilde{P}(T)$. The kernels $p(Y|X)$ in which $|Y| = 1$ are the atoms. The join and meet operations are defined by formulas that duplicate (14) and (15) in terms of probability kernels.

The preceding argument has been cast in a top-down perspective, as a full kernel $p(T)$ has been assumed available, from which a complete collection $P(T)$ of mutually consistent kernels may be derived. But, in the construction of probabilistic models, a perspective somewhat opposite to this is actually taken. Indeed, in constructing a model, kernels of low rank are first specified — that is, kernels $p(Y|X)$ in which Y is a variable of small cardinality, possibly an elementary variable, for simplicity. These are the building blocks of the model, from which other kernels of higher rank are obtained — by combining the building blocks through promotion or multiplication under suitable assumptions of stochastic independence — and then other lower rank kernels can be JC-derived for the necessities of the modeling (Koller & Friedman, 2009, pp. 4–5).

Illustrations of this circumstance are offered by hierarchical Bayesian models in statistics. Let us consider, for example, the following assignment formulas, in which D is an observable random variable (it could be the mean of a sample of data), whose distribution is assumed to involve a location parameter μ and a precision parameter λ, which themselves are conceived of as random variables with distributions depending on hyper-parameters ν, ξ, α, and β, these being provided with definite hyper-prior distributions (a model compatible with Bernardo & Smith, 2000, p. 440):

$$\begin{aligned} &D \sim \text{Normal}(\mu, \lambda), \quad \mu \sim \text{Normal}(\nu, \xi), \quad \lambda \sim \text{Gamma}(\alpha, \beta), \\ &\nu \sim \text{Uniform}(0, 50), \quad \xi \sim \text{Uniform}(0, 1), \quad \alpha \sim \text{Uniform}(0, 1), \quad \beta \sim \text{Uniform}(0, 1). \end{aligned} \tag{24}$$

In the terms we are using in this paper, these formulas specify seven primitive kernels $p(D|\mu, \lambda)$, $p(\mu|\nu, \xi)$, $p(\lambda|\alpha, \beta)$, $p(\nu|\varnothing)$, $p(\xi|\varnothing)$, $p(\alpha|\varnothing)$, and $p(\beta|\varnothing)$, which express postulates in the model (their variable pairs are atoms in a lattice, as the conditioned variables are one-dimensional). For purposes of statistical inference, the modeler may be interested in determining the kernel $p(\mu|D)$ that is implied by the assumed postulates, that is, the family of posterior distributions of the parameter μ given the observable quantity D. But for determining such low rank kernel (itself an atom) one must first ascend from the given primitive kernels to the top kernel $p(D, \mu, \lambda, \nu, \xi, \alpha, \beta|\varnothing)$ and then descend from this to $p(\mu|D)$ through suitable conditioning and projections. Of course, such a procedure is only plausible if a kernel $p(D, \mu, \lambda, \nu, \xi, \alpha, \beta|\varnothing)$ that $\preceq$-dominates all seven primitive kernels really exists (and is

[3] In numerical operations, the null kernel $\sharp$ acts as the number 1, which is the neutral term of multiplication (cf. Studený, 2005, p. 20).

reachable from these by ascending operations, possibly supported by suitable independence assumptions). In general, however, if the primitive kernels in a model are specified separately from one another, there is no a priori guarantee that there is a "consensus kernel" covering all of them, so that the stated problem might fail to have a solution.

In addressing this topic, use must be made of the concept of "compatibility" as understood in the discussions concerning conditional specified distributions and, more generally, conditionally specified statistical models (Arnold, Castillo, & Sarabia, 1999, 2001). The concept can consistently be framed within the theory developed so far in this paper.

Definition 2. Given kernels $p(Y_1|X_1), \ldots, p(Y_m|X_m)$ are *compatible* with one another if there is a kernel $p(Z|W)$ from which all of them are JC-derivable, that is, $p(Y_i|X_i) \preceq p(Z|W)$ for all $i = 1, \ldots, m$.

Note that if $p(Z|W)$ is a kernel that satisfies this condition, then $(Y_i|X_i) \preceq (Z|W)$ for all $i = 1, \ldots, m$, so that $(Y_1|X_1) \vee \cdots \vee (Y_m|X_m) \preceq (Z|W)$. Thus, for verifying whether m given kernels are mutually compatible, the standard approach would be to verify whether on the *join* of their variable pairs a kernel may be constructed from which all of them can be JC-derived. Also note that this compatibility relation fails to be universally transitive. For example, if X and Y are disjoint discrete variables, then any kernel $p(X|\varnothing)$ is compatible both with any kernel $p(Y|X)$ and with any kernel $p(Y|\varnothing)$, but these two could be incompatible with each other — certainly they are compatible if $p(Y|\varnothing) = J[M[p(Y|X), p(X|\varnothing)], X]$.

For applications, one wants to have conditions that are easily testable and sufficient to ensure compatibility. The next proposition provides such a condition, which is rather general, as it *only* concerns the *variable pairs* on which the kernels are acting. No mention is made of the specific form of the probability functions in the kernels.

Proposition 2. *Let* $(p(Y_1|X_1), \ldots, p(Y_m|X_m))$ *be a list of* $m \geq 2$ *probability kernels of the same type* (*i.e.*, *either all of discrete type*, *or all of continuous type*) *such that their variable pairs satisfy this condition*:

$$Y_i \cap (Y_{i-1} \cup X_{i-1} \cup \cdots \cup Y_1 \cup X_1) = \varnothing, \text{ for all } i = 2, \ldots, m. \tag{25}$$

Then the m kernels are compatible.

Proof. The proof is by induction on the number $m \geq 2$ of kernels. *First step*: For $m = 2$, let any two variable pairs $(Y_1|X_1) = (Y|X)$ and $(Y_2|X_2) = (V|U)$ be given such that $V \cap (Y \cup X) = \varnothing$, that is $F \cup G = \varnothing$ in the terms of Figure 2, so that $(Y|X) \vee (V|U) = (A \cup B \cup C \cup D \cup E|H)$ according to (17). Let $p(Y|X) = p(A \cup E|C \cup H)$ and $p(V|U) = p(B|D \cup E \cup H)$ be arbitrary kernels of the same measure-theoretic type on the indicated variable pairs. We may extend these kernels into $p(A \cup E \cup D|C \cup H)$ and $p(B|A \cup C \cup D \cup E \cup H)$ by setting

$$p(a, e, d|c, h) = p(a, e|c, h) \cdot p(d), \text{ for all } (a, e, d, c, h) \in (A, E, D, C, H)^\circ$$
$$p(b|a, c, d, e, h) = p(b|d, e, h), \text{ for all } (b, a, c, d, e, h) \in (B, A, C, D, E, H)^\circ$$

where $p(D)$ is a freely chosen kernel of the same type as the two given kernels. Then we may combine the extended kernels by promotion to obtain the following kernel on the join $(Y|X) \vee (V|U)$:

$$p(A \cup B \cup C \cup D \cup E|H) = p(B|A \cup C \cup D \cup E \cup H) \; M \; p(A \cup E \cup D|C \cup H) \; M \; p(C|H)$$

where $p(C|H)$ is itself a freely chosen kernel. Note that the double promotion in this equation is syntactically regular, as the third kernel is a promoter of the second, and in turn this is a promoter of the first. Kernel $p(A \cup E|C \cup H)$ is derivable from $p(A \cup E \cup D|C \cup H)$ by projection, and this is derivable from $p(A \cup B \cup C \cup D \cup E|H)$ because of (10) and (11). Furthermore, the remark associated to (23) ensures that $p(B|D \cup E \cup H)$ is derivable from $p(B|A \cup C \cup D \cup E \cup H)$, and this is derivable from $p(A \cup B \cup C \cup D \cup E|H)$ because of (10). Therefore, $p(Y|X) = p(A \cup E|C \cup H)$ and $p(V|U) = p(B|D \cup E \cup H)$ are compatible kernels, as there exists a kernel from which both of them are derivable. Note that $p(A \cup B \cup C \cup D \cup E|H)$ does not involve any variable besides those involved in $p(A \cup E|C \cup H)$ or in $p(B|D \cup E \cup H)$. *Inductive step*: Let us now consider any list $(p(Y_1|X_1), \ldots, p(Y_{m-1}|X_{m-1}), p(Y_m|X_m))$ of $m > 2$ kernels whose variable pairs comply with condition (25), and suppose (as inductive hypothesis) that the first $m - 1$ members in the list are compatible with one another, so that there is a kernel $p(Z|W)$ from which all of them are derivable. Based on the remark that concludes the first step of the current proof, we may presume $Z \cup W \subseteq Y_{m-1} \cup X_{m-1} \cup \cdots \cup Y_1 \cup X_1$, so that $Y_m \cap (Z \cup W) = \varnothing$ because of hypothesis (25). Thus, the conditions are satisfied that make it possible to apply the argument in the first step of the current proof to the kernels $p(Z|W)$ and $p(Y_m|X_m)$. The argument ensures the existence of a kernel from which both $p(Z|W)$ (hence $p(Y_1|X_1), \ldots, p(Y_{m-1}|X_{m-1})$) and $p(Y_m|X_m)$ are derivable. Therefore, the m kernels are all compatible with one another. □

The proposition thus proved directly shows the internal consistency of the Bayesian model specified in (24). Indeed, if the primitive kernels in the model are listed in the order

$$(p(\beta|\varnothing),\ p(\alpha|\varnothing),\ p(\xi|\varnothing),\ p(\nu|\varnothing),\ p(\lambda|\alpha, \beta),\ p(\mu|\nu, \xi),\ p(D|\mu, \lambda)),$$

then it appears that the conditioned variable in each kernel falls out of the collection of the variables involved in the kernels preceding that kernel in the list, so that condition (25) for mutual compatibility of the kernels is satisfied. More generally, suppose that $(T_{r(1)}, \ldots, T_{r(n)})$ is an ordering of the elementary variables that form the full variable $T = \{T_1, \ldots, T_n\}$ of a probabilistic model, and that $(T_{r(1)}|X_1), \ldots, (T_{r(n)}|X_n)$ are variable pairs — specifically, atoms in the lattice $\widetilde{O}(T)$ — such that $X_i \subseteq T_{r(1)} \cup \ldots \cup T_{r(i-1)}$, for each $i = 1, \ldots, n$ (in particular, $X_1 = \varnothing$). Then Proposition 2 implies, as a corollary, that *any* kernels $p(T_{r(1)}|X_1), \ldots, p(T_{r(n)}|X_n)$ of the same measure-theoretic type (e.g., all discrete or all continuous kernels) on those variable pairs are compatible with one another, and such kernels uniquely determine (through multiple promotion, possibly supported by assumptions of stochastic independence) a distribution $p(T)$ on the full variable of the model. Thus, possible kernels on the atom variable pairs $(T_{r(1)}|X_1), \ldots, (T_{r(n)}|X_n)$ form a system of independent and minimum-rank generators of possible full distributions for the model.

The corollary now highlighted corresponds to a key result in the theory of Bayesian networks. Indeed, if the elementary variables in a collection $T = \{T_1, \ldots, T_n\}$ are represented as nodes of an acyclic directed graph (DAG), then a list $((T_{r(1)}|X_1), \ldots, (T_{r(n)}|X_n))$ of variable pairs with the stated characteristics can be formed, in which $(T_{r(1)}, \ldots, T_{r(n)})$ is a suitable permutation of T and X_i (for $i = 1, \ldots, n$) is the set of "parents" of the variable $T_{r(i)}$ in the graph. Therefore, if $p(T_{r(1)}|X_1), \ldots, p(T_{r(n)}|X_n)$ are the elementary probability kernels postulated in the Bayesian network, then by virtue of the stated corollary a joint "consensus" distribution $p(T)$ does exist and this may be uniquely inferred (also thanks to the

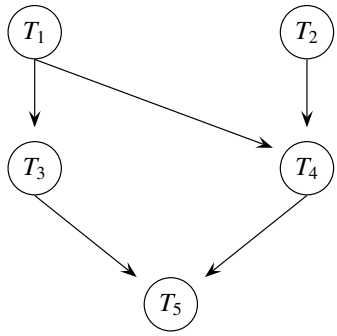


Figure 4. The DAG of a Bayesian network on five elementary variables (from Kjærulff & Madsen, 2008, p. 13).

conditional independences represented in the DAG) through the "chain rule", which corresponds to multiple promotion in our terms (Pearl, 1988, pp. 119–120; Darwiche, 2009, pp. 57–58). The DAG in Figure 4 offers the basis for an illustration of the stated property. By associating to each elementary variable the set of its parents in the graph — that is, the variables emitting an arrow towards that variable — the following list of variable pairs is obtained:

$$(T_1|\varnothing),\ (T_2|\varnothing),\ (T_3|T_1),\ (T_4|T_1,T_2),\ (T_5|T_3,T_4).$$

It is seen that, in this ordering, the variable pairs do comply with condition (25) (a circumstance ultimately due to the acyclic character of the graph), so that Proposition 2 ensures mutual compatibility among the probability kernels that a modeler may specify in defining a Bayesian network on the DAG:

$$p(T_1|\varnothing),\ p(T_2|\varnothing),\ p(T_3|T_1),\ p(T_4|T_1,T_2),\ p(T_5|T_3,T_4). \tag{26}$$

Furthermore, the DAG represents a system of conditional stochastic independences, which are assumed true by the modeler. For example, it implies that the variable T_5 is assumed to be conditionally independent of the variable $\{T_1, T_2\}$ given the variable $\{T_3, T_4\}$, this being the set of parents of T_5. Because of such independences encoded in the hypothesized graph, the specification (by the modeler) of the kernels (26) directly entails (on account of (23)) the specification of these kernels:

$$p(T_5|T_1,T_2,T_3,T_4),\ p(T_4|T_1,T_2,T_3),\ p(T_3|T_1,T_2),\ p(T_2|T_1),\ p(T_1|\varnothing).$$

It is seen that, in this ordering (which for convenience is the opposite of that in (26)), each kernel is a regular promoter of its immediate predecessor in the list. Thus, multiple promotion can be applied, whose result will be the full distribution $p(T_1, T_2, T_3, T_4, T_5)$ the existence of which is ensured by the configuration of the variable pairs owing to Proposition 2. Such multiple promotion amounts to an ascension from five atoms in a lattice of kernels to their join (which, in this special case, equals the supremum of the lattice itself).

4. Order of variable pairs and compatibility of kernels

In this section, we present a result that illustrates the diagnostic power of the order $\preceq$ between variable pairs as concerns a special aspect of compatibility between probability kernels. More precisely, we will show that, for all variable pairs $(V|U)$ and $(Y|X)$, one has $(V|U) \preceq (Y|X)$ if and only if *each* kernel $p(Y|X)$ on the latter pair *uniquely determines* a kernel $p(V|U)$ on the former. In other words, if $p(Y|X)$ is given and $(V|U) \preceq (Y|X)$, then there exists one single kernel on $(V|U)$ compatible with it, whereas if not $(V|U) \preceq (Y|X)$, then there exist different kernels on $(V|U)$ compatible with the same $p(Y|X)$. As in the preceding section, implicit in the following arguments is the assumption that the kernels to be combined or compared are of the same measure-theoretic type (either discrete or continuous).

In proving the indicated result, use will be made of one further operation on kernels, called *multiplication*, which is defined as follows (cf. Koski & Noble, 2009, pp. 55–56):

$$\begin{aligned}
&\text{for all } p(Y|X) \text{ and } p(Z|W) \text{ such that } Y \cap (Z \cup W) = \varnothing = Z \cap (Y \cup X) \\
&p(Y|X) \times p(Z|W) = p(Y \cup Z|X \cup W) \\
&\text{in which } p(y, z|t, u, v) = p(y|t, u) \cdot p(z|u, v) \\
&\text{for all } (y, z) \in (Y, Z)^\circ \text{ and } (t, u, v) \in (X \setminus W, X \cap W, W \setminus X)^\circ.
\end{aligned} \tag{27}$$

This formula is applicable also to cases in which $W = \varnothing$ (so that $p(Z|W) = p(Z|\varnothing)$ is one single probability function on the range Z°) or $Z = \varnothing$ (so that $p(Z|W) = p(\varnothing|W)$ stands for the null kernel $\sharp$; see footnote 3). In these special cases, the definition becomes

$$\begin{aligned}
&p(Y|X) \times p(Z|\varnothing) = p(Y \cup Z|X) \\
&\qquad \text{in which } p(y, z|x) = p(y|x) \cdot p(z) \text{ for all } (y, z, x) \in (Y, Z, X)^\circ \\
&p(Y|X) \times p(\varnothing|W) = p(Y|X \cup W) \\
&\qquad \text{in which } p(y|x, w) = p(y|x) \text{ for all } (y, x, w) \in (Y, X, W)^\circ.
\end{aligned}$$

It is easily shown that the product $p(Y|X) \times p(Z|W)$ defined by (27) is itself a probability kernel; more specifically, it is a family (indexed by $(X \cup W)^\circ$) of probability functions on $(Y, Z)^\circ$. Equally it can be shown that multiplication is an associative and commutative operation, and for every variable $S \subseteq Y$ it satisfies these equations:

$$J[p(Y|X) \times p(Z|W), S] = (J[p(Y|X), S]) \times p(Z|W) \tag{28}$$

$$C[p(Y|X) \times p(Z|W), S] = (C[p(Y|X), S]) \times p(Z|W). \tag{29}$$

In other words, a kind of commutativity holds between multiplication $\times$ on the one hand and projection J and conditioning C on the other hand.

Now we state and prove the main result in this section.

Proposition 3. *Let $(V|U)$ and $(Y|X)$ be any two non-null members of a lattice $\widetilde{O}(T)$ of variable pairs. Then $(V|U) \preceq (Y|X)$ if and only if for all kernels $p(V|U)$, $p(Y|X)$, $q(V|U)$, and $q(Y|X)$ such that the*

first two are compatible, *and also the other two are compatible*, *the equality* $p(Y|X) = q(Y|X)$ *implies the equality* $p(V|U) = q(V|U)$.

Proof. The "only if" part of this proposition is easily shown, on considering that if $(V|U) \preceq (Y|X)$, then the stated compatibility hypotheses imply $p(V|U) \preceq p(Y|X)$ and $q(V|U) \preceq q(Y|X)$, that is

$$p(V|U) = J[C[p(Y|X), U \setminus X], Y \setminus (V \cup U)] \text{ and } q(V|U) = J[C[q(Y|X), U \setminus X], Y \setminus (V \cup U)]$$

so that the equality $p(Y|X) = q(Y|X)$ obviously implies the equality $p(V|U) = q(V|U)$. To prove the "if" part is tantamount to proving that

$$\begin{aligned} &\text{if not } (V|U) \preceq (Y|X) \\ &\text{then there are kernels } p(V|U),\ p(Y|X), q(V|U), \text{ and } q(Y|X) \text{ such that} \\ &p(V|U) \text{ and } p(Y|X) \text{ are compatible, } q(V|U) \text{ and } q(Y|X) \text{ are compatible, and} \\ &p(Y|X) = q(Y|X) \text{ but } p(V|U) \neq q(V|U). \end{aligned} \tag{30}$$

In the terms of Figure 2 and on account of (13), the antecedent "not $(V|U) \preceq (Y|X)$" in this implication means that the condition $B \cup C \cup D \cup G \neq \varnothing$ holds true. Hereafter we separately discuss (using the labels in Figure 2) three cases that exhaustively cover this condition. *Case* $B \cup G \neq \varnothing$. Choose any two probability functions $p(B \cup G)$ and $q(B \cup G)$ that are *different* from each other — such functions do exist, simply because $B \cup G \neq \varnothing$. Then consider any probability function $r(A \cup C \cup D \cup E \cup F \cup H)$, combine it with $p(B \cup G)$ and $q(B \cup G)$ by multiplication, thus obtaining

$$\begin{aligned} p(W) &= p(A \cup B \cup C \cup D \cup E \cup F \cup G \cup H) = p(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H) \\ q(W) &= q(A \cup B \cup C \cup D \cup E \cup F \cup G \cup H) = q(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H) \end{aligned}$$

JC-derive the following kernels from $p(W)$

$$\begin{aligned} p(Y|X) &= p(A \cup E \cup F|C \cup G \cup H) = J[C[p(W), C \cup G \cup H], B \cup D] \\ p(V|U) &= p(B \cup F \cup G|D \cup E \cup H) = J[C[p(W), D \cup E \cup H], A \cup C] \end{aligned}$$

and similarly the kernels $q(Y|X)$ and $q(V|U)$ from $q(W)$. Kernels $p(Y|X)$ and $p(V|U)$ are compatible, because they are JC-derived from the *same* $p(W)$, and for a similar reason also $q(Y|X)$ and $q(V|U)$ are compatible. Furthermore, on account of (28) and (29),

$$\begin{aligned} &p(Y|X) = \\ &J[C[p(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H), C \cup G \cup H], B \cup D] = \\ &J[C[p(B \cup G), G] \times C[r(A \cup C \cup D \cup E \cup F \cup H), C \cup H], B \cup D] = \\ &J[p(B|G) \times r(A \cup D \cup E \cup F|C \cup H), B \cup D] = \\ &J[p(B|G), B] \times J[r(A \cup D \cup E \cup F|C \cup H), D] = \\ &p(\varnothing|G) \times r(A \cup E \cup F|C \cup H) \end{aligned}$$

and

$$
\begin{aligned}
&p(V|U) = \\
&J[C[p(B \cup G) \times r(A \cup C \cup D \cup E \cup F \cup H),\, D \cup E \cup H],\, A \cup C] = \\
&J[p(B \cup G) \times C[r(A \cup C \cup D \cup E \cup F \cup H),\, D \cup E \cup H],\, A \cup C] = \\
&J[p(B \cup G) \times r(A \cup C \cup F|D \cup E \cup H),\, A \cup C] = \\
&p(B \cup G) \times J[r(A \cup C \cup F|D \cup E \cup H),\, A \cup C] = \\
&p(B \cup G) \times r(F|D \cup E \cup H).
\end{aligned}
$$

Similar procedures show that $q(Y|X) = q(\varnothing|G) \times r(A \cup E \cup F|C \cup H)$ and $q(V|U) = q(B \cup G) \times r(F|D \cup E \cup H)$. Therefore, $p(Y|X) = q(Y|X)$ (because $p(\varnothing|G) = \sharp = q(\varnothing|G)$), but $p(V|U) \neq q(V|U)$ (because $p(B \cup G) \neq q(B \cup G)$ by the initial choice), so that the four kernels do comply with the requirements in the consequent of implication (30). *Case* $B \cup G = \varnothing$ and $D \neq \varnothing$. As $(V|U)$ is presumed different from $\perp$ (the null variable pair), the hypothesis $B \cup G = \varnothing$ implies that $F \neq \varnothing$. It is well known that, for any two (non-empty) disjoint variables F and D, probability functions $p(F \cup D)$ and $q(F \cup D)$ can be constructed such that $p(F) = J[p(F \cup D),\, D]$ and $q(F) = J[q(F \cup D),\, D]$ are *equal*, but $p(F|D) = C[p(F \cup D),\, D]$ and $q(F|D) = C[q(F \cup D),\, D]$ are *different*. Given such functions, by multiplication let us construct the following:

$$
\begin{aligned}
p(W) &= p(A \cup C \cup D \cup E \cup F \cup H) = p(F \cup D) \times r(A \cup C \cup E \cup H) \\
q(W) &= q(A \cup C \cup D \cup E \cup F \cup H) = q(F \cup D) \times r(A \cup C \cup E \cup H)
\end{aligned}
$$

where $r(A \cup C \cup E \cup H)$ is a freely chosen probability function. By applying the same method used above, the following results are obtained:

$$
\begin{aligned}
p(Y|X) &= p(A \cup E \cup F|C \cup H) = J[C[p(W),\, C \cup H],\, D] = p(F) \times r(A \cup E|C \cup H) \\
p(V|U) &= p(F|D \cup E \cup H) = J[C[p(W),\, D \cup E \cup H],\, A \cup C] = p(F|D) \times r(\varnothing|E \cup H)
\end{aligned}
$$

and similarly $q(Y|X) = q(F) \times r(A \cup E|C \cup H)$ and $q(V|U) = q(F|D) \times r(\varnothing|E \cup H)$. Thus, the equality $p(F) = q(F)$ implies $p(Y|X) = q(Y|X)$, and the inequality $p(F|D) \neq q(F|D)$ implies $p(V|U) \neq q(V|U)$, so that the four kernels satisfy what the implication (30) requires. *Case* $B \cup D \cup G = \varnothing$ and $C \neq \varnothing$. A well-known fact, symmetric to that mentioned above, is that for any two (non-empty) disjoint variables F and C, probability functions $p(F \cup C)$ and $q(F \cup C)$ can be constructed such that $p(F) = J[p(F \cup C),\, C]$ and $q(F) = J[q(F \cup C),\, C]$ are *different*, whereas $p(F|C) = C[p(F \cup C),\, C]$ and $q(F|C) = C[q(F \cup C),\, C]$ are *equal*. Drawing on this property, the third case in question can be solved by the same method used for the previous ones, on taking account that $(Y|X) = (A \cup E \cup F|C \cup H)$ and $(V|U) = (F|E \cup H)$ in the presumed situation. □

The proposition thus proved shows that the order relation $\preceq$ between variable pairs constitutes a general criterion for deducibility between probability kernels. For illustrating the meaning of this statement, let us assume $T = \{T_1, T_2, T_3\}$ as the full variable of a model and consider the following three variable pairs

in the lattice $\widetilde{O}(T)$ (see Figure 3 on page 86, top left):

$$(T_1, T_2|T_3),\ (T_1|T_2, T_3),\ (T_3|T_1, T_2).$$

In principle, on each of these variable pairs various (infinitely many) alternative probability kernels may be defined, but suppose that compatibility between kernels is required (cf. Definition 2). We may then ask: given the compatibility requirement, does a deterministic constraint exist between how a kernel on $(T_1, T_2|T_3)$ is chosen and how kernels on $(T_1|T_2, T_3)$ and $(T_3|T_1, T_2)$ may be chosen? Proposition 3 allows us to give the following answers. For *any* possible kernel $p(T_1, T_2|T_3)$ there is *one single* kernel $p(T_1|T_2, T_3)$ compatible with it (and derivable from it by a C operation), because the relation $(T_1|T_2, T_3) \preceq (T_1, T_2|T_3)$ between their variable pairs is true. On the contrary, for *any* possible kernel $p(T_1, T_2|T_3)$ there are *several* kernels $p(T_3|T_1, T_2)$ compatible with it, because the relation $(T_3|T_1, T_2) \preceq (T_1, T_2|T_3)$ between their variable pairs is false. It is remarkable that both answers can be given only considering the candidate variable pairs, irrespectively of the specific kernel chosen for the pair $(T_1, T_2|T_3)$. In general, for any fixed variable pair $(Y|X)$ in a lattice $\widetilde{O}(T)$, we may ask: which are the variable pairs $(V|U)$ such that, for *any* kernel $p(Y|X)$ there exists *exactly one* kernel $p(V|U)$ *compatible* with it? Proposition 3 answers that such variable pairs $(V|U)$ are precisely those such that $(V|U) \preceq (Y|X)$, that is, the members of the ideal generated by the member $(Y|X)$ within the lattice $\widetilde{O}(T)$.

5. Concluding remarks

The motivations for this study arose from the examination of probabilistic models that involve *several* variables and assign a prominent role to *conditional* probability distributions on them. Examples are provided by Bayesian networks, probabilistic graphical models, hierarchical Bayesian models in statistics, and Bayesian and Markov models in experimental sciences, which we mentioned in the Introduction along with few selected references to the corresponding literature. In models of these kinds, distinct *levels* of conditional probabilities are generally involved, in which the role played by any one variable in the system may vary. An elementary illustration of such duplicity or reversibility of roles is provided by the Bayes rule itself, as it intervenes in basic statistical models. Within the equation $p(\theta|D) = p(\theta) \cdot p(D|\theta)/p(D)$ that expresses the rule, the parameter quantity θ is a conditioning variable (placed on the right of the bar) in the likelihood term $p(D|\theta)$, and becomes a conditioned variable (placed on the left of the bar) in the posterior term $p(\theta|D)$. The opposite is true of the variable D representing the data.

With this study, we contribute ideas for a general framework in which the key components of such probabilistic models may be represented and interrelated for analysis. For representing the key components of a model, the comprehensive concept of "probability kernel" has been adopted, characterized as a family of probability functions on one (possible multiple) variable that are indexed by some other (possible multiple) variable. The analysis has been focused on such pairs of multiple variables and on the set-theoretic relations and operations applicable to them. This analysis is bent toward generality, as it is independent of the peculiar characteristics of the probability functions collected in a kernel. Working in this perspective, significant implications of algebraic character have been found, especially relating to the concept of a lattice.

Although focused on the variables, our analysis has meaningful consequences concerning the probability kernels acting on the given variables. Proposition 1 reveals a kind of algebraic structure into which the kernels in a complex probabilistic model may be mapped and interrelated through operations, in a way suggested by the Hasse diagrams in Figure 3. Proposition 2 provides a general criterion of compatibility between kernels, that makes it possible to test whether given low rank kernels may be the building blocks of a consistent probabilistic model. Proposition 3 provides a criterion for finding which kernels are uniquely determined by a given kernel and may thus be unambiguously deduced from it. The criterion is quite general and becomes strengthened when assumptions are introduced that specify the kind of probability functions forming the kernels in a model — for example, deducibility between kernels of Gaussian form.

References

[1] B. C. Arnold, E. Castillo, and J. M. Sarabia, *Conditional specification of statistical models*, Springer, 1999.

[2] B. C. Arnold, E. Castillo, and J. M. Sarabia, "Conditionally specified distributions: an introduction", *Statist. Sci.* **16**:3 (2001), 249–274.

[3] J.-M. Bernardo and A. F. M. Smith, *Bayesian theory*, Wiley, Chichester, 2000.

[4] P. Billingsley, *Probability and measure*, 3rd ed., Wiley, New York, 1995.

[5] A. Blake, P. Kohli, and C. Rother (editors), *Markov random fields for vision and image processing*, MIT Press, Cambridge, MA, 2011.

[6] A. Darwiche, *Modeling and reasoning with Bayesian networks*, Cambridge University Press, 2009.

[7] A. P. Dawid, "Conditional independence in statistical theory", *J. Roy. Statist. Soc. Ser. B* **41**:1 (1979), 1–31.

[8] A. Gelman, J. B. Carlin, H. S. Stern, D. B. Dunson, A. Vehtari, and D. B. Rubin, *Bayesian data analysis*, 3rd ed., CRC Press, Boca Raton, FL, 2014.

[9] D. Kersten, P. Mamassian, and A. Yuille, "Object perception as Bayesian inference", *Annu. Rev. Psychol.* **55** (2004), 271–304.

[10] U. B. Kjærulff and A. L. Madsen, *Bayesian networks and influence diagrams: a guide to construction and analysis*, Springer, 2008.

[11] D. Koller and N. Friedman, *Probabilistic graphical models: principles and techniques*, MIT Press, Cambridge, MA, 2009.

[12] T. Koski and J. M. Noble, *Bayesian networks: an introduction*, Wiley, Chichester, 2009.

[13] S. L. Lauritzen, *Graphical models*, Oxford Statistical Science Series **17**, Oxford University Press, New York, 1996.

[14] S. P. Meyn and R. L. Tweedie, *Markov chains and stochastic stability*, Springer, 1993.

[15] J. Pearl, *Probabilistic reasoning in intelligent systems: networks of plausible inference*, Morgan Kaufmann, San Mateo, CA, 1988.

[16] D. Pollard, *A user's guide to measure theoretic probability*, Cambridge Series in Statistical and Probabilistic Mathematics **8**, Cambridge University Press, 2002.

[17] J. N. Rouder, R. D. Morey, and M. S. Pratte, "Bayesian hierarchical models of cognition", pp. 504–551 in *New handbook of mathematical psychology*, vol. 1, edited by W. H. Batchelder et al., Cambridge University Press, 2017.

[18] M. Studený, *Probabilistic conditional independence structures*, Springer, 2005.

Received 2018-09-06. Revised 2019-09-13. Accepted 2019-10-09.

LUIGI BURIGANA: luigi.burigana@unipd.it
Department of General Psychology, University of Padua, I-35131 Padova, Italy

MICHELE VICOVARO: michele.vicovaro@unipd.it
Department of General Psychology, University of Padua, I-35131 Padova, Italy

Algebraic Statistics
Vol. 11, No. 1, 2020
https://doi.org/10.2140/astat.2020.11.99

MINIMAL EMBEDDING DIMENSIONS OF CONNECTED NEURAL CODES

RAFFAELLA MULAS AND NGOC M. TRAN

A receptive field code is a recently proposed deterministic model of neural activity patterns in response to stimuli. The main question is to characterize the set of realizable codes, and their minimal embedding dimensions with respect to a given family of receptive fields. Here we answer both of these questions when the receptive fields are connected. In particular, we show that all connected codes are realizable in dimension at most three. To our knowledge, this is the first family of receptive field codes for which both the exact characterization and minimal embedding dimension are known.

1. Introduction

A *receptive field code* is a deterministic model of neural activity patterns in response to stimuli defined by Curto, Itskov, Veliz-Cuba and Youngs [5]. It consists of $n \in \mathbb{N}$ neurons, each neuron $i \in [n] = \{1, 2, \ldots, n\}$ has a receptive field $U_i \subseteq \mathbb{R}^d$. Given a stimulus $x \in \mathbb{R}^d$, the neurons generate a codeword $\sigma(x) \in 2^{[n]}$ via

$$i \in \sigma(x) \iff x \in U_i. \tag{1}$$

A receptive field code $\mathcal{C}(\mathcal{U}) \subseteq 2^{[n]}$ is the set of all possible codewords generated from the collection of receptive fields $\mathcal{U} = (U_1, \ldots, U_n)$. Without loss of generality, we can assume that every receptive field code includes the empty set ($\varnothing \in \mathcal{C}$), which is equivalent to assuming that $\bigcup_{i\in[n]} U_i \subsetneq \mathbb{R}^d$.

The *minimal embedding problem* is the following: given a code $\mathcal{C} \subseteq 2^{[n]}$ and a family $\mathcal{F} = (\mathcal{F}_d, d \geq 1)$, where $\mathcal{F}_d$ is a collection of sets in $\mathbb{R}^d$, find the smallest d such that $\mathcal{C} = \mathcal{C}(\mathcal{U})$ for some $\mathcal{U} \subseteq \mathcal{F}_d$. Call this smallest d the minimal embedding dimension of the code $\mathcal{C}$ with respect to the family $\mathcal{F}$, denoted $d^*(\mathcal{C}, \mathcal{F})$.

This paper focuses on *connected codes*. These are codes realizable by connected sets in $\mathbb{R}^d$, for some $d \in \mathbb{N}$, which are either all closed or all open, following the convention of [5]. Our main results completely characterize realizability and minimal embedding dimensions of connected codes. In particular, it is easy to check if a code is connected, and if it is, then the minimal embedding dimension is at most three. We state our main results below. Characterization of the minimal embedding dimension for the case $d^* = 2$ is given in terms of the graph of a family of connected sets $\mathcal{U}$ (cf. Definition 9).

Proposition 1 (Realizability of connected codes). A code $\mathcal{C}$ is connected if and only if for each $\sigma, \tau \in \mathcal{C}$ and for each $i \in \sigma \cap \tau$, there exists a sequence of distinct codewords $\nu_1, \ldots, \nu_m \in \mathcal{C}$ such that:

• $\sigma = \nu_1$; • $\nu_j \subseteq \nu_{j+1}$ or $\nu_{j+1} \subseteq \nu_j$, for every $j \in [m-1]$; • $\nu_m = \tau$; • $i \in \nu_j$ for each $j \in [m]$.

MSC2010: 92B20, 05C65, 05C10.
Keywords: none, receptive field code, minimal embedding, realizability.

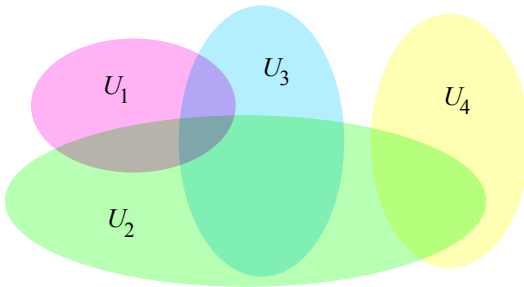


Figure 1. Four receptive fields generating the connected code $\mathcal{C}(\mathcal{U}) = \{\varnothing, 1, 2, 3, 4, 12, 13, 23, 24, 123\}$.

Theorem 2 (minimal embedding of connected codes). Suppose $\mathcal{C}$ is a connected code. Let $d^*(\mathcal{C})$ denote its minimal embedding dimension with respect to the family of connected sets.

- $d^*(\mathcal{C}) = 1$ if and only if the sensor graph of $\mathcal{C}$ is bipartite [22].
- Else, $d^*(\mathcal{C}) = 2$ if and only if there exists a realization $\mathcal{C}(\mathcal{U}) = \mathcal{C}$ by connected sets $\mathcal{U}$ in dimension 3 such that the graph of $\mathcal{U}$ is planar.
- Else, $d^*(\mathcal{C}) = 3$.

1.1. ***Related literature.*** The minimal embedding dimension $d^*(\cdot, \mathcal{F})$ and the family $\mathcal{F}$ of receptive fields form a trade-off in measuring the complexity of the signal encoded by the neurons, and is thus of particular interest in receptive field coding. In the extreme case where $\mathcal{F}_d$ is the collection of all sets in $\mathbb{R}^d$, then any code is realizable in dimension one [5].

Lemma 3. Let $\mathcal{C}$ be a code on n neurons. For any $d \geq 1$, $\mathcal{C} = \mathcal{C}(\mathcal{U})$ for some sequence of sets $\mathcal{U}$ in $\mathbb{R}^d$.

There has been a number of work on criteria for realizability and bounds for $d^*(\cdot, \mathcal{F})$ when the set $\mathcal{F}$ consists of (open or closed) convex sets [3; 6; 4; 5; 10; 11; 12; 13; 15; 16; 17; 22; 18]. However, complete characterization and the exact minimal embedding dimension of convex codes remain a problem. Giusti and Itskov [12] found necessary conditions for a code to be realizable with open convex sets, and proved lower bounds on the embedding dimensions of such codes. In [3], Cruz, Giusti, Itskov and Kronholm proved that there exists a family of codes, called max-intersection-complete codes, that are both open convex and closed convex, and they gave an upper bound for their embedding dimension. To the best of our knowledge, connected codes form the first family of receptive field codes for which an intrinsic characterization and the exact embedding dimension is known. Furthermore, our proof gives explicit constructions for the code realization in each dimension.

There are biologically observed receptive fields which are connected but not convex. A prominent example are the center-surround fields in the ganglion cells on the retina [7; 1], which approximately have the shape of a torus in dimension two. In some other cases, such as the grid cells in entorhinal cortex, the receptive field of a single cell is neither convex nor connected [14]. Even amongst place cells of the hippocampus, which generally have a single convex receptive field, it is shown that certain cells have disconnected receptive fields [20]. Such a code with disconnected receptive fields may be realized as the

projection of a connected code in higher dimensions. In comparison, these codes cannot be realized as projections of convex codes in any dimension, because such projections preserve convexity. This gives us motivations for studying connected codes.

In dimension 1, connected codes equal convex codes. This was completely characterized by Rosen and Zhang [22] via the sensor graph and it is included in Theorem 2 for completeness. We do not define the sensor graph of a code here, but note that it is an intrinsic property of the code, independent of any realization $\mathcal{U}$.

Jeffs [16, Theorem 4.1] gave an alternative characterization of connected codes, which is distinct from our approach as it is formulated in terms of the links in the code, that we do not define here.

Receptive field codes are closely related to Euler diagrams, which found applications in information systems, statistics and logic [8; 21]. Since their main applications are in visualization, the literature on Euler diagrams focus exclusively on 2 and 3 dimensions. Translated to our setup, an Euler diagram in $\mathbb{R}^3$ is a collection $\mathcal{U} = (U_1, \dots, U_n)$ of closed, orientable surfaces embedded in $\mathbb{R}^3$. An Euler diagram in $\mathbb{R}^2$ is a similar collection of closed curves embedded in $\mathbb{R}^2$. A diagram description is a code $\mathcal{C}$ such that $\varnothing \in \mathcal{C}$. The description of an Euler diagram $\mathcal{U}$ is the code $\mathcal{C}(\mathcal{U}^\circ)$, where $\mathcal{U}^\circ := (U_1^\circ, \dots, U_n^\circ)$ consists of the relative interior of the sets U_i's. The main problem in this literature is realizability: given a code $\mathcal{C}$, is there an Euler diagram $\mathcal{U}$ such that $\mathcal{C} = \mathcal{C}(\mathcal{U}^\circ)$? Every code $\mathcal{C}$ can be realized by an Euler diagram in dimension 2 [21], and by an Euler diagram in dimension 3 with connected sets U_i's [9]. Note the crucial difference to receptive field codes: in an Euler diagram, codewords are generated by intersection of the relative interior of the U_i's. In particular, all codes $\mathcal{C}$ which fail the condition of Proposition 1 satisfy $\mathcal{C} = \mathcal{C}(\mathcal{U}^\circ)$ for some tuple of closed connected sets $\mathcal{U}$ in $\mathbb{R}^3$, but $\mathcal{C} \neq \mathcal{C}(\mathcal{U})$ for any tuple of closed connected sets in any dimension.

2. Proof of the main results

We shall first give one definition and state a lemma that can be found in [5] and that will be useful for the proof of our main results.

Definition 4. Let $\mathcal{C}$ be a code on n neurons. We say that $\mathcal{C}$ is realizable by an atom sequence $\mathcal{A} = (A_\sigma \subseteq \mathbb{R}^d, \sigma \subseteq [n])$ if $A_\sigma \neq \varnothing \iff \sigma \in \mathcal{C}$. In this case, write $\mathcal{C} = \mathcal{C}(\mathcal{A})$.

Lemma 5. Let $\mathcal{C}$ be a code on n neurons. Then $\mathcal{C} = \mathcal{C}(\mathcal{A})$ if and only if $\mathcal{C} = \mathcal{C}(\mathcal{U})$, where

$$U_i = \bigcup_{i \in \sigma} A_\sigma, \tag{2}$$

or equivalently,

$$A_\sigma = \Big(\bigcap_{i \in \sigma} U_i \Big) \setminus \bigcup_{j \notin \sigma} U_j, \tag{3}$$

with the convention that $A_\varnothing = \mathbb{R}^d \setminus \bigcup_{i \in [n]} U_i$.

In other words, $\mathcal{A}$ and $\mathcal{U}$ determine each other.

Lemma 6. A code is realizable with closed connected sets in $\mathbb{R}^d$ if and only if it is realizable with open connected sets in $\mathbb{R}^d$.

Proof. Given $\mathcal{C} = \mathcal{C}(\mathcal{U})$ where $\mathcal{U} = (U_1, \ldots, U_n)$ is a family of closed connected sets in $\mathbb{R}^d$, we can always construct a family of open connected sets in $\mathbb{R}^d$, $\hat{\mathcal{U}} = (\hat{U}_1, \ldots, \hat{U}_n)$, such that $U_i \subset \hat{U}_i$ for each $i \in [n]$. The fact that the U_i's are contained in the $\hat{U}_i$'s implies that the old intersections are preserved and, furthermore, we can take the $\hat{U}_i$'s small enough to avoid the formation of new atoms. In this way we get a family of open connected sets in $\mathbb{R}^d$ with $\mathcal{C}(\mathcal{U}) = \mathcal{C}(\hat{\mathcal{U}})$.

Vice versa, if $\mathcal{C} = \mathcal{C}(\mathcal{U})$ where $\mathcal{U} = (U_1, \ldots, U_n)$ is given by open connected sets, we can let $\tilde{\mathcal{U}} = (\tilde{U}_1, \ldots, \tilde{U}_n)$ be a family of closed connected sets in $\mathbb{R}^d$ such that $\tilde{U}_i \subset U_i$ for each $i \in [n]$. The fact that the $\tilde{U}_i$'s are contained in the U_i's implies that no new intersections are created, and we can take the $\tilde{U}_i$'s big enough to preserve the old atoms. In this way we get a family of closed connected sets in $\mathbb{R}^d$ such that $\mathcal{C}(\mathcal{U}) = \mathcal{C}(\tilde{\mathcal{U}})$. □

Example 7. As an example of how Lemma 6 works, let $\mathcal{U}$ consists of three closed line segments in the plane meeting at a "T", so that $\mathcal{C}(\mathcal{U}) = \{\varnothing, 1, 2, 3, 123\}$. Since we are in $\mathbb{R}^2$, open sets $\hat{U}_i$'s containing the line segments U_i's must be full-dimensional. Therefore we can take open rectangles $\hat{U}_i$'s containing the old U_i's, such that they intersect all together but not pairwise. In this way we have that $\mathcal{C}(\mathcal{U}) = \mathcal{C}(\hat{\mathcal{U}})$.

Definition 8. We say that two sets $A, B \subset \mathbb{R}^d$ are adjacent if $A \cap B = \varnothing$ and either $\overline{A} \cap B \neq \varnothing$ or $A \cap \overline{B} \neq \varnothing$, where $\overline{A}$ denotes the closure of A in the Euclidean topology.

Definition 9 (graph of a realization). Let $\mathcal{C} = \mathcal{C}(\mathcal{U})$ be a connected code with realization $\mathcal{U}$. Let $\mathcal{A}$ be the atoms defined via $\mathcal{U}$ in (3). The graph of $\mathcal{U}$, denoted $\mathbb{G}(\mathcal{U})$, is a graph with one vertex for every connected component of each atom A_σ with $\sigma \neq \varnothing$, and an edge for every pair of connected components of atoms that are adjacent.

Lemma 10. Let $\mathcal{C} = \mathcal{C}(\mathcal{U})$ be a connected code with realization $\mathcal{U}$ in dimension d. If $\mathbb{G}(\mathcal{U})$ can be embedded in $\mathbb{R}^{d'}$, then $\mathcal{C}$ can also be realized by connected sets in dimension d'.

Proof. Take an embedding of $\mathbb{G}(\mathcal{U})$ in $\mathbb{R}^{d'}$. Let $\mathcal{A}$ be the atoms defined via $\mathcal{U}$ in (3). Let $v_{\sigma j} \in \mathbb{R}^{d'}$ be the realization of the vertex of $\mathbb{G}(\mathcal{U})$ indexed by the j-th component of the atom A_σ. For each pair of nodes $v_{\sigma j}$ and $v_{\tau k}$, let $e_{\sigma j, \tau k} \subset \mathbb{R}^{d'}$ be the realization of the edge between these nodes. If they are not connected, define $e_{\sigma j, \tau k} = \varnothing$. Now define atoms $\mathcal{A}'$ in $\mathbb{R}^{d'}$ via

$$A'_\sigma := \bigcup_j \left(v_{\sigma j} \cup \bigcup_{\tau k : \sigma \subset \tau} e_{\sigma j, \tau k} \right) \subset \mathbb{R}^{d'},$$

for $\sigma \in \mathcal{C} \backslash \{\varnothing\}$, and

$$A'_\varnothing := \mathbb{R}^{d'} \backslash \bigcup_{\sigma \in \mathcal{C} \backslash \{\varnothing\}} A'_\sigma.$$

It is easy to check that $\mathcal{C} = \mathcal{C}(\mathcal{A}')$, so $\mathcal{C}$ is realizable in dimension d', as needed. □

2.1. ***Proof of Proposition 1.*** Let $\mathcal{C}$ be a code on n neurons. By Lemma 3, $\mathcal{C} = \mathcal{C}(\mathcal{U}) = \mathcal{C}(\mathcal{A})$ for some $U_i, A_\sigma \subseteq \mathbb{R}^d$, $i \in [n]$, $\sigma \subseteq [n]$. For each $i \in [n]$, U_i is connected if and only if for every $\sigma, \tau \subseteq [n]$ such that $i \in \sigma \cap \tau$, from each connected component C_σ of A_σ to each connected component C_τ of A_τ there is a path $C_\sigma = C_{\nu_1} \to C_{\nu_2} \ldots \to C_{\nu_{m-1}} \to C_{\nu_m} = C_\tau$ in $\mathbb{G}(\mathcal{C}(\mathcal{U}))$, where $C_{\nu_j} \subseteq A_{\nu_j}$ is a connected component of A_{ν_j}, such that $A_{\nu_j} \subseteq U_i$ for each $j \in [m]$. Note that, in order to have the receptive fields either all open or all closed, two connected components $C_\sigma \subseteq A_\sigma$, $C_\tau \subseteq A_\tau$ are allowed to be adjacent if and only if either $\sigma \subseteq \tau$ or $\tau \subseteq \sigma$. Hence U_i is allowed to be connected if and only if for every σ, τ such that $i \in \sigma \cap \tau$, there exists a sequence of distinct codewords $\nu_1, \ldots, \nu_m \in \mathcal{C}$ such that:

- $\sigma = \nu_1$,
- either $\nu_j \subseteq \nu_{j+1}$ or $\nu_{j+1} \subseteq \nu_j$, for every $j \in [m-1]$,
- $\nu_m = \tau$ and
- $i \in \nu_j$ for each $j \in [m]$.

This proves the proposition. □

2.2. ***Proof of Theorem 2.*** We split the statement of Theorem 2 into two parts, and prove them separately. The first part, Proposition 11 states that the minimal embedding dimension for a connected code is at most 3. The second part, Proposition 13 gives a characterization for connected codes with $d^* = 2$. For the case $d^* = 1$, see [22, Proposition 1.9 and Theorem 3.4].

Proposition 11. Let $\mathcal{C}$ be a connected code on n neurons. Then $\mathcal{C}$ is realizable by connected sets in dimension 3.

Proof. For all $\sigma \in \mathcal{C}\backslash\{\varnothing\}$, choose disjoint balls $B_\sigma \subset \mathbb{R}^3$ and for all $\sigma, \tau \in \mathcal{C}$ such that $\sigma \subset \tau$, let $T_{\sigma,\tau} \subset \mathbb{R}^3$ be a tube that connects B_σ and B_τ. Since we are in $\mathbb{R}^3$, the tubes can always be arranged so that they do not intersect with each other and this can be proved by induction the number of tubes. Given m disjoint tubes between $|\mathcal{C}\backslash\{\varnothing\}|$ balls, suppose we need to construct a tube $T_{\sigma,\tau}$ joining the balls B_σ and B_τ. Since the number m of existing tubes is finite, we can pick a point $s \in B_\sigma$ and a point $t \in B_\tau$ such that their projections in the $(0, 0, 1)$ direction is larger than that of any other point on the m existing tubes. Now join s and t by a tube $T_{\sigma,\tau}$ such that its projection onto the $(0, 0, 1)$ direction is larger than that of all other tubes. Thus, $T_{\sigma,\tau}$ is disjoint from the first m tubes, completing the induction argument. Now, let

$$A_\sigma := B_\sigma \cup \bigcup_{\sigma \subset \tau} T_{\sigma,\tau}$$

for $\sigma \in \mathcal{C}\backslash\{\varnothing\}$ and let

$$A_\varnothing := \mathbb{R}^3 \setminus \bigcup_{\sigma \in \mathcal{C}\backslash\{\varnothing\}} A_\sigma.$$

Then $\mathcal{C} = \mathcal{C}(\mathcal{A})$. Define $\mathcal{U}$ from $\mathcal{A}$ via (2). By Lemma 5, $\mathcal{C} = \mathcal{C}(\mathcal{U})$. By construction of $\mathcal{A}$ and since we are assuming that $\mathcal{C}$ satisfies Proposition 1, the U_i's are connected. This completes the proof. □

Remark 12. The proof of Proposition 11 uses a classical technique that has been used for example also in [3, Lemma 2.2], [16, Lemma 4.4] and [9, Lemma 4.1].

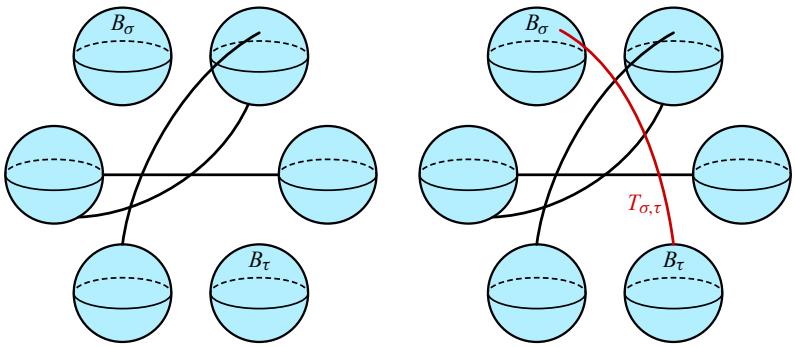


Figure 2. Illustration of the construction in Proposition 11. Given m disjoint tubes between $|\mathcal{C}\backslash\{\varnothing\}|$ balls (picture on the left), construct $T_{\sigma,\tau}$ such that its projection onto the $(0, 0, 1)$ direction is larger than that of all other tubes (right).

Proposition 13. Let $\mathcal{C}$ be a connected code on n neurons. Then $d^*(\mathcal{C}) = 2$ if and only if there exists a realization $\mathcal{C} = \mathcal{C}(\mathcal{U})$ by connected sets in $\mathbb{R}^3$ such that $\mathbb{G}(\mathcal{U})$ is planar.

Proof. Suppose $d^*(\mathcal{C}) = 2$. Then there exists a realization $\mathcal{C} = \mathcal{C}(\mathcal{U})$ with $\mathcal{U}$ a collection of connected sets $\mathcal{U}$ in $\mathbb{R}^2$. The graph of $\mathcal{U}$, $\mathbb{G}(\mathcal{U})$, is by construction also embedded in $\mathbb{R}^2$. One can trivially embed a realization in $\mathbb{R}^2$ into $\mathbb{R}^3$ without changing the graph $\mathbb{G}(\mathcal{U})$, so we are done. Conversely, suppose that $\mathcal{C} = \mathcal{C}(\mathcal{U})$ for some $\mathcal{U}$ in $\mathbb{R}^3$ such that $\mathbb{G}(\mathcal{U})$ is planar. By Lemma 10, $\mathcal{C}$ is realizable in dimension 2. □

We conclude our paper with two examples.

Example 14 (connected code with $d^* = 3$). Consider the following code

$$\mathcal{C} = \{\varnothing, 1, 2, 3, 4, 5, 12, 13, 14, 15, 23, 24, 25, 34, 35, 45\}. \tag{4}$$

This satisfies Proposition 1, so $\mathcal{C}$ is a connected code. It's easy to see that every graph $\mathbb{G}(\mathcal{U})$ associated to this code must be a subdivision of the graph in Figure 3; i.e., if $\mathcal{C} = \mathcal{C}(\mathcal{U})$, then $\mathbb{G}(\mathcal{U})$ must be either the graph in in Figure 3 or it can be obtained from it by subdividing some of its edges into two new edges, which must be connected to a new vertex. This is due to the fact that $\mathcal{C}$ is the code that contains exactly every $i \in [5]$ and every pair ij of distinct $i, j \in [5]$. This implies, by Kuratowski's theorem [2], that every graph associated to $\mathcal{C}$ is not planar. By Theorem 2, $\mathcal{C}$ has minimal embedding dimension 3.

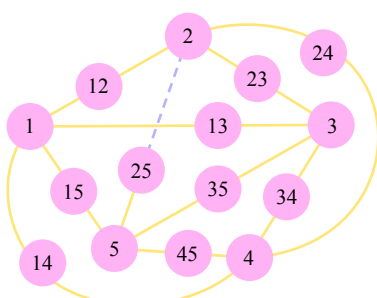


Figure 3. The graph of a connected realization of a code $\mathcal{C}$ with $d^*(\mathcal{C}) = 3$ in Example 14.

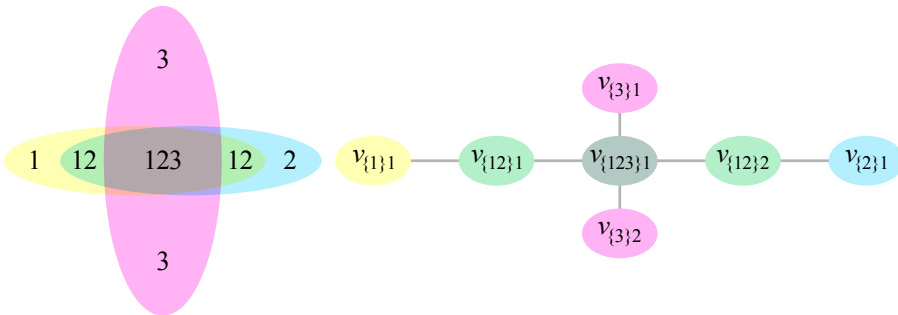


Figure 4. The realization of a connected code $\mathcal{C}$ with $d^*(\mathcal{C}) = 2$ in Example 15 and its graph.

Example 15 (connected code with $d^* = 2$). Let $\mathcal{C} = \{\varnothing, 1, 2, 3, 12, 123\}$ be a code on 3 neurons. By Proposition 1, this code is connected. Figure 4 shows its realization by connected sets in $\mathbb{R}^2$, and the corresponding graph. We claim that the minimal embedding dimension of this code is 2. One could verify by computing the sensor graph of $\mathcal{C}$. Alternatively, note that for the code to be realizable by connected sets, we must have $U_1 \cap U_2 \cap U_3 \neq \varnothing$ and U_i can not be contained in U_j for every $i, j \in [3]$, $i \neq j$. This is clearly not possible in dimension 1.

3. Open directions

In practice, neural firing is stochastic. One could incorporate noise to the receptive field code by replacing the deterministic equation (1) with some parametrization of the firing probability $\mathbb{P}(i \in \sigma(x) | x \in U_i)$. To be well-defined, this model needs further specifications, such as the distribution of the signal on $\mathbb{R}^d$. In this formulation, the minimal embedding dimension is a difficult and poorly formed statistical problem. Furthermore, it is clear that the minimal embedding dimension depends heavily on such details. However, underlying such models the assumption that there is a set of true receptive fields $\mathcal{U}$. Knowing the minimal embedding dimension for the deterministic model ensures that the neuroscientists do not have excessively many parameters, which can lead to ill-defined estimation problems. From this view, Theorem 2 states that if the true receptive fields are only required to be connected, one can assume that they are in dimension 3.

Apart from connected and convex sets, there are many biologically relevant models for receptive fields. Finding the minimal embedding dimension of receptive field codes realizable by any given family is an interesting and challenging problem. To be concrete, we propose another simple family motivated by observations from neuroscience. In experiments, one often encounter a group of neurons which all have the same receptive field up to translation, such as the retinal ganglion cells, head direction cells [7], place cells and grid cells [19; 14]. This corresponds to the case where $\mathcal{F}_d$ consists of all possible translations of some set $S \subset \mathbb{R}^d$. We call this the *shift* code. Thus, a concrete open problem is: which shift codes can be realized, and what would be their minimal embedding dimensions?

Acknowledgments

The authors thank the anonymous referee for constructive comments.

References

[1] M. R. Blackburn, "A simple computational model of center-surround receptive fields in the retina", technical report ADA264723, Defense Technical Information Center, 1993, https://apps.dtic.mil/sti/citations/ADA264723.

[2] J. A. Bondy and U. S. R. Murty, *Graph theory*, Graduate Texts in Mathematics **244**, Springer, 2008.

[3] J. Cruz, C. Giusti, V. Itskov, and B. Kronholm, "On open and closed convex codes", *Discrete Comput. Geom.* **61**:2 (2019), 247–270.

[4] C. Curto, "What can topology tell us about the neural code?", *Bull. Amer. Math. Soc.* (*N.S.*) **54**:1 (2017), 63–78.

[5] C. Curto, V. Itskov, A. Veliz-Cuba, and N. Youngs, "The neural ring: an algebraic tool for analyzing the intrinsic structure of neural codes", *Bull. Math. Biol.* **75**:9 (2013), 1571–1611.

[6] C. Curto, E. Gross, J. Jeffries, K. Morrison, M. Omar, Z. Rosen, A. Shiu, and N. Youngs, "What makes a neural code convex?", *SIAM J. Appl. Algebra Geom.* **1**:1 (2017), 222–238.

[7] P. Dayan and L. F. Abbott, *Theoretical neuroscience: computational and mathematical modeling of neural systems*, MIT Press, Cambridge, MA, 2001.

[8] J. Flower, A. Fish, and J. Howse, "Euler diagram generation", *J. Vis. Lang. Comput.* **19**:6 (2008), 675–694.

[9] J. Flower, G. Stapleton, and P. Rodgers, "On the drawability of 3D Venn and Euler diagrams", *J. Vis. Lang. Comput.* **25**:3 (2014), 186–209.

[10] M. K. Franke and M. Hoch, "Investigating an algebraic signaturefor max intersection-complete codes", REU report, Texas A&M University, 2017, http://see-math.math.tamu.edu/ugs/research/REU/results/REU_2017/Hoch.pdf.

[11] M. Franke and S. Muthiah, "Every binary code can be realized by convex sets", *Adv. in Appl. Math.* **99** (2018), 83–93.

[12] C. Giusti and V. Itskov, "A no-go theorem for one-layer feedforward networks", *Neural Comput.* **26**:11 (2014), 2527–2540.

[13] E. Gross, N. Obatake, and N. Youngs, "Neural ideals and stimulus space visualization", *Adv. in Appl. Math.* **95** (2018), 65–95.

[14] T. Hafting, M. Fyhn, S. Molden, M.-B. Moser, and E. I. Moser, "Microstructure of a spatial map in the entorhinal cortex", *Nature* **436** (2005), 801–806.

[15] M. Hoch, S. Muthiah, and N. Obatake, "On the identification of k-inductively pierced codes using toric ideals", preprint, 2018. arXiv 1807.02390

[16] R. A. Jeffs, *Convexity of neural codes*, undergraduate thesis, Harvey Mudd College, 2016, https://scholarship.claremont.edu/hmc_theses/87/.

[17] R. A. Jeffs, M. Omar, and N. Youngs, "Homomorphisms preserving neural ideals", *J. Pure Appl. Algebra* **222**:11 (2018), 3470–3482.

[18] C. Lienkaemper, A. Shiu, and Z. Woodstock, "Obstructions to convexity in neural codes", *Adv. in Appl. Math.* **85** (2017), 31–59.

[19] J. O'Keefe, "Place units in the hippocampus of the freely moving rat", *Exp. Neurol.* **51**:1 (1976), 78–109.

[20] E. Park, D. Dvorak, and A. A. Fenton, "Ensemble place codes in hippocampus: CA1, CA3, and dentate gyrus place cells have multiple place fields in large environments", *PLOS ONE* **6**:7 (2011), art. id. e22349.

[21] P. Rodgers, "A survey of Euler diagrams", *J. Vis. Lang. Comput.* **25**:3 (2014), 134–155.

[22] Z. Rosen and Y. X. Zhang, "Convex neural codes in dimension 1", preprint, 2017. arXiv 1702.06907

Received 2017-10-12. Revised 2018-07-13. Accepted 2019-02-18.

RAFFAELLA MULAS: raffaella.mulas@gmail.com
MPI MiS-Leipzig, Germany

NGOC M. TRAN: ntran@math.utexas.edu
University of Texas at Austin, Austin, TX, United States

and

Hausdorff Center for Mathematics, Bonn, Germany

Guidelines for Authors

Authors may submit manuscripts in PDF format on-line at the submission page.

Originality. Submission of a manuscript acknowledges that the manuscript is original and and is not, in whole or in part, published or under consideration for publication elsewhere. It is understood also that the manuscript will not be submitted elsewhere while under consideration for publication in this journal.

Language. Articles are usually in English or French, but articles written in other languages are welcome.

Required items. A brief abstract of about 150 words or less must be included. It should be self-contained and not refer to bibliography keys. If the article is not in English, two versions of the abstract must be included, one in the language of the article and one in English. Also required are keywords and a Mathematics Subject Classification for the article, and, for each author, affiliation (if appropriate) and email address.

Format. Authors are encouraged to use LaTeX and the standard amsart class, but submissions in other varieties of TeX, and exceptionally in other formats, are acceptable. Initial uploads should normally be in PDF format; after the refereeing process we will ask you to submit all source material.

References. Bibliographical references should be complete, including article titles and page ranges. All references in the bibliography should be cited in the text. The use of BibTeX is preferred but not required. Tags will be converted to the house format, however, for submission you may use the format of your choice. Links will be provided to all literature with known web locations and authors are encouraged to provide their own links in addition to those supplied in the editorial process.

Figures. Figures must be of publication quality. After acceptance, you will need to submit the original source files in vector graphics format for all diagrams in your manuscript: vector EPS or vector PDF files are the most useful.

Most drawing and graphing packages — Mathematica, Adobe Illustrator, Corel Draw, MATLAB, etc. — allow the user to save files in one of these formats. Make sure that what you are saving is vector graphics and not a bitmap. If you need help, please write to graphics@msp.org with as many details as you can about how your graphics were generated.

Bundle your figure files into a single archive (using zip, tar, rar or other format of your choice) and upload on the link you been provided at acceptance time. Each figure should be captioned and numbered so that it can float. Small figures occupying no more than three lines of vertical space can be kept in the text ("the curve looks like this:"). It is acceptable to submit a manuscript with all figures at the end, if their placement is specified in the text by means of comments such as "Place Figure 1 here". The same considerations apply to tables.

White Space. Forced line breaks or page breaks should not be inserted in the document. There is no point in your trying to optimize line and page breaks in the original manuscript. The manuscript will be reformatted to use the journal's preferred fonts and layout.

Proofs. Page proofs will be made available to authors (or to the designated corresponding author) at a Web site in PDF format. Failure to acknowledge the receipt of proofs or to return corrections within the requested deadline may cause publication to be postponed.

Algebraic Statistics

2020 11 : 1