Journal of Symbolic Computation 45 (2010) 1296–1315

Contents lists available at ScienceDirect

Journal of Symbolic Computation

journal homepage: www.elsevier.com/locate/jsc

An implicitization challenge for binary factor analysis

María Angélica Cueto[a,1], Enrique A. Tobis[b], Josephine Yu[c]

[a] Department of Mathematics, University of California, Berkeley. 970 Evans Hall #3840, Berkeley, CA 94720-3840, USA

[b] Departamento de Matemática, FCEN - Universidad de Buenos Aires, Pabellón I - Ciudad Universitaria, C1428EGA, Buenos Aires, Argentina

[c] School of Mathematics, Georgia Institute of Technology, Atlanta, GA 30332, USA

ARTICLE INFO

Article history:
Received 26 October 2009
Accepted 4 May 2010
Available online 22 June 2010

Keywords:
Factor analysis
Tropical geometry
Hadamard products
Newton polytope

ABSTRACT

We use tropical geometry to compute the multidegree and Newton polytope of the hypersurface of a statistical model with two hidden and four observed binary random variables, solving an open question stated by Drton, Sturmfels and Sullivant in (Drton et al., 2009, Ch. VI, Problem 7.7). The model is obtained from the undirected graphical model of the complete bipartite graph $K_{2,4}$ by marginalizing two of the six binary random variables. We present algorithms for computing the Newton polytope of its defining equation by parallel walks along the polytope and its normal fan. In this way we compute vertices of the polytope. Finally, we also compute and certify its facets by studying tangent cones of the polytope at the symmetry classes of vertices. The Newton polytope has 17 214 912 vertices in 44 938 symmetry classes and 70 646 facets in 246 symmetry classes.

Published by Elsevier Ltd

1. Introduction

In recent years, a fruitful interaction between (computational) algebraic geometry and statistics has emerged, under the form of algebraic statistics. The main objects studied by this field are probability distributions that can be described by means of polynomial or even rational maps. Among them, an important source of examples are the so called graphical models. In this paper, we focus our attention on a special model: the *undirected* $(4, 2)$-*binary factor analysis model* $\mathcal{F}_{4,2}$.

E-mail addresses: macueto@math.berkeley.edu (M.A. Cueto), etobis@dc.uba.ar (E.A. Tobis), josephine.yu@math.gatech.edu (J. Yu).

URLs: http://math.berkeley.edu/~macueto/ (M.A. Cueto), http://www.tobis.com.ar/ (E.A. Tobis), http://people.math.gatech.edu/~jyu67/ (J. Yu).

[1] Tel.: +1 510 642 6550; fax: +1 510 642 8204.

0747-7171/$ – see front matter. Published by Elsevier Ltd
doi:10.1016/j.jsc.2010.06.011

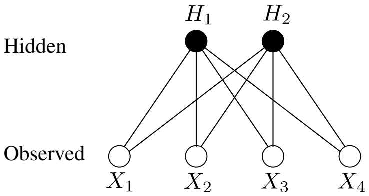

Fig. 1. The model $\mathcal{F}_{4,2}$. Each node represents a binary random variable.

First, let us describe our main player. Consider the complete undirected bipartite graph $K_{2,4}$ with four *observed* nodes X_1, X_2, X_3, X_4 and two *hidden* nodes H_1, H_2 (cf. Fig. 1). Each node represents a binary random variable and each edge represents a dependency between two random variables. In other words, if there is no edge between two random variables, then they are conditionally independent given the rest of the variables. We obtain a hidden model from this undirected graphical model by marginalizing over H_1 and H_2. This model is the discrete undirected version of the factor analysis model discussed in (Drton et al., 2009, Section 4.2). The model and its immediate generalization $\mathcal{F}_{m,n}$ are closely related to the statistical model describing the behavior of restricted Boltzmann machines (Le Roux and Bengio, 2008), which are widely discussed in the machine learning literature. Here, $\mathcal{F}_{m,n}$ is the binary undirected graphical model with n hidden variables and m observed variables encoded in the complete bipartite graph $K_{m,n}$. The main invariant of interest in these models is the expected dimension, and, furthermore, lower bounds on n such that the probability distributions are a dense subset of the probability simplex Δ_{2^m-1}. By direct computation, it is easy to show that $\mathcal{F}_{2,2}$ and $\mathcal{F}_{3,2}$ are dense subsets of the corresponding probability simplices, so $\mathcal{F}_{4,2}$ is the first interesting example worth studying. Understanding the model $\mathcal{F}_{4,2}$ can pave the way for the study of restricted Boltzmann machines in general (Cueto et al., 2010).

The set of all possible joint probability distributions (X_1, X_2, X_3, X_4) that arise in this way forms a semialgebraic set $\mathcal{M}$ in the probability simplex Δ_{15}. To simplify our construction, we disregard the inequalities defining the model and we extend our parameterization to the entire affine space $\mathbb{C}^{16}$. In other words, we consider the Zariski closure of the joint probability distributions in $\mathbb{C}^{16}$. As a result of this, we obtain an algebraic subvariety of $\mathbb{C}^{16}$ which carries the core information of our model. In turn, we projectivize the model by considering its associated projective variety. This variety is expected to have codimension one and be defined by a homogeneous polynomial in 16 variables.

Problem (*An Implicitization Challenge, (Drton et al., 2009, Ch. VI, Problem7.7)*)**.** Find the degree and the defining polynomial of the model $\mathcal{M}$.

Our main results state that the variety $\mathcal{M}$ is a hypersurface of degree 110 in $\mathbb{P}^{15}$ (Theorem 15) and explicitly enumerate all vertices and facets of the polytope (Theorem 14). Our methods are based on tropical geometry. Since the polynomial is multihomogeneous, we get its *multidegree* from just one vertex. Interpolation techniques will allow us to compute the corresponding irreducible homogeneous polynomial in 16 variables, using the lattice points in the Newton polytope. However, this polytope will turn out to be too big for interpolation to be practically feasible.

The paper is organized as follows. In Section 2 we describe the parametric form of our model and we express our variety as the Hadamard square of the first secant of the Segre embedding $\mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \times \mathbb{P}^1 \hookrightarrow \mathbb{P}^{15}$. In Section 3 we present the tropical interpretation of our variety. By means of the nice interplay between the construction described in Section 2 and its tropicalization, we compute this tropical variety as a collection of cones with multiplicities. We should remark that we do not obtain a fan structure, but, nonetheless, our characterization is sufficient to fulfill the goal of the paper. The key ingredient is the computation of multiplicities by the so called push-forward formula (Sturmfels et al., 2007, Theorem 3.12) which we generalize to match our setting (Theorem 7). We finish Section 3 by describing the effective computation of the tropical variety and discussing some of the underlying combinatorics.

In Section 4 we compute the multidegree of our model with respect to a natural 5-dimensional grading, which comes from the tropical picture in Section 3. Once this question is answered, we shift gears and move to the study of the Newton polytope of our variety. We present two algorithms that