İstanbul Üniversitesi İşletme Fakültesi Dergisi
Istanbul University Journal of the School of Business Administration
Cilt/Vol:41, Sayı/No:2, 2012, 278-283
ISSN: 1303-1732 – www.ifdergisi.org © 2012

An invitation to algebraic statistics: New outlook and opportunities

Eyüp Çetin[1]
Department of Quantitative Methods,
School of Business Administration
Istanbul University, Istanbul, Turkey

Abstract

Algebra, a branch of pure mathematics, now advances statistics and operations research of applied mathematics. This synergy is called algebraic statistics as a new discipline. Algebraic statistics offers statisticians, management scientists, business researchers, econometricians and algebraists new opportunities, horizons and connections to advance their fields and related application areas. In this effort, this young, vibrant, quickly growing, and active discipline is briefly discussed and some major application areas are given and also an illustrative example is presented.

Keywords: *Statistics, Algebra, Operations Research, Optimization, Business Applications*

Cebirsel istatistiğe bir davet: Yeni bakış açısı ve olanaklar

Özet

Teorik matematiğin bir dalı olan cebir son zamanlarda uygumalı matematiğin dalları olan istatistik ve yöneylem araştırmasının gelişimine katkıda bulunmaktadır. Bu sinerji yeni bir disiplin şeklinde cebirsel istatistik olarak tanımlanmaktadır. Cebirsel istatistik istatistikçilere, yönetim bilimcilerine, işletme araştırmacılarına, ekonometrisyenlere ve cebircilere yeni ufuklar, olanaklar ve kendi alanlarını geliştirmeleri için işbirlikleri sunmaktadır. Bu çalışmada, bu genç, parlak, hızla gelişen ve aktif disiplin kısaca ele alınmış, bazı temel uygulama alanları ve tanımlayıcı bir örnek ele alınmıştır.

Anahtar Kelimeler: *İstatistik, Cebir, Yöneylem Araştırması, Optimizasyon, İşletme Uygulamaları*

1. Introduction

The synergistic syntheses of the disciplines have accelerated over the recent years. One of the most interesting mergers has been revealed in mathematical sciences as well. This interdisciplinary area is algebraic statistics. Algebraic statistics is at the intersection of algebra that includes theory of group, ring, field, module, lattice, representation and coding and statistics. Even though the name of this new discipline is algebraic statistics, it also covers the problems in operations research and optimization as well.

Algebraic statistics is a new paradigm and may open new horizons for algebraists, statisticians, operations and business researchers. In this effort, it is aimed to introduce the interdisciplinary field and point some opportunities. This paper is organized as follows; firstly the discipline of algebraic statistics is defined, then some application areas are presented briefly. The first and unique journal in the world entitled *Journal of Algebraic Statistics* is introduced as a milestone in the field. A representative example is discussed, and some conclusions are drawn.

[1] *eycetin@istanbul.edu.*tr *(E. Çetin)*

2. What is Algebraic Statistics?

Algebra, a branch of pure mathematics, now advances statistics and operations research of applied mathematics. This young, vibrant, quickly growing, and active discipline focused on the applications of algebraic geometry and its computational tools in statistical models [1], is entitled "algebraic statistics". More extensively, algebraic statistics is concerned with the development of techniques in algebraic geometry, commutative algebra, and combinatorics, to address problems in statistics and its applications [2]. It is clearly seen in the literature that algebraic statistics also deals with some operations research / management science issues, which are widely industrial engineering topics, such as Markov chains, optimization, and network analysis stemming from its philosophy.

Algebraic statistics, as a defined field of study, is a relatively new discipline that has developed and changed rather rapidly over the last fifteen years [2]. Two papers by Pistone and Wynn [3] and Diaconis and Sturmfels [4] set the foundations of algebraic statistics [5]. Thus, we can claim that algebraic statistics has two origins. For a detailed discussion on historical background of algebraic statistics, see [5].

By these motivations, many researchers from algebra, statistics, operations research, quantitative methods, engineering and some related application areas have been contributing to this important interdisciplinary field.

3. Application Areas

Algebraic statistics has been applied in disclosure limitation, design of experiments, graphical models, hypothesis testing for log-linear models, maximum likelihood estimation, and approximations to Bayesian integrals, and in the last decade it finds its applications in several other areas, such as computational biology, chemical networks, robotic inspection and finance. For example, the algebraic approach to phylogenetics has caused many applications in computational biology. Besides phylogenetics, this approach has seen biological applications in inferring the progression to drug resistance in HIV, determining the parametric behaviors of sequence alignments, and studying the geometry of fitness landscapes.

Example applications include mathematical modeling, engineering, operations research, optimization, model identification, system analysis and design, system verification, and system synthesis. Application areas also include systems biology, genomics, proteomics, and evolutionary biology, finance, engineering, to name a few. Areas in algebra include polynomial methods, commutative algebra and algebraic geometry, group theory, string rewriting, automated reasoning, automata theory, formal languages, combinatorics, graph theory, and artificial intelligence, among others [1,6]. For a detailed literature review see [7].

4. The Journal as a Milestone in the Field

The researchers in this emerging area have been publishing their works in statistics and mathematics journals. As Prof. Fabrizio Catanese*, Editor of the Journal of Algebraic Geometry and one of the frequent contributors to the field, says "the present moment seems a very appropriate one to launch a new journal on algebraic statistics"* [8], the first and unique journal in the world dedicated to algebraic statistics entitled *Journal of Algebraic Statistics* was fortunately founded by the author and Prof. Unsal Tekir, an algebraist, in 2009. The annual journal published its inaugural issue including some opening messages in June 2010. This international rostrum that welcomes papers at the

intersection of algebra, statistics and operations research/management science may be determined as a milestone in the advance of algebraic statistics.

5. An Illustrative Example

Statistics is mainly concerned with models, submodels and the relations between them. As a general rule, we consider models and submodels that can be specified by an *algebraic variety*, *Θ* with respect to some set of parameters with ideal denoted by Ideal (*Θ*). These sort of models are called *algebraic statistical models*. Being an algebraic model depends on the parameters used in the probability description. In particular, an algebraic model can be linear when the algebraic variety is an affine subspace of the space of parameters in the saturated model.

The most basic example consists in using as parameters the value of the density function itself, that is, we have a vector parameter $p = (p_i: i = 1...N)$ with

$$p_i = \mathrm{P}(a_i),\ \ a_i \subset D,\ \ \sum_i^N p_i = 1$$

where a_i , *i = 1...N* is a numbering of design points, and *p* is restricted to some algebraic variety *Π* described by an ideal in the ring *k[p]*. The ideal of this variety is specified by giving equations in addition to the normalization condition.

Because the statistical models are described as algebraic varieties, the problem of finding a minimal set of free parameters, that is, a proper parameterization, could be discussed in the framework of the parametric (rational) representation of algebraic varieties [9].

The following example, appears in [9], is devoted to the illustration of sampling and sufficiency reduction regarding independent marginals when the polynomial form is used. The treatment of such items as independence and sufficiency are particularly interesting in the polynomial encoding.

Let us consider the simplest possible model of a two-dimensional sampling distribution, that is, two independent Bernoulli variables. We denote the success probabilities by *p*1, *p*2. The generic joint probability in polynomial form is

$$p(x, y; \theta) = \theta_{00} + \theta_{10}x + \theta_{01}y + \theta_{11}xy$$

with marginals

$$p_1(x; \theta) = \left(\theta_{00} + \frac{1}{2}\,\theta_{01}\right) + \left(\theta_{10} + \frac{1}{2}\,\theta_{11}\right)x$$

$$p_2(y; \theta) = \left(\theta_{00} + \frac{1}{2}\,\theta_{10}\right) + \left(\theta_{01} + \frac{1}{2}\,\theta_{11}\right)y$$

The ideal of the model is obtained by the normalizing equation and the four equations obtained by equating the coefficients in

$$p(x, y; \theta) = p_1(x; \theta)p_2(y; \theta)$$

namely,

$$\begin{cases} 4 = \theta_{00} + 2\theta_{10} + 2\theta_{01} + \theta_{11} \\ \theta_{00} = \left(\theta_{00} + \frac{1}{2}\,\theta_{01}\right)\left(\theta_{00} + \frac{1}{2}\,\theta_{10}\right) \\ \theta_{10} = \left(\theta_{00} + \frac{1}{2}\,\theta_{10}\right)\left(\theta_{10} + \frac{1}{2}\,\theta_{11}\right) \\ \theta_{01} = \left(\theta_{00} + \frac{1}{2}\,\theta_{01}\right)\left(\theta_{01} + \frac{1}{2}\,\theta_{11}\right) \\ \theta_{11} = \left(\theta_{10} + \frac{1}{2}\,\theta_{11}\right)\left(\theta_{01} + \frac{1}{2}\,\theta_{11}\right) \end{cases} \quad (1)$$

A proper parameterization can be given in terms of the success probability as

$$\begin{cases} p_1 = p_1(1;\theta) = \left(\theta_{00} + \frac{1}{2}\,\theta_{01}\right) + \left(\theta_{10} + \frac{1}{2}\,\theta_{11}\right) \\ p_2 = p_2(1;\theta) = \left(\theta_{00} + \frac{1}{2}\,\theta_{10}\right) + \left(\theta_{01} + \frac{1}{2}\,\theta_{11}\right) \end{cases} \quad (2)$$

Solving Equations (1) and (2) by reduction to triangular form for the monomial ordering *plex* with initial ordering

$$\theta_{00} \succ \theta_{10} \succ \theta_{01} \succ \theta_{11} \succ p_1 \succ p_2$$

we obtain

$$\begin{cases} -4\theta_{00} - 2\theta_{10} - 2\theta_{01} + 4 = 0, \\ -\frac{1}{2}\,\theta_{10} + \frac{1}{2}\,\theta_{01} + p_1 - p_2 = 0, \\ \frac{1}{2}\,\theta_{01} + 1/4\theta_{11} - p_2 + 1 = 0, \\ -\theta_{11} + 4p_1p_2 - 4p_1 - 4p_2 + 4 = 0 \end{cases}$$

From this we could solve for the θ's and by substitution obtain the properly parameterized version of the density.

Let us see what happens in the case $\theta = p1 = p2$. If we sum the sample space ideal, the model ideal and the probability ideal, we obtain the ideal generated by the polynomials

$$\begin{cases} x^2 - x, \\ y^2 - y, \\ 4 - (4\theta_{00} + 2\theta_{10} + 2\theta_{01} + \theta_{11}), \\ \theta_{00} - \left(\theta_{00} + \frac{1}{2}\,\theta_{01}\right)\left(\theta_{00} + \frac{1}{2}\,\theta_{10}\right), \\ \theta_{10} - \left(\theta_{00} + \frac{1}{2}\,\theta_{10}\right)\left(\theta_{10} + \frac{1}{2}\,\theta_{11}\right), \\ \theta_{01} - \left(\theta_{00} + \frac{1}{2}\,\theta_{01}\right)\left(\theta_{01} + \frac{1}{2}\,\theta_{11}\right), \\ \theta_{11} - \left(\theta_{10} + \frac{1}{2}\,\theta_{11}\right)\left(\theta_{01} + \frac{1}{2}\,\theta_{11}\right), \\ p - (\theta_{00} + \theta_{10}x + \theta_{01}y + \theta_{11}xy) \end{cases}$$

A Gröbner basis is

$$\begin{cases} -4\theta_{00} - 2\theta_{10} - 2\theta_{01} - \theta_{11} + 4, \\ -\frac{1}{2}\,\theta_{10} + \frac{1}{2}\,\theta_{01}, \\ \frac{1}{2}\,\theta_{01} + 1/4\theta_{11} - \theta + 1, \\ \theta_{11} + 4\theta^2 - 8\theta + 4, \\ 4xy\theta^2 - 8xy\theta + 4xy - 2x\theta^2 + 6x\theta - \\ 4x - 2y\theta^2 + 6y\theta - 4y + \theta^2 - \theta - p + 4 \end{cases}$$

The last polynomial is the representation of the probability density. In the corresponding equation, the part equal to *p* factors out as follows:

$$p = (2x\theta - 2x - \theta + 2)(2y\theta - 2y - \theta + 2)$$

If we order with respect to the powers of θ, we obtain the representation

$$\begin{aligned} p = \quad & (4xy - 4x - 4y + 4) + \\ & (-8xy + 6x + 6y - 1)\theta + \\ & (4xy - 2x - 2y + 1)\theta^2 \end{aligned}$$

here the polynomial coefficients are a set of sufficient statistics. Actually an elementary analysis shows that $T(x+y) = x + y$ is a sufficient statistic because on *D* we have $(x+y)^2 = (x+y) + 2xy$, that is $xy = \frac{1}{2}T(x+y)^2 - \frac{1}{2}T(x+y)$ [9].

6. Conclusions

In this paper, we highlight the new interdisciplinary field of algebraic statistics and its applications. Firstly, the field is defined, and some major application areas are given, then the journal of the field is introduced and finally an illustrative example is discussed.

Algebraic statistics as a fertile, quickly growing and synergistic field offers statisticians, operations researchers, management scientists, engineers, econometricians and algebraists new opportunities, horizons and connections to advance their fields and related application areas.

Algebraic statistics invites researchers interested to break boundaries by their close collaborations. It also offers a nice ground where algebraists, statisticians, operations researchers and related scientists can talk about the same problems. It seems we will hear algebraic statistics and its successes more frequently in the near future.

References

[1] U. Tekir, E. Çetin, R. Yoshida, S. Petrović, J. A. Howe, B. Kiremitci, Letter from the Editors, *Journal of Algebraic Statistics*, 1, 1, 1-2 (2010).

[2] M. Drton, B. Sturmfels, S. Sullivant, Lectures on Algebraic Statistics, Oberwolfach Seminars 39, Birkhäuser (2009).

[3] G. Pistone, H.P. Wynn, Generalised Confounding with Gröbner Bases, *Biometrika*, 83, 3, 653–666 (1996).

[4] P. Diaconis, B. Sturmfels, Algebraic Algorithms for Sampling From Conditional Distributions, *Annals of Statistics*, 26,1, 363–397 (1998).

[5] E. Riccomagno, A Short History of Algebraic Statistics, Metrika, 69, 2-3, 397-418 (2010).

[6] JA De Loera, R Hemmecke, M Köppe, Algebraic and Geometric Ideas in the Theory of Discrete Optimization, MOS-SIAM Series on Optimization, available in December 2012.

[7] M. Drton, S. Sullivant, Algebraic Statistical Models, *Statististica Sinica*, 17, 4, 1273–1297 (2007).

[8] Editorial Messages, *Journal of Algebraic Statistics*, 1, 1, 3-5 (2010).

[9] G. Pistone, E. Riccomagno, H.P. Wynn, Algebraic Statistics: Computational Commutative Algebra in Statistics, Chapman & Hall/CRC, USA, p. 120-124, 2001.