week 8 day 2

Exact testing for model/data fit for log-linear models
Part four
Algebraic & Geometric Methods in Statistics

Sonja Petrović
Created for Math/Stat 561

Mar 1, 2023.

Reminder

- Definition of Markov bases (recall from Lecture 13)

MB definition

Given: A, any two tables u, v for which $Au = Av$
Markov basis $= \{b_1 \ldots, b_n\} \subset \ker A$
* there exists a choice of basis vectors satisfying

$$u + b_{i_1} + \cdots + b_{i_N} = v,$$

* *each* partial sum must result in a non-negative vector.

Before we state the fundamental theorem of MB, let's look at two examples

- Note: we will discuss the meaning of *all* terms in the theorem!

2-way tables

Let $A : \mathbb{Z}^{k_1 \times k_2} \to \mathbb{Z}^{k_1+k_2}$ defined by

$$\begin{aligned} A(u) &= \left(u_{1+}, \ldots, u_{k_1+}; u_{+1}, \ldots, u_{+k_2}\right) \\ &= \text{vector of row and column sums of } u \end{aligned}$$

$$\ker_{\mathbb{Z}}(A) = \left\{u \in \mathbb{Z}^{k_1 \times k_2} \mid \text{row and column sums of } u \text{ are } 0\right\}$$

Markov basis consists of the $2\binom{k_1}{2}\binom{k_2}{2}$ moves like

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

Luis García-Puente (SHSU) Algebraic Statistics 2015 16 / 29

Figure 1: Example from Garcia-Puente

3-way tables

Let $A : \mathbb{Z}^{k_1 \times k_2 \times k_3} \to \mathbb{Z}^{k_1k_2+k_1k_3+k_2k_3}$ defined by

$$A(u) = \left(\left(\sum_{i_3} u_{i_1i_2i_3} \right)_{i_1,i_2} ; \left(\sum_{i_2} u_{i_1i_2i_3} \right)_{i_1,i_3} ; \left(\sum_{i_1} u_{i_1i_2i_3} \right)_{i_2,i_3} \right)$$

$$= \text{all 2-way margins of the 3-way table } u$$

Markov basis depends on k_1, k_2, k_3, contains moves like:

$$\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix} \quad \begin{pmatrix} -1 & 1 \\ 1 & -1 \end{pmatrix}$$

but also non-obvious moves like:

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

The weight of the evidence: conditional p-value

Prob(observing v at least as 'extreme' as u | given marginals Au).

- compute all such tables
- give each a *score* : $\chi^2(u) = \sum_{ij} \frac{(u_{ij} - E_{ij})^2}{E_{ij}}$, where $E_{ij} = \mathbb{E}(u_{ij})$.
- count the fraction *more extreme* than u.

	M	F	Other	
≤1,2M	1	9	3	13
>1,2M	9	1	3	13
	10	10	6	26

$$\chi^2 = \frac{4^2}{5} + \frac{4^2}{5} + 0 + \frac{4^2}{5} + \frac{4^2}{5} + 0 = 12.8$$

	M	F	Other	
≤1,2M	5	5	3	13
>1,2M	5	5	3	13
	10	10	6	26

$$\chi^2 = 0$$

	M	F	Other	
≤1,2M	9	2	2	13
>1,2M	1	8	4	13
	10	10	6	26

$$\chi^2 = \frac{4^2}{5} + \frac{3^2}{5} + \frac{1^2}{3} + \frac{3^2}{5} + \frac{3^2}{5} + \frac{1^2}{3} = 10.667$$

	M	F	Other	
≤1,2M	10	1	2	13
>1,2M	0	9	4	13
	10	10	6	26

$$\chi^2 = \frac{5^2}{5} + \frac{4^2}{5} + \frac{1^2}{3} + \frac{5^2}{5} + \frac{4^2}{5} + \frac{1^2}{3} = 17.0667$$

The Fundamental Theorem of Markov bases (FTMB)

Theorem (Diaconis-Sturmfels, AOS '98)

A set of moves is a **Markov basis** for the log-linear model A **if and only if** the corresponding set of binomials is a **generating set** of the ideal I_A.

$$\begin{array}{|c|c|c|}\hline 1 & -1 & 0 \\ \hline -1 & 1 & 0 \\ \hline 0 & 0 & 0 \\ \hline \end{array}\,, \quad \begin{array}{|c|c|c|}\hline -1 & 0 & 1 \\ \hline 0 & 0 & 0 \\ \hline 1 & 0 & -1 \\ \hline \end{array}\,, \quad \begin{array}{|c|c|c|}\hline 0 & 0 & 0 \\ \hline 0 & 1 & -1 \\ \hline 0 & -1 & 1 \\ \hline \end{array}\,.$$

$x_{11}x_{22} - x_{12}x_{21}$, $x_{13}x_{31} - x_{11}x_{33}$, $x_{22}x_{33} - x_{23}x_{32}$.

`Macaulay2`: toricMarkov(A)

Do we know how to compute this ideal?

[What is... a Markov basis?, AMS Notices, August 2019]

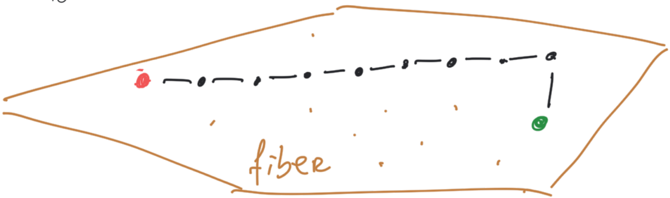
	M	F	Other	
≤1,2M	1	9	3	13
>1,2M	9	1	3	13
	10	10	6	26

+1	-1	0
-1	+1	0

0	-1	+1
0	+1	-1

+8	-7	-1
-8	+7	+1

	M	F	Other	
≤1,2M	9	2	2	13
>1,2M	1	8	4	13
	10	10	6	26



The algebra

Definition

Let $A : \mathbb{Z}^n \to \mathbb{Z}^d$. The **toric ideal** I_A is the ideal

$$\langle p^u - p^v \mid u, v \in \mathbb{N}^n, Au = Av \rangle \subset \mathbb{K}[p_1, \ldots, p_n],$$

where $p^u = p_1^{u_1} p_2^{u_2} \cdots p_n^{u_n}$.

Theorem (Diaconis-Sturmfels 1998)

*The set of moves $\mathcal{B} \subset \ker_{\mathbb{Z}} A$ is a **Markov basis** for A if and only if the set of binomials $\{p^{b^+} - p^{b^-} \mid b \in \mathcal{B}\}$ generates I_A.*

$$\begin{pmatrix} 0 & 0 & 0 & 0 \\ 1 & 0 & -1 & 0 \\ -1 & 0 & 1 & 0 \end{pmatrix} \longrightarrow p_{21}p_{33} - p_{23}p_{31}$$

Metropolis-Hastings

- Show example from Garcia-Puente: slides 22-23

Link to slides 22-23

- Show algorithm from Danai G: slide 27

Summary: testing goodness of fit of a model

Goal

Test model goodness of fit ("Model validation problem' ')

- Given: candidate model $\mathcal{P}$ + *one* g_{obs},
- decide (w/ high degree of confidence) whether g_{obs} can be regarded as a draw from some distribution $P_{\theta_0} \in \mathcal{P}$.

Requires:

- A valid GoF statistic (measure of distance between g_{obs} and P_{θ_0}).
- Distribution of GoF must not depend on unknown parameters

Conditioning on the sufficient statistics $t(g)$
$\implies$ distribution independent of parameters.

For log-linear models, Markov bases are use to sample from the conditional distribution given observed sufficient statistics.

Reading

- Markov bases and Metropolis-Hastings - that is the start of Section 9.2.
  - include example 201-202 culminating with Proposition 9.2.10.
  - look out for Felix's talk in april!

The usual. . . license

This document is created for Math/Stat 561, Spring 2023.