

Markov bases for noncommutative analysis of ranked data

Nicholas Eriksson

Department of Mathematics
University of California, Berkeley

2 March 2006
Joint work with Persi Diaconis
math.AC/0405060

Outline

Ranked data

Fourier analysis

Toric ideals

Statistics

Markov bases for
noncommutative
analysis of ranked
data

Nicholas Eriksson

Ranked data

Fourier analysis

Toric ideals

Statistics

Summary

The data

5738 voters each ranked 5 candidates for the APA
presidency.

Ranking	Votes	Ranking	Votes	Ranking	Votes
54321	29	43521	91	32541	41
54312	67	43512	84	32514	64
54231	37	43251	30	32451	34
54213	24	43215	35	32415	75
54132	43	43152	38	32154	82
54123	28	43125	35	32145	74
53421	57	42531	58	31542	30
53412	49	42513	66	31524	34

The first-order summary

The percentage of voters who ranked candidate i in position j .

Candidate	Rank				
	1	2	3	4	5
1	18.3	26.4	22.8	17.4	14.8
2	13.5	18.7	24.6	24.6	18.3
3	28.0	16.7	13.8	18.2	23.1
4	20.4	16.9	18.9	20.2	23.3
5	19.6	21.0	19.6	19.2	20.3

The Fourier transform

The Fourier transform of $f \in \mathbb{R}[G]$ at ρ

$$\hat{f}(\rho) = \sum_{g \in G} f(g)\rho(g).$$

The first-order summary is just the Fourier transform of our data $f \in \mathbb{R}[S_5]$ at the permutation representation.

Our data decomposes into irreps $f(g) = \sum_{\rho \in \hat{G}} f|_{V_\rho}(g)$,
with

$$f|_{V_\rho}(g) = \frac{d_\rho}{|G|} \sum_{h \in G} \chi_\rho(h) f(gh).$$

Representation theory

Markov bases for
noncommutative
analysis of ranked
data

Nicholas Eriksson

Ranked data

Fourier analysis

Toric ideals

Statistics

Summary

- ▶ The irreducible representations S^λ are indexed by partitions λ of n .
- ▶ The dimension of S^λ is the number of standard Young tableaux of shape λ .

The decomposition

ρ	S^5	$S^{4,1}$	$S^{3,2}$	$S^{3,1,1}$	$S^{2,2,1}$	S^{2111}	S^{11111}
d_ρ^2	1	16	25	36	25	16	1
Data	2286	298	459	78	27	7	0

Data is concentrated in S^5 , $S^{4,1}$, $S^{3,2}$. Is this a coincidence?

What do these representations mean?

Second order summary

The projection into $S^{3,2}$ in a natural basis.

	Rank									
	1,2	1,3	1,4	1,5	2,3	2,4	2,5	3,4	3,5	4,5
1,2	-137	-20	18	140	111	22	4	6	-97	-46
1,3	476	-88	-179	-209	-147	-169	-160	107	128	241
1,4	-189	51	113	24	-9	98	99	-65	23	-146
1,5	-150	57	47	45	43	49	56	-48	-53	-48
2,3	-42	84	19	-61	30	-16	82	-76	-39	72
2,4	157	-20	-43	-25	-93	-76	-56	8	38	112
2,5	22	-44	7	15	-117	69	25	62	99	-138
3,4	-265	-7	72	199	39	140	85	19	-52	-233
3,5	-169	10	88	70	78	44	47	-51	-36	-80
4,5	296	-24	-142	-130	-5	-163	-128	38	-9	267

Markov bases

A *Markov basis* for S_n is a finite subset of moves $g_1, \dots, g_B \in \mathbb{Z}[S_n]$ such that any two elements in $\mathbb{N}[S_n]$ with the same first order summary can be connected by a sequence of moves in that subset.

Markov bases

A *Markov basis* for S_n is a finite subset of moves $g_1, \dots, g_B \in \mathbb{Z}[S_n]$ such that any two elements in $\mathbb{N}[S_n]$ with the same first order summary can be connected by a sequence of moves in that subset.

Example (S_3)

$$\begin{bmatrix} 123 \\ 231 \\ 312 \end{bmatrix} - \begin{bmatrix} 132 \\ 213 \\ 321 \end{bmatrix}$$

It's a valid move since both elements have Fourier transform

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

The kernel of the map

$$\begin{aligned}\phi_{\mathcal{S}_n}: \mathbb{C}[x_\pi \mid \pi \in \mathcal{S}_n] &\longrightarrow \mathbb{C}[t_{ij} \mid 1 \leq i, j \leq n] \\ x_\pi &\longmapsto \prod_{1 \leq i, j \leq n} t_{i\pi(i)}\end{aligned}$$

is a toric ideal.

Example (\mathcal{S}_3)

$$I_{\mathcal{S}_3} = \langle x_{123}x_{231}x_{312} - x_{132}x_{213}x_{321} \rangle$$

Gröbner basis calculation — Singular

Markov bases for
noncommutative
analysis of ranked
data

Nicholas Eriksson

```
ring r=0, (t11,t12,t13,t21,t22,t23,t31,t32,t33,  
          x123,x132,x213,x231,x312,x321), dp;  
ideal i =  
x123 - t11*t22*t33,  
x132 - t11*t23*t32,  
x213 - t12*t21*t33,  
x231 - t12*t21*t32,  
x312 - t13*t23*t31,  
x321 - t13*t22*t31;  
  
eliminate(i,t11*t12*t13*t21*t22*t23*t31*t32*t33);  
  
_[1]=x123*x231*x312-x132*x213*x321
```

Ranked data

Fourier analysis

Toric ideals

Statistics

Summary

Example: S_4

Ideal has 178 minimal generators in three symmetry classes

$$\begin{array}{c} \begin{bmatrix} 1234 \\ 2143 \end{bmatrix} - \begin{bmatrix} 1243 \\ 2134 \end{bmatrix} \\ \begin{bmatrix} 2314 \\ 2431 \\ 4123 \end{bmatrix} - \begin{bmatrix} 2134 \\ 2413 \\ 4321 \end{bmatrix} \quad \begin{bmatrix} 1324 \\ 2134 \\ 3214 \end{bmatrix} - \begin{bmatrix} 1234 \\ 2314 \\ 3124 \end{bmatrix} \end{array}$$

How to compute a set of generators

1. Gröbner basis computation of $\ker(\phi_{S_n})$ — an ideal in $n!$ indeterminates.
2. Fiber-by-fiber calculation:

$$\rho: \mathbb{N}[S_n] \rightarrow \mathbb{N}^{n^2}$$

Theorem

A set of moves is a Markov basis if every fiber $\rho^{-1}(\mathbf{b})$ for $\mathbf{b} \in \mathbb{N}^{n^2}$ is connected using these moves.

The elements $\mathbf{b} \in \mathbb{N}^{n^2}$ with $\rho^{-1}(\mathbf{b})$ non-empty are the *magic squares*.

Degree bounds

Theorem (Diaconis-Sturmfels)

Every element of a reverse lexicographic Gröbner basis of I_{S_n} has degree at most n .

Theorem (Diaconis-Eriksson)

I_{S_n} is generated in degree $n - 1$ for $n \geq 4$.

Conjecture

Degree 3 suffices.

6×6 magic squares

Degree	Number
1	720
2	202410
3	20933840
4	1047649905
5	30767936616
6	602351808741

Too many squares to check!

Symmetry

The group $S_n \times S_n$ acts on \mathbb{N}^{n^2} by permuting rows and columns.

- ▶ Reduces the computation for S_6 to about 75000 cases.
- ▶ Lifts to a group action on the Markov basis of I_{S_n} .
- ▶ In terms of tableaux, one copy of S_n acts by permuting columns of the tableau, the other acts by permuting the labels in the tableau.
- ▶ The symmetrized bases are remarkably small - only 14 for S_5 and 60 for S_6 .

Markov bases

Markov bases for
noncommutative
analysis of ranked
data

Nicholas Eriksson

Ranked data

Fourier analysis

Toric ideals

Statistics

Summary

n	Degree 2		Degree 3		Degree ≥ 4	
	all	sym	all	sym	all	sym
3	0	0	1	1	0	0
4	18	1	160	2	0	0
5	1050	2	28840	12	0	0
6	57150	7	6694240	53	0	0
7	3567690	12	?	?	?	?

Exponential families

Exponential family of probability distributions on S_n :

$$P_{\Theta}(\pi) = Z^{-1} e^{\text{Tr}(\Theta\rho(\pi))},$$

where $Z = \sum_{\pi \in S_n} e^{\text{Tr}(\Theta\rho(\pi))}$.

A sufficient statistic for Θ based on data $f(\pi)$ is the Fourier transform $\hat{f}(\rho)$.

Statistical test

Null hypothesis: our data comes from the exponential family P_{Θ} .

To test goodness of fit, we want to sample from the conditional distribution of the data f given the sufficient statistic $\hat{f}(\rho)$:

$$P_{\Theta}(f|\hat{f}(\rho)) = w^{-1} \prod_{\sigma \in \mathcal{S}_n} \frac{1}{f(\sigma)!},$$

where

$$w = \sum_{\substack{g \in \mathbb{Z}[\mathcal{S}_n] \\ \hat{g}(\rho) = \hat{f}(\rho)}} \prod_{\sigma \in \mathcal{S}_n} \frac{1}{g(\sigma)!}.$$

We can sample from this hypergeometric distribution using the Markov basis and the Metropolis algorithm.

Random walks

The averages of the projection into the 7 isotypic subspaces of S_5 for 100 random draws for the data and 3 perturbations.

ρ	S^5	$S^{4,1}$	$S^{3,2}$	$S^{3,1,1}$	$S^{2,2,1}$	S^{2111}	S^{11111}
Data	2286	298	459	78	27	7	0
Hypergeom.	2286	298	16	19	10	6	0
Uniform	2286	298	511	672	436	295	25
Bootstrap	2286	303	469	93	37	13	1

The projection to $S^{3,2}$ is much larger in the data than in the hypergeometric distribution, affirming a strong rejection of the null model.

Thus we need to look at the structure of the higher order projection on its own terms.

Summary

Problems in statistics can lead to interesting problems in combinatorics and algebra.

Example

For any finite group and representation, we can define a toric ideal. Are these ideals nice?