

Markov basis for testing homogeneity of Markov chains

Akimichi Takemura and Hisayuki Hara

University of Tokyo

July 6, 2009

Table of contents

- 1 Introduction and Notation
- 2 Properties of the configuration and the toric ideal
- 3 MB for two-state case with arbitrary length
- 4 MB for arbitrary finite state space and length of three
- 5 MB for three-state, length of four
- 6 Discussions and conjectures

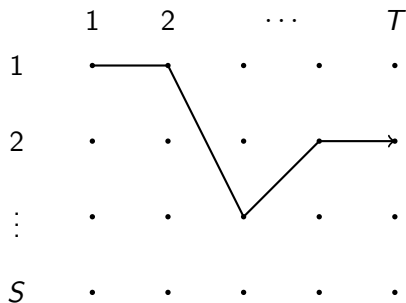
Introduction and Notation

Introduction and Notation

- We study Markov basis for testing homogeneity of Markov chains.
- Markov chain models are used in various fields, such as models for social mobility studies or state switching models in econometrics.
- We give a complete description of Markov basis for the following cases:
 - two-state, arbitrary length.
 - arbitrary finite state space and length of three.
 - three-state, length of four (already very hard!)
- The general case remains to be a conjecture.

Introduction and Notation

- Consider a Markov chain X_t , $t = 1, \dots, T$ (≥ 3), over a finite state space $\mathcal{S} = \{1, \dots, S\}$ ($S \geq 2$).
- Each observed Markov chain is a path $\omega = (s_1, \dots, s_T) \in \mathcal{S}^T$.



Introduction and Notation

- Let $p_{ij}^{(t)} = P(X_{t+1} = j \mid X_t = i)$ denote the transition probability from time t to $t + 1$.
- Let $\{\pi_i\}$ denote the initial distribution of X_1 .
- The probability of a path $\omega = (s_1, s_2, \dots, s_T)$ is written as

$$p(\omega) = \pi_{s_1} p_{s_1 s_2}^1 \cdots p_{s_{T-1} s_T}^{T-1}. \quad (1)$$

- We consider the null hypothesis of homogeneity:

$$H : p_{ij}^{(t)} = p_{ij}, \quad t = 1, \dots, T - 1. \quad (2)$$

Introduction and Notation

- Then the probability of the path is

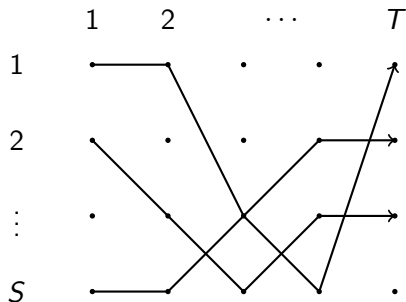
$$p(\omega) = \pi_{s_1} p_{s_1 s_2} \cdots p_{s_{T-1} s_T},$$

which is a toric model (i.e. the right hand side is a monomial).

- So we can test H by a Markov basis approach.
- The parameters of the model under H is the set of transition probabilities $\{p_{ij}\}_{i,j \in \mathcal{S}}$ and the initial distribution $\{\pi_i\}_{i \in \mathcal{S}}$.

Introduction and Notation

- Suppose that we observe a (multi)set of N paths $W = \{\omega_1, \dots, \omega_N\}$ of the Markov chain.



Introduction and Notation

- We identify the set of paths W with a T -way contingency table $\mathbf{x} = \{x(\omega), \omega \in \mathcal{S}^T\}$, where $x(\omega)$ denotes the frequency of the path ω in W .
- We are going to count the number of transitions in W .
- x_{ij}^t : the number of transitions from $s_t = i$ to $s_{t+1} = j$ in W .
- x_i^t : the frequency of the state $s_t = i$ in W .
- In particular x_i^1 is the frequency of the initial state $s_1 = i$.
- The set of the numbers of transitions $\{x_{ij}^t\}$, $i, j \in \mathcal{S}$, $t = 1, \dots, T - 1$ forms a sufficient statistic for non-homogeneous model (conditional independence model)

$$p(\omega) = \pi_{s_1} p_{s_1 s_2}^1 \cdots p_{s_{T-1} s_T}^{T-1}.$$

Introduction and Notation

- Let

$$x_{ij}^+ = \sum_{t=1}^{T-1} x_{ij}^t$$

denote the total number of transitions from i to j in W (ignoring the time t).

- The sufficient statistic under H is given by

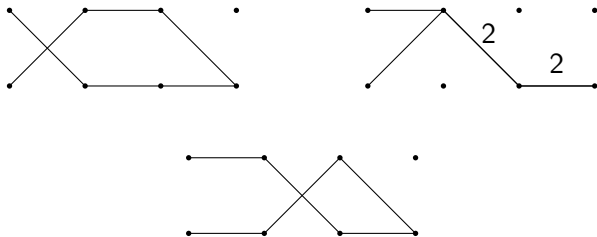
$$\mathbf{b} = \mathbf{b}(\mathbf{x}) = \{x_s^1, s \in \mathcal{S}\} \cup \{x_{ij}^+, i, j \in \mathcal{S}\}.$$

- Fiber:

$$\mathcal{F}_{\mathbf{b}} = \{\mathbf{x} \in \mathbb{N}^{S^T} \mid \mathbf{b}(\mathbf{x}) = \mathbf{b}\}.$$

Introduction and Notation

- Example of elements of a fiber:



- We want to obtain a set of “moves” for transforming a set of paths to another set in the same fiber: **Markov basis**.

Properties of the configuration and the toric ideal

“homogeneous Markov chain toric ideal”
(HMC toric ideal)

Properties of HMC toric ideal

- For illustration we write out the configuration A for $S = 2$ and $T = 4$.


	1111	1112	1121	1122	1211	1212	1221	1222	2111	2112	2121	2122	2211	2212	2221	2222
	1	1	1	1	1	1	1	1	0	0	0	0	0	0	0	0
	0	0	0	0	0	0	0	0	1	1	1	1	1	1	1	1
"11"	3	2	1	1	1	0	0	0	2	1	0	0	1	0	0	0
"12"	0	1	1	1	1	2	1	1	0	1	1	1	0	1	0	0
"21"	0	0	1	0	1	1	1	0	1	1	2	1	1	1	1	0
"22"	0	0	0	1	0	0	1	2	0	0	0	1	1	1	2	3

- The columns 1121 and 1211 are identical!

Properties of HMC toric ideal

•

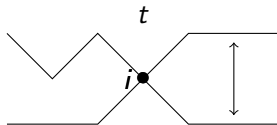


(alternative figure: )

- We need to consider degree 1 moves.
- We could choose just one column from A among identical columns. But it is not clear which column to choose.

Properties of HMC toric ideal

- “Crossing path swapping”



- Data with crossing path swappings look the same:



corresponds either to $\{111, 212\}$ or to $\{112, 211\}$.

Properties of HMC toric ideal

- **FACT:** crossing path swappings are the square-free degree two moves for the non-homogeneous model (linearly ordered conditional independence model):

$$p(\omega) = \pi_{s_1} p_{s_1 s_2}^1 \cdots p_{s_{T-1} s_T}^{T-1}.$$

(π_{s_1} can be absorbed into $p_{s_1 s_2}^1$.)

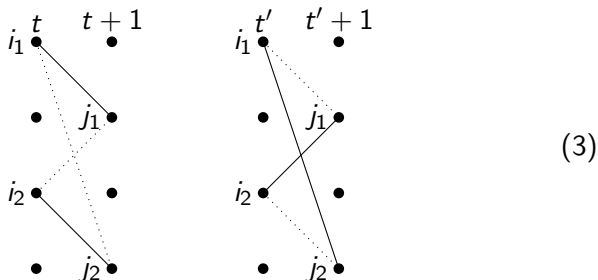
Proposition 1 (Dobra)

The set of crossing path swappings constitutes a Markov basis for the linearly ordered conditional independence model.

- (Roughly) we will be identifying all the data sets with the same picture.

Properties of HMC toric ideal

- Example of a move: “2 by 2 swap”



- (Roughly again) Note that this “move” corresponds to a collection of moves, because in the picture the other parts of the paths are not specified.

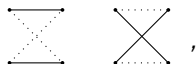
Markov basis for two-state case (arbitrary length)

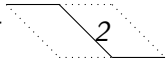
MB for two-state case

Theorem 1

A Markov basis for $S = 2$, $T \geq 4$, consists of the following moves.

- ① cross path swappings,
- ② degree one moves,
- ③ 2 by 2 swaps of the following form:
- ④ moves of the following form



(the case of  2 included).

MB for two-state case

- For $S = 2$, the complexity of Markov basis is the “same” for arbitrary T . (Looks somewhat like “Markov complexity” result.)
- Actually for $T = 3$, we do not need the fourth type.
- Moves are sums of at most two “loops”.
- Proof of the theorem is not too difficult, but not too easy.

Markov basis for arbitrary finite state space and length of three

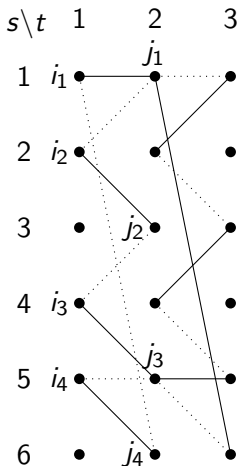
MB for $T = 3$

- Here we give an explicit form of a Markov basis for the case of $T = 3$. The number of states $S = |\mathcal{S}|$ is arbitrary.
- Let $t:ij$ denote the transition from i to j at time t to $t + 1$.
- For $T = 3$ we only need to consider $t = 1, 2$.
- Let i_1, \dots, i_m , $m \leq S$, be distinct elements of \mathcal{S} . Similarly let j_1, \dots, j_m , $m \leq S$, be distinct elements of \mathcal{S} .
- Define a move (a “permutation”) $Z(i_1, \dots, i_m; j_1, \dots, j_m)$ by

$$Z(i_1, \dots, i_m; j_1, \dots, j_m) : \\ \{1:i_1j_1, 1:i_2j_2, \dots, 1:i_mj_m\} \leftrightarrow \{2:i_1j_m, 2:i_2j_1, \dots, 2:i_mj_{m-1}\}. \quad (4)$$

MB for $T = 3$

A typical move for $S = 6$ with $m = 4$, $(i_1, i_2, i_3, i_4) = (1, 2, 4, 5)$,
 $(j_1, j_2, j_3, j_4) = (1, 3, 5, 6)$.



MB for $T = 3$

Theorem 2

A Markov basis for HMC toric ideal with $T = 3$ is given by the set of crossing path swappings and moves corresponding to m times m permutations $Z(i_1, \dots, i_m; j_1, \dots, j_m)$ in (4), where $m = 2, \dots, S$, i_1, \dots, i_m are distinct, and j_1, \dots, j_m are distinct.

Proof of this theorem is “easy”.

Markov basis for three-state, length of four

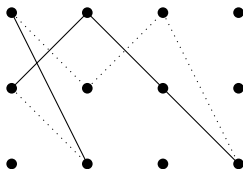
MB for $S = 3, T = 4$

- We wanted to check the case $S = 3, T = 4$, assuming that it is not too complicated.
- However this case turned out to be incredibly difficult.
- We now have a theorem, but the proof at the moment exists in my hand-written memo of 150 pages with hundreds of pictures.
- To state our theorem we need some more definitions.

MB for $S = 3, T = 4$

An extended simple loop

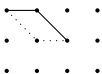

- Intuitively, it is a loop, such that when we are moving towards the future we follow positive edges (solid lines) and when we are moving towards the past we follow negative edges (dotted lines).
- Also we require that each node is passed at most once.
- An important example of an extended simple loop for $S = 3$ and $T = 4$ is depicted as follows.

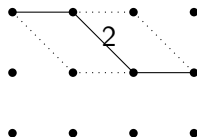


MB for $S = 3, T = 4$

Sign-conformal sum of extended simple loops

- Two extended simple loops L_1, L_2 are *sign-conformal* if each edge $t:ij$ belonging to both L_1 and L_2 has the same sign in two loops.
- A sign-conformal sum of two extended simple loops L_1, L_2 is a graph where each edge is weighted by a non-zero integer.
- The weight of an edge is given by its sign in L_1 and the number of L_k 's containing the edge.

- Example: sign-conformal sum of  and  is



MB for $S = 3, T = 4$

- Now our summary result for 3×4 case is stated as follows.

Theorem 3

A Markov basis for $S = 3$ and $T = 4$ is given by the set of crossing path swappings and the set of moves which are sign-conformal sums of at most three extended simple loops.

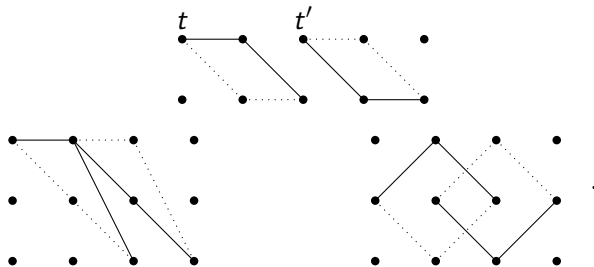
- Actually this statement is composed in hindsight. We are actually proving the following theorem:

Theorem 4

A Markov basis for $S = 3$ and $T = 4$ consists of moves in the following list (16 types of moves).

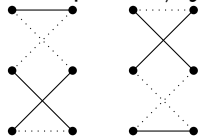
MB for $S = 3, T = 4$

- 0 [Type 0] Crossing path swapping.
- 1 [Type 1, "Deg1"] Degree one moves.
- 2 [Type 2, "2:2swap"] 2 by 2 swaps of (3).
- 3 [Type 3, "1*1swap"] Degree 2 move of partial path swapping with $m = 1$ in the following picture:

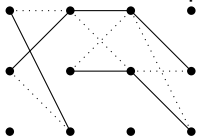


MB for $S = 3, T = 4$

- ④ [Type 4, “3:3 permutation”] Since $S = 3$, the 3 by 3 permutation is written as $\{t:1i_1, t:2i_2, t:3i_3\} \leftrightarrow \{t':1j_1, t:2j_2, t:3j_3\}$ where $1 \leq t < t' \leq 4$ and both $\{i_1, i_2, i_3\}$ and $\{j_1, j_2, j_3\}$ are permutations of $\{1, 2, 3\}$. Further, for this move to be degree three, we require $i_1 \neq j_1$, $i_2 \neq j_2$ and $i_3 \neq j_3$. An example of this move is

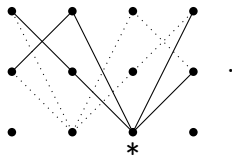


- ⑤ [Type 5, “2:2:2 swap”] Sum of three loops of length 4:

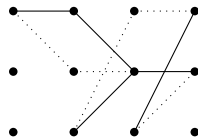
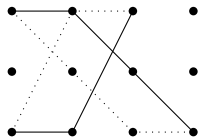
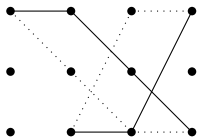


MB for $S = 3, T = 4$

- 6 [Type 6, "2 by 2 swap of unequal length", "2:2 uneq"] A swap of partially specified partial paths for 3×4 case such as

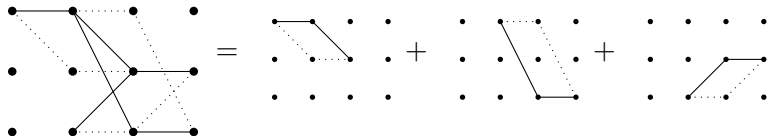


- 7 [Type 7, "2+3loop", sum of loops of length 4 and 6] There are three types, 7A, 7B, 7C:

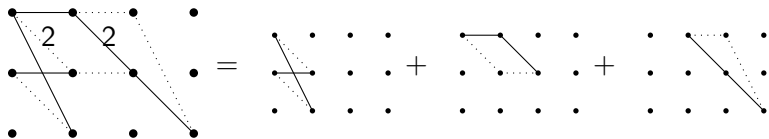


MB for $S = 3, T = 4$

8 [Type 8, "3loops", sum of three loops of length 4]

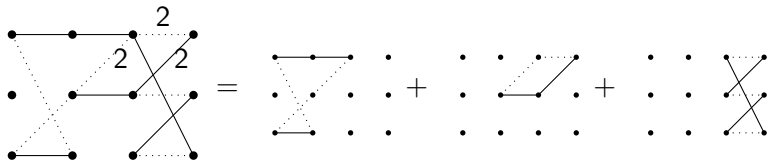


9 [Type 9, "3loopsW", sum of three loops of length 4 with overlapped edges]

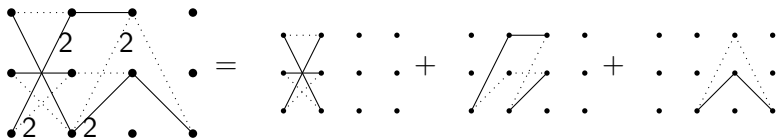


MB for $S = 3, T = 4$

- 10 [Type 10, "2+3+3loopW3", sum of three loops of lengths 4, 6 and 6 with 3 overlapped edges]

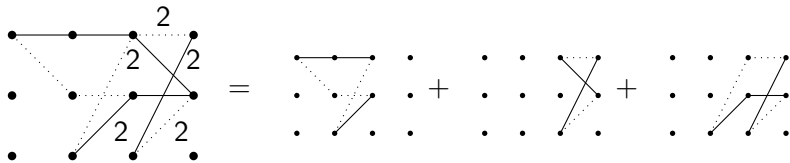


- 11 [Type 11, "2+2+3loopW4", sum of three loops of lengths 4, 6 and 6 with 4 overlapped edges]

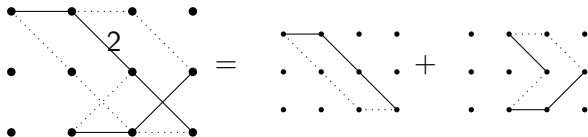


MB for $S = 3, T = 4$

- 12 [Type 12, "2+3+3loopW5", sum of three loops of lengths 4, 6 and 6 with 5 overlapped edges]

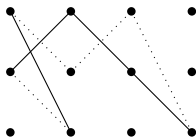


- 13 [Type 13, "3+3loop", sum of two loops of length 6]

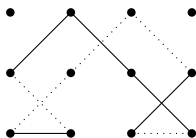


MB for $S = 3, T = 4$

- 14 [Type 14, "LL1", a long extended simple loop of length 8]



- 15 [Type 15, "LL2", a long extended simple loop of length 10]



Discussions and conjectures

Discussions and conjectures

- From the case of $S = 2$, we expect that there exists some kind of “Markov complexity”, i.e., the complexity of MB does not depend on T .
- (Bold) Conjecture: For general \mathcal{S} , MB consists of at most $|\mathcal{S}|$ sign-conformal sums of extended simple loops.
- Somehow, roughly speaking, the moves for “pictures” look very similar to “toric ideals” with respect to these pictures.

Bibliography I

- Akimichi Takemura and Hisayuki Hara (2009). Markov chain Monte Carlo test of homogeneity of Markov chains. In preparation.