Theory and Applications of Random Partition Processes

Harry Crane

Department of Statistics
Rutgers University

October 11, 2012
Random partitions
- population genetics
- ecology
- physical science
- clustering
- machine learning/statistics.

Fragmentation trees
- phylogenetics
- linguistics

Complex networks
- physics
- population biology
- epidemiology
Partitions

\[ [n] := \{1, \ldots, n\} \text{ (set of labels)} \]

A partition \( B \) of \([n]\) is

- a set of non-empty disjoint subsets (blocks) \( b \subset [n] \) such that
  \[ \bigcup_{b \in B} b = [n], \text{ e.g. } B = 124|35|6 \equiv 35|6|124 \equiv \{\{1, 2, 4\}, \{3, 5\}, \{6\}\}; \]
- an equivalence relation \( B : [n] \times [n] \to \{0, 1\} \) with \( B(i, j) = 1 \iff i \sim_B j \);
- a symmetric Boolean matrix \((B_{ij}) := (B(i, j))\), e.g.

\[
\begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix}.
\]

For \( B \in \mathcal{P} \), \( \#B \) is number of blocks of \( B \), e.g. \( \#B = 3 \) above;
For \( b \in B \), \( \#b \) is the number of elements of \( b \subset \mathbb{N} \). e.g. \( \#\{1, 2, 4\} = 3 \).
$\mathcal{P}[n]$: set partitions of $[n]$

$\mathcal{P}[n]$ denotes the set of partitions of $[n]$

\[
\begin{align*}
\mathcal{P}[1] & : 1 \\
\mathcal{P}[2] & : 12 \ 1|2 \\
\mathcal{P}[3] & : 123 \ 1|23 \ 12|3 \ 13|2 \ 1|2|3
\end{align*}
\]

Action by permutation: $\sigma = (12)(3)$, $\pi = 13|2 \implies \pi^\sigma = 1|23$.

Restriction maps: $D_{m,n} : \mathcal{P}[n] \rightarrow \mathcal{P}[m]$, $D_{m,n}B := B_{[m]}$ ($1 \leq m \leq n$), e.g.

\[
D_{5,6}(1256|3|4) = 125|3|4.
\]

$\mathcal{P}_\infty$ is the collection $(\mathcal{P}[n], n \geq 1)$ together with deletion ($D_{m,n}, m \leq n$) and permutation maps, and all composite mappings, i.e. partitions of $\mathbb{N}$.
Exchangeable Feller Chains

\( \Pi := (\Pi_m, m \geq 0) \) is an exchangeable Feller chain on \( \mathcal{P}_\infty \) if

- **exchangeable**: \( D_n \Pi = L(D_n \Pi)^\sigma \) for all permutations \( \sigma : [n] \to [n] \).
- **Feller**: \( D_n \Pi \) is a Markov chain for all \( n \geq 1 \);

For example,

\[
\{1|2|34 \mapsto 134|2\} = L\{14|2|3 \mapsto 134|2\} = L\{14|2|3 \mapsto 124|3\}.
\]

\[
1|2|34 \mapsto \begin{cases} 
134|2 & 1/9 \\
13|24 & \text{w.p. } 1/18 \\
13|2|4 & 1/18 
\end{cases}
\]

\[
1|2|3 \mapsto 13|2 \quad \text{w.p. } \frac{1}{9} + \frac{1}{18} + \frac{1}{18} = \frac{2}{9}.
\]

\[
14|2|3 \mapsto \begin{cases} 
134|2 & 1/9 \\
13|24 & \text{w.p. } 1/18 \\
13|2|4 & 1/18 
\end{cases}
\]

\[
1|2|3 \mapsto 13|2 \quad \text{w.p. } \frac{1}{9} + \frac{1}{18} + \frac{1}{18} = \frac{2}{9}.
\]
Motivation: Mitochondrial DNA (mtDNA) sequences

mtDNA sequences for 9 species (snake, iguana, lizard, crocodile, bird, whale, cow, human, monkey)

<table>
<thead>
<tr>
<th></th>
<th></th>
<th>snake</th>
<th>T A G G A T T G A T A C C C C</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td></td>
<td>iguana</td>
<td>T A G G A T T G A T A C C C C</td>
</tr>
<tr>
<td>2</td>
<td></td>
<td>lizard</td>
<td>T A G G A T T G A T A C C C C</td>
</tr>
<tr>
<td>3</td>
<td></td>
<td>crocodile</td>
<td>T A G G A T T G A T A C C C C</td>
</tr>
<tr>
<td>4</td>
<td></td>
<td>bird</td>
<td>T G G G A T T G A T A C C C C</td>
</tr>
<tr>
<td>5</td>
<td></td>
<td>whale</td>
<td>T G G G A T T G A T A C C C C</td>
</tr>
<tr>
<td>6</td>
<td></td>
<td>cow</td>
<td>A A G C A T C T A C A C C C C</td>
</tr>
<tr>
<td>7</td>
<td></td>
<td>human</td>
<td>A A C C C C C C C C A T C C C</td>
</tr>
<tr>
<td>8</td>
<td></td>
<td>monkey</td>
<td>T G G G A T T G A T A C C C C</td>
</tr>
</tbody>
</table>

1234569|78 → 123478|569 → 12345679|8 → ···

How to model this sequence of partitions?
\( \mathcal{P}_{[\infty]:k} \), \( k \)-colorings of \( \mathbb{N} \) and partition matrices

\( \mathcal{P}_{[\infty]:k} \): partitions with at most \( k \) blocks  
\( \mathcal{L}_{[n]:k} \): \( k \)-colorings of \([n]\) (labeled partitions)  

- \( x \in \mathcal{L}_{[n]:k}: x = x^1 x^2 \cdots x^n \), e.g. \( x = 12112 \Rightarrow (134, 25) \).  
- Write a \( k \)-coloring as a set-valued vector \( L = (L_1, \ldots, L_k) \).  
- Natural map \( B_n : \mathcal{L}_{[n]:k} \to \mathcal{P}_{[n]:k} \) by removing colors  
  
\[
(34, 1, 256) \mapsto_{B_6} 1|256|34.
\]

- DNA example: with A, C, G, T as 1, 2, 3, 4:  
  \( x = TTTTTTAAT \Rightarrow (78, \emptyset, \emptyset, 1234569) \mapsto_{B_9} 1234659|78 \).

\( \mathcal{M}_{[n]:k} : k \times k \) partition matrices

\[
\begin{pmatrix}
234 & 1456 & 2 \\
15 & \emptyset & 146 \\
6 & 23 & 35
\end{pmatrix}
\begin{pmatrix}
34 \\
1 \\
256
\end{pmatrix}
=
\begin{pmatrix}
1234 \\
6 \\
5
\end{pmatrix}.
\]

In general,

\[
\begin{pmatrix}
M_{11} & M_{12} & \cdots & M_{1k} \\
M_{21} & M_{22} & \cdots & M_{2k} \\
\vdots & \vdots & \ddots & \vdots \\
M_{k1} & M_{k2} & \cdots & M_{kk}
\end{pmatrix}
\begin{pmatrix}
L_1 \\
L_2 \\
\vdots \\
L_k
\end{pmatrix}
=
\begin{pmatrix}
\bigcup_{j=1}^k (M_{1j} \cap L_j) \\
\bigcup_{j=1}^k (M_{2j} \cap L_j) \\
\vdots \\
\bigcup_{j=1}^k (M_{kj} \cap L_j)
\end{pmatrix}.
\]
Constructing Markov chains on $\mathcal{L}[\infty]:k$

Let:

- $\Lambda_0$ be an exchangeable initial state
- $\chi$ be a probability measure on $\mathcal{M}[\infty]:k$
- $M_1, M_2, \ldots$ be i.i.d. random partition matrices with distribution $\chi$ (independent of $\Lambda_0$).

For each $m \geq 1$, put

$$\Lambda_m := M_m \Lambda_{m-1}^T = M_m M_{m-1} \cdots M_1 \Lambda_0^T.$$

$\Lambda := (\Lambda_m, m \geq 0)$ is a Markov chain on $k$-colorings of $\mathbb{N}$.

Example, $\Lambda_0 = (1345, 26); ~ M_1 = \begin{pmatrix} 2345 & 256 \\ 16 & 134 \end{pmatrix}; ~ M_2 = \begin{pmatrix} 1345 & 24 \\ 26 & 1356 \end{pmatrix}$.

$\Lambda_0 = (1345, 26)$

$\Lambda_1 = M_1 \Lambda_0^T = \begin{pmatrix} 2345 & 256 \\ 16 & 134 \end{pmatrix} \begin{pmatrix} 1345 \\ 26 \end{pmatrix} = (23456, 1)$

$\Lambda_2 = M_2 \Lambda_1^T = M_2 M_1 \Lambda_0^T = \begin{pmatrix} 1345 & 24 \\ 26 & 1356 \end{pmatrix} \begin{pmatrix} 23456 \\ 1 \end{pmatrix} = (345, 126)$

$\vdots$
Theorem (C. 2012)

Every exchangeable Feller chain $\Lambda$ on $\mathcal{L}_{[\infty]:k}$ can be constructed from an i.i.d. sequence $M_1, M_2, \ldots$ so that

$$\Lambda_m = M_m M_{m-1} \cdots M_1 \Lambda_0, \quad m \geq 1.$$ 

Corollary (C. 2012)

Every exchangeable Feller chain $\Pi$ on $\mathcal{P}_{[\infty]:k}$ can be obtained as the projection $\mathcal{B}_\infty(\Lambda)$, where $\Lambda$ is an exchangeable Feller chain on $\mathcal{L}_{[\infty]:k}$. 

Matrix permanents

Recall: we can regard a partition $B$ as a symmetric Boolean matrix $(B_{ij}) := (B(i,j))$, e.g.

$$
\begin{pmatrix}
1 & 1 & 0 & 1 & 0 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
1 & 1 & 0 & 1 & 0 & 0 \\
0 & 0 & 1 & 0 & 1 & 0 \\
0 & 0 & 0 & 0 & 0 & 1
\end{pmatrix} = 124|35|6.
$$

For an $n \times n$ matrix $X$, the $\alpha$-permanent of $X$ is given by

$$
\text{per}_\alpha X := \sum_{\sigma \in \text{Sym}_n} \alpha^{\#\sigma} \prod_{j=1}^{n} X_{j\sigma(j)}.
$$

Hard to compute, but for a partition $B$, we have

$$
\text{per}_\alpha B = \prod_{b \in B} \alpha^{\#b}.
$$

Moreover, there is the identity

$$
\text{per}_\alpha X = \sum_{B \in \mathcal{P}_{[n]:k}} \frac{k!}{(k - \#B)!} \text{per}_{\alpha/k}(X \cdot B),
$$

$X \cdot B$ is the Hadamard product.
For $X$ a non-negative $n \times n$ matrix with positive diagonal entries and $\alpha > 0$, we have a general class of partition-valued Markovian transition probabilities on $\mathcal{P}_{[n]:k}$:

$$P_n(B, B') = \frac{k!}{(k - \#B')!} \frac{\text{per}_{\alpha/k}(X \cdot B \cdot B')}{\text{per}_\alpha(X \cdot B)}, \quad B, B' \in \mathcal{P}_{[n]:k}.$$ 

- Gives a parametric statistical model for dependent sequences of partitions.
- In cases of interest, $X$ is a discrete parameter $\implies$ hard to estimate.
Example: Phylogenetic inference

Use permanental partition transition probabilities with $X$ as a rooted tree matrix in likelihood-based inference of the unknown tree.

Given sequence $B = (B_1, B_2, \ldots, B_m)$, obtain a likelihood

$$
\mathcal{L}(X, \alpha; B) = \frac{k^{\downarrow \#B}}{\text{per}_{k\alpha} X} \left( \prod_{j=1}^{m-1} \frac{k^{\downarrow \#B_{j+1}}}{\text{per}_{\alpha/k} (X \cdot B_j \cdot B_{j+1})} \right) \frac{\text{per}_{\alpha} (X \cdot B_j)}{\text{per}_{\alpha} (X \cdot B_j)}
$$

How to (approximately) optimize with respect to $X$ (restricted to the space of rooted trees)?
References