Nonparametric Bayesian statistics with exchangeable random structures

Daniel M. Roy

University of Cambridge

October 2014

Thanks to

Nate Ackerman (Harvard) James R. Lloyd (Cambridge)
Cameron Freer (MIT/Analog) Zoubin Ghahramani (Cambridge)
Peter Orbanz (Columbia) Sara Wade (Cambridge)

Olav Kallenberg (Auburn) Mark Schervish (CMU)

Statistical Machine Learning

Given some data $X_1, X_2, \ldots$
identify hidden structure/patterns
in order to predict future or missing data.

1. (height, weight, age) of students

$$X = \begin{matrix} 171\text{cm} & 65\text{kg} & 21 \\ 182\text{cm} & 70\text{kg} & 19 \\ 170\text{cm} & ? & 20 \\ \vdots & \vdots & \vdots \end{matrix}$$

2. daily average (temperature, humidity, rainfall) in Guanajuato

$$X = \begin{matrix} 21^\circ\text{C} & 40\% & 0\text{cm} \\ 27^\circ\text{C} & 50\% & 0\text{cm} \\ 30^\circ\text{C} & 60\% & ? \\ \vdots & \vdots & \end{matrix}$$

3. X = scores/ratings for movies by users

$$X = \begin{matrix} 5 & 5 & 5 & 3 & 2 & .. \\ 2 & 1 & 2 & 5 & 5 & .. \\ 5 & ? & 5 & 2 & 3 & .. \\ \vdots & \vdots & \vdots & \vdots & \vdots & \end{matrix}$$

Is there structure in the data?

What structure is in the data?

1.

$$\{\, \square, \quad \square, \quad \square, \quad \square, \quad \square, \quad \cdots \}$$

2.

$$\square \to \square \to \square \to \square \to \square \quad \cdots$$

3.

$$\text{users} \left\{ \begin{array}{cccccc} \multicolumn{6}{c}{\overbrace{\hphantom{\square \quad \square \quad \square \quad \square \quad \square \quad \cdots}}^{\text{movies}}} \\ \square & \square & \square & \square & \square & \cdots \\ \square & \square & \cdots & & & \\ \square & \vdots & \ddots & & & \\ \vdots & & & & & \end{array} \right.$$

Is there even more structure in the data?

1. $$\mathbb{P}\begin{pmatrix} 171\text{cm} & 65\text{kg} & 21 \\ 182\text{cm} & 70\text{kg} & 19 \\ 170\text{cm} & ? & 20 \end{pmatrix} \stackrel{?}{=} \mathbb{P}\begin{pmatrix} 182\text{cm} & 70\text{kg} & 19 \\ 171\text{cm} & 65\text{kg} & 21 \\ 170\text{cm} & ? & 20 \end{pmatrix}$$

(Swapped top and middle row.)

2. $$\mathbb{P}\begin{pmatrix} 21^\circ\text{C} & 40\% & 0\text{cm} \\ 27^\circ\text{C} & 50\% & 0\text{cm} \\ 30^\circ\text{C} & 60\% & ? \end{pmatrix} \stackrel{?}{=} \mathbb{P}\begin{pmatrix} 27^\circ\text{C} & 50\% & 0\text{cm} \\ 21^\circ\text{C} & 40\% & 0\text{cm} \\ 30^\circ\text{C} & 60\% & ? \end{pmatrix}$$

(Swapped top and middle row.)

3. $$\mathbb{P}\begin{pmatrix} 5 & 5 & 5 & 3 & 2 \\ 2 & 1 & 2 & 5 & 5 \\ 5 & 4 & 5 & 2 & 3 \end{pmatrix} \stackrel{?}{=} \mathbb{P}\begin{pmatrix} 2 & 1 & 2 & 5 & 5 \\ 5 & 5 & 5 & 3 & 2 \\ 5 & 4 & 5 & 2 & 3 \end{pmatrix}$$

(Swapped top and middle row.)

$$\stackrel{?}{=} \mathbb{P}\begin{pmatrix} 1 & 2 & 2 & 5 & 5 \\ 5 & 5 & 5 & 3 & 2 \\ 4 & 5 & 5 & 2 & 3 \end{pmatrix}$$

(Swapped first and second columns.)

$$\stackrel{?}{=} \mathbb{P}\begin{pmatrix} 5 & 5 & 2 & 2 & 1 \\ 5 & 5 & 5 & 3 & 2 \\ 5 & 5 & 4 & 3 & 2 \end{pmatrix}$$

(Sorted each row.)

Probabilistic symmetries

1. **Exchangeable sequence**
   $\mathbb{P}[(X_1, X_2, \ldots, X_n)] = \mathbb{P}[(X_4, X_n, \ldots, X_1)] = \mathbb{P}[(X_{\pi(1)}, \ldots, X_{\pi(n)})]$
   Invariance to permutation.
2. **Stationary sequence**
   $\mathbb{P}[(X_1, X_2, \ldots)] = \mathbb{P}[(X_4, X_5, \ldots)] = \mathbb{P}[(X_{k+1}, X_{k+2}, \ldots)]$
   Invariance to shift.
3. **Exchangeable array**
   $$\mathbb{P}\left[\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots \\ X_{2,1} & X_{2,2} & \cdots \\ X_{3,1} & X_{3,2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}\right] = \mathbb{P}\left[\begin{pmatrix} X_{\pi(1),\tau(1)} & X_{\pi(1),\tau(2)} & \cdots \\ X_{\pi(2),\tau(1)} & X_{\pi(2),\tau(2)} & \cdots \\ X_{\pi(3),\tau(1)} & X_{\pi(3),\tau(2)} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix}\right]$$
   Invariance to separate permutation of rows and columns.

What is the most general way to model data assuming these symmetries?

Goals of this tutorial

1. Give a review of exchangeability in several forms.
2. Link each type of exchangeability to a representation theorem.
3. Explain how to interpret these representation theorems in their various forms.
4. Convey that probabilistic symmetries are an important consideration when constructing a statistical model.

Tutorial Outline

1. Exchangeable sequences.
2. Exchangeable graphs and arrays.

CAVEATS

A rigorous account requires measure theory.

1. All spaces are complete, separable metric spaces, equipped with their Borel σ-algebras.
2. All functions and sets are measurable.
3. Some necessary details will appear in a light gray color like this. Ignore these for now, and go back and study them later to understand the material at a greater depth.
4. I will define the "naturals" to be $\mathbb{N} := \{1, 2, \dots\}$.

Definition (exchangeable sequences)

Let $X_1, X_2, \dots$ be a sequence of random variables taking values in a space S. We say the sequence is **exchangeable** when, for every permutation π of $\mathbb{N}$,

$$(X_1, X_2, \dots) \overset{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots) \tag{1}$$

What does Eq. (1) mean?

$$(X_1, X_2, \dots) \overset{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots) \tag{2}$$

$\iff$ for all $n \in \mathbb{N}$,

$$(X_1, \dots, X_n) \overset{d}{=} (X_{\pi(1)}, \dots, X_{\pi(n)}). \tag{3}$$

$\iff$ for all $n \in \mathbb{N}$, for all subsets $A_1, \dots, A_n \subseteq S$,

$$\mathbb{P}\{X_1 \in A_1, \dots, X_n \in A_n\} = \mathbb{P}\{X_{\pi(1)} \in A_1, \dots, X_{\pi(n)} \in A_n\}. \tag{4}$$

$\iff$ for all $n \in \mathbb{N}$, for all subsets $A_1, \dots, A_n \subseteq S$,

$$\mathbb{P}\{X_1 \in A_1, \dots, X_n \in A_n\} = \mathbb{P}\{X_1 \in A_{\pi(1)}, \dots, X_n \in A_{\pi(n)}\}. \tag{5}$$

Definition (exchangeable sequences)

Let $X_1, X_2, \dots$ be a sequence of random variables taking values in a space S. We say the sequence is **exchangeable** when, for every permutation π of $\mathbb{N}$,

$$(X_1, X_2, \dots) \overset{\mathrm{d}}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots) \tag{1}$$

What does Eq. (1) mean?

$$(X_1, X_2, \dots) \overset{\mathrm{d}}{=} (X_{\pi(1)}, X_{\pi(2)}, \dots) \tag{2}$$

$\iff$ for all $n \in \mathbb{N}$,

$$(X_1, \dots, X_n) \overset{\mathrm{d}}{=} (X_{\pi(1)}, \dots, X_{\pi(n)}). \tag{3}$$

$\iff$ for all $n \in \mathbb{N}$, and distinct $k_1, \dots, k_n \in \mathbb{N}$,

$$(X_1, \dots, X_n) \overset{\mathrm{d}}{=} (X_{k_1}, \dots, X_{k_n}). \tag{4}$$

$\iff$ for all $n \in \mathbb{N}$, and permutations π of $[n] := \{1, 2, \dots, n\}$,

$$(X_1, \dots, X_n) \overset{\mathrm{d}}{=} (X_{\pi(1)}, \dots, X_{\pi(n)}). \tag{5}$$

i.i.d. sequences are exchangeable

Recall that a sequence $X_1, X_2, \ldots$ of random variables is **independent** when

$$\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_i \in A_i\}, \tag{6}$$

for all $n \in \mathbb{N}$ and subsets $A_1, \ldots, A_n \subseteq S$,
and is **independent and identically distributed (i.i.d.)** when

$$\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \prod_{i=1}^{n} \mathbb{P}\{X_1 \in A_i\} = \prod_{i=1}^{n} \mu(A_i), \tag{7}$$

where $\mu = \mathbb{P}\{X_1 \in \cdot\,\}$ is the marginal distribution of every element.
We will say that $X_1, X_2, \ldots$ is i.i.d.-μ.

For every permutation π of $[n] = \{1, 2, \ldots, n\}$.

$$\prod_{i=1}^{n} \mu(A_{\pi(i)}) = \prod_{i=1}^{n} \mu(A_i) \tag{8}$$

conditionally i.i.d. sequences are exchangeable

Recall that a sequence $X_1, X_2, \ldots$ of random variables is **conditionally independent** when there is a random variable θ such that a.s.

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n \mid \theta] = \prod_{i=1}^{n} \mathbb{P}[X_i \in A_i \mid \theta], \tag{9}$$

for all $n \in \mathbb{N}$ and $A_1, \ldots, A_n \subseteq S$, and is **conditionally i.i.d.** given θ when a.s.

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n \mid \theta] = \prod_{i=1}^{n} \mathbb{P}[X_1 \in A_i \mid \theta] = \prod_{i=1}^{n} \nu(A_i), \tag{10}$$

where $\nu := \mathbb{P}[X_1 \in \cdot \mid \theta]$ is the (random) marginal distribution, conditioned on θ. Note that the sequence is also conditionally i.i.d.-ν given ν.

Let $X_1, X_2, \ldots$ be conditionally i.i.d.-ν given ν. Let $n \in \mathbb{N}$, π a permutation of $[n]$.

$$\mathbb{P}\{X_{\pi(1)} \in A_1, \ldots, X_{\pi(1)} \in A_n\} \tag{11}$$

$$= \mathbb{E}\Big(\mathbb{P}[X_{\pi(1)} \in A_1, \ldots, X_{\pi(1)} \in A_n \mid \nu]\Big) = \mathbb{E}\Big(\prod_{i=1}^{n} \nu(A_i)\Big) \tag{12}$$

Pölya's urn

Let $S = \{0, 1\}$.
Let $\mathbb{P}\{X = 1\} = \mathbb{P}\{X = 0\} = 1$. In other words,

$$X_1 \sim \text{Bernoulli}(1/2) \tag{13}$$

Let $S_n = X_1 + \cdots + X_n$.

$$\mathbb{P}[X_{n+1} = 1 \mid X_1, \ldots, X_n] = \frac{S_n + 1}{n + 2}. \tag{14}$$

In other words,

$$X_{n+1} \mid X_1, \ldots, X_n \sim \text{Bernoulli}(\tfrac{S_n+1}{n+2}) \tag{15}$$

$1\ 0\ 1\ 1\ 0\ ?$ $\qquad$ $\mathbb{P}\{1 \text{ next} \mid \text{seen } 10110\} = 3 + 15 + 2.$

Pölya's urn

$$X_1 \sim \text{Bernoulli}(1/2) \tag{16}$$

$$X_{n+1} \mid X_1, \ldots, X_n \sim \text{Bernoulli}(\tfrac{S_n+1}{n+2}), \qquad \text{for } n \in \mathbb{N}. \tag{17}$$

Note that each element X_{n+1} depends on all previous elements. No independence!

Let $x_1, \ldots, x_n \in \{0, 1\}$ and define $s_j = x_1 + \cdots + x_j$.

$$\mathbb{P}\{X_1 = x_1, X_2 = x_2, \ldots, X_n = x_n\} \tag{18}$$

$$= \mathbb{P}\{X_1 = x_1\} \cdot \mathbb{P}\{X_2 = x_2 \mid X_1 = x_1\} \cdots \mathbb{P}\{X_n = x_n \mid X_1 = x_1, \ldots, X_{n-1} = x_{n-1}\} \tag{19}$$

$$= \underbrace{\left(\frac{1}{2}\right)}_{\mathbb{P}\{X_1=x_1\}} \cdot \underbrace{\left(\frac{s_1+1}{3}\right)^{x_2}\left(\frac{2-s_1}{3}\right)^{1-x_2}}_{\mathbb{P}\{X_2=x_2 \mid X_1=x_1\}} \cdots \underbrace{\left(\frac{s_{n-1}+1}{n+1}\right)^{x_n}\left(\frac{n-1-s_{n-1}}{n+1}\right)^{1-x_n}}_{\mathbb{P}\{X_n=x_n \mid X_1=x_1,\ldots,X_{n-1}=x_{n-1}\}} \tag{20}$$

$$= \frac{(s_n)!(n-s_n)!}{(n+1)!} \qquad \text{this is invariant to permutation, hence exchangeable!} \tag{21}$$

Theorem (de Finetti; Hewitt-Savage)

Let $X = (X_1, X_2, \ldots)$ be an infinite sequence of random variables in a space S. The following are equivalent:

1. *X is exchangeable.*
2. *X is conditionally i.i.d.*

Exchangeable:

$$(X_1, X_2, \ldots) \overset{d}{=} (X_{\pi(1)}, X_{\pi(2)}, \ldots), \qquad \text{for } \pi \in S_\infty \tag{22}$$

Conditionally i.i.d.: $\exists$ random variable θ s.t., for all n and A_i,

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n \,|\, \theta] = \prod_{i=1}^n \mathbb{P}[X_1 \in A_i \,|\, \theta] \text{ a.s.} \tag{23}$$

Equivalently, $\exists$ random probability measure ν s.t., for all n and A_i,

$$\mathbb{P}[X_1 \in A_1, \ldots, X_n \in A_n \,|\, \nu] = \prod_{i=1}^n \nu(A_i) \text{ a.s.} \tag{24}$$

Taking expectations, $\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \mathbb{E}\Big(\prod_{i=1}^n \nu(A_i)\Big)$. (25)

Mixed i.i.d.: $\exists$ distribution μ (the de Finetti mixing measure) s.t., for all n and A_i,

$$\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \int \prod_{i=1}^n v(A_i)\, \mu(\mathrm{d}v) \tag{26}$$

PÖLYA'S URN

$$X_1 \sim \text{Bernoulli}(1/2) \tag{27}$$

$$X_{n+1} \mid X_1, \dots, X_n \sim \text{Bernoulli}(\tfrac{S_n+1}{n+2}), \qquad \text{for } n \in \mathbb{N}. \tag{28}$$

Exchangeable:

$$\mathbb{P}\{X_1 = x_1, \dots, X_n = x_n\} = \tfrac{(s_n)!(n-s_n)!}{(n+1)!} \tag{29}$$

Conditionally i.i.d.: There is a random variable θ s.t. a.s.

$$\mathbb{P}[X_1 = x_1, \dots, X_n = x_n \mid \theta] = \prod_{i=1}^{n} \mathbb{P}[X_1 = x_i \mid \theta] \tag{30}$$

$$= \mathbb{P}[X_1 = 1 \mid \theta]^{s_n} \mathbb{P}[X_1 = 0 \mid \theta]^{n-s_n} \tag{31}$$

$$= \nu_1^{s_n} (1-\nu_1)^{n-s_n} \tag{32}$$

where $\nu_1 = \mathbb{P}[X_1 = 1 \mid \theta]$. Note that $(X_n)_{n\in\mathbb{N}}$ is conditionally i.i.d. given ν_1.
Are θ and ν_1 different? How are they related?

Mixed i.i.d.: Let μ be the distribution of ν_1. Taking expectations,

$$\mathbb{P}\{X_1 = x_1, \dots, X_n = x_n\} = \int_0^1 \vartheta^{s_n} (1-\vartheta)^{n-s_n} \, \mu(\mathrm{d}\vartheta) \tag{33}$$

Equating expressions for $\mathbb{P}\{X_1 = x_1, \dots, X_n = x_n\}$ implies $\mu =$Uniform[0,1]!

DE FINETTI'S THEOREM IN THE CASE OF PÖLYA'S URN

Pölya's urn:

$$X_1 \sim \text{Bernoulli}(1/2) \tag{34}$$

$$X_{n+1} \mid X_1, \dots, X_n \sim \text{Bernoulli}(\tfrac{S_n+1}{n+2}), \qquad \text{for } n \in \mathbb{N}. \tag{35}$$

Beta-Bernoulli process:

$$\nu_1 \sim \mu \tag{36}$$

$$X_n \mid \nu_1 \overset{iid}{\sim} \text{Bernoulli}(\nu_1), \qquad \text{for } n \in \mathbb{N}. \tag{37}$$

How are $X = (X_n)_{n \in \mathbb{N}}$, $\mathbb{P}[X]$ and ν_1, $\mu = \mathbb{P}[\nu_1]$ related?

In what sense are ν_1 and μ uniquely determined by X and $\mathbb{P}[X]$?

- By the law of large numbers, and the disintegration theorem,

$$\frac{1}{n} \sum_{j=1}^{n} X_j \longrightarrow \nu_1 \qquad \text{as} \qquad n \to \infty \quad \text{a.s.} \tag{38}$$

- If θ renders X conditionally i.i.d., then $\nu_1 = g(\theta)$ a.s. for some function g.
- If μ' is a measure such that

$$\mathbb{P}\{X_1 = x_1, \dots, X_n = x_n\} = \int v^{s_n} (1-v)^{n-s_n} \, \mu'(\mathrm{d}v) \tag{39}$$

then $\mu' = \mathbb{P}[\nu_1]$.

DE FINETTI'S THEOREM IN GENERAL

Let $X = (X_n)_{n \in \mathbb{N}}$ be an exchangeable sequence in a space S.
Let $\hat{P}_n = \frac{1}{n}(\delta_{X_1} + \cdots + \delta_{X_n})$ be the empirical measure.
Define a random measure ν on S by

$$\nu(A) = \lim_{n\to\infty} \hat{P}_n(A) \quad \text{a.s.} \qquad A \subseteq S. \tag{40}$$

Informally, $\nu = \lim_{n\to\infty} \hat{P}_n$. Let $\mu = \mathbb{P}[\nu]$.
Then X is conditionally i.i.d.-ν given ν. That is

$$X_n \,|\, \nu \overset{iid}{\sim} \nu \tag{41}$$

Uniqueness?

- If θ renders X conditionally i.i.d., then $\nu = g(\theta)$ a.s. for some function g.
- If μ' is a measure such that

$$\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \int \prod_{i=1}^{n} v(A_i)\, \mu'(\mathrm{d}v) \tag{42}$$

then $\mu' = \mathbb{P}[\nu]$.

Definition (Statistical inference base/model)

1. Sample space $\mathcal{X}$.
2. Parametric family $\mathcal{P}_0 := \{P_\theta\}_{\theta \in T}$ of probability distributions on $\mathcal{X}$ indexed by elements of T called parameters. T is called the parameter space.
3. Observed data $x^* \in \mathcal{X}$.
4. Loss function $L : T \times T \to \mathbb{R}$.

The *risk* of an estimator $\delta : S \to T$ for $t \in T$, is

$$R(\delta, \theta) = \mathbb{E}_{X \sim P_\theta}\{L(\theta, \delta(X))\} := \int_{\mathcal{X}} L(\theta, \delta(x))\, P_\theta(\mathrm{d}x) \tag{43}$$

Definition (Bayesian estimator)

Let θ be a random variable in T, with prior distribution π, and let $X \mid \theta \ \sim \ P_\theta$.
The *Bayesian estimator* minimizes the posterior expected loss:

$$\delta_\pi(x) := \arg\min_{\theta^* \in T} \mathbb{E}_{\theta \,\sim\, \mathbb{P}[\theta \mid X = x]}\{L(\theta, \theta^*)\} \tag{44}$$

EXCHANGEABILITY AND STATISTICS: NOT JUST BAYESIAN!

Classic i.i.d. framework

Let $X_1, \ldots, X_n \overset{iid}{\sim} F$, for an unknown distribution $F \in \mathcal{F}_0 := \{F_\theta\}_{\theta \in T}$.

Say observations are $\mathbb{R}$-valued. We can formalize this as follows:

1. Sample space $\mathcal{X} = \mathbb{R}^n$
2. Parametric family $\mathcal{F}_0^n := \{F_\theta^n : \theta \in T\}$, where F^n is the n-fold product.

Let $(X_1, \ldots, X_n) \sim F^n$, for an unknown $F^n \in \mathcal{F}_0^n$.

Exchangeable observation

Let $(X_1, \ldots, X_n) \sim P$ be conditionally i.i.d., for an unknown distribution P.

Conditionally i.i.d. observation

Let $(X_1, \ldots, X_n) \,|\, F \sim F^n$ and $F \sim \mu$, for an unknown distribution μ.

Even as $n \to \infty$, data reflects only one realization F from μ. The problem of estimating μ is “ill-posed” to the frequentist. The Bayesian gets 1 data point.

Bayesian approach in the i.i.d. framework

Let θ be a random variable with some prior. Then F_θ is a random measure, and $X_1, X_2, \ldots$ is an exchangeable sequence.

"DE FINETTI'S THEOREM JUSTIFIES ..." NOT THE POINT.

de Finetti's philosophy

1. de Finetti rejected the idea of a parameter and argued there was no need to assume their existence.
2. He thought that probabilities should be specified only on observable quantities.
3. de Finetti's theorem shows that conserved quantities (like the limiting frequency of 1's) arise from symmetries and are random variables.
4. One can then interpret this in the classic sense (specifying a parameter and placing a prior on it), but there's no need to do that. The underlying random measure ν is there whether you like it or not.

Subjectivism

1. Distribution represents subjective (personal) uncertainty.
2. Exchangeability $\iff$ *certainty* that order of the data is irrelevant.
3. subjective distribution on data alone $+$ exchangeability $\implies$ conditionally i.i.d. distribution on ν is subjective as well.
4. Note: de Finetti's holds only for infinite sequences, but subjectivist need only be unwilling to posit an upper bound on the data size and projectivity.

CONNECTION WITH NONPARAMETRIC INFERENCE

What does exchangeability have to do with *nonparametric* Bayesian statistics?

Bayesian

1. Model: $X_1, X_2, \ldots \overset{iid}{\sim} F$ for unknown $F \in \mathcal{F}_0$. Need prior on $\mathcal{F}_0$.
2. If $\mathcal{F}_0 = \{\text{Gaussian distributions on } \mathbb{R}\}$ then $\mathcal{F}_0 \cong \mathbb{R}^2$.

   Finite-dimensional space, hence parametric.
3. If $\mathcal{F}_0 = \{\text{All Borel probability measures on } \mathbb{R}\}$ then $\mathcal{F}_0 \cong \mathbb{R}^{\mathbb{N}}$.

   Infinite-dimensional space, hence nonparametric.
4. Hard. But if you have a specific question to ask (e.g., what is $F(A)$ for some set A?) then it's possible.
5. Dirichlet process, Polya trees, Normalized Completely Random Measures, etc.

Subjectivist

1. Need a model for our data $X_1, X_2, \ldots$.
2. If we believe order is irrelevant, by exchangeability, it suffices to specify a prior measure μ on space of probability measures.
3. No further assumptions: support of μ must be all distributions.

   Hence μ will be a nonparametric prior.

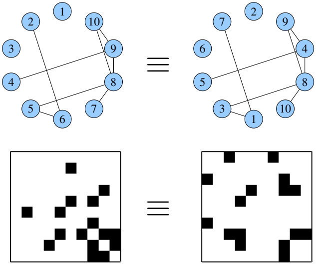
So far we have...

1. Reviewed exchangeability for sequences.
2. Presented de Finetti's representation theorem.
   Exchangeable if and only if conditionally i.i.d.
3. Discussed how to interpret de Finetti's theorems in its various forms.
4. Shown that probabilistic symmetries lead to statistical models.
   Exchangeability leads to a Bayesian approach to the classic i.i.d. framework.

Tutorial Outline

1. ~~Exchangeable sequences.~~
2. Exchangeable graphs and arrays.

Let $X = (X_{i,j})_{i,j \in \mathbb{N}}$ be the adjacency matrix of an undirected graph on $\mathbb{N}$.



Definition (jointly exchangeable array)

Call X **(jointly) exchangeable** when, for every permutation π of $\mathbb{N}$,

$$(X_{i,j})_{i,j \in \mathbb{N}} \overset{\text{d}}{=} (X_{\pi(i),\pi(j)})_{i,j \in \mathbb{N}}. \tag{45}$$

Most figures by James Lloyd (Cambridge) and Peter Orbanz (Columbia)

Definition (jointly exchangeable array)

Call X **(jointly) exchangeable** when, for every permutation π of $\mathbb{N}$,

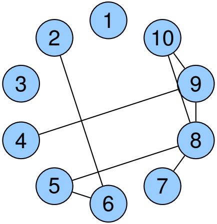
$$(X_{i,j})_{i,j \in \mathbb{N}} \overset{d}{=} (X_{\pi(i),\pi(j)})_{i,j \in \mathbb{N}}. \tag{46}$$

equivalently

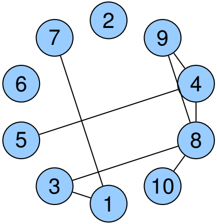
$$\begin{pmatrix} X_{1,1} & X_{1,2} & \cdots \\ X_{2,1} & X_{2,2} & \cdots \\ X_{3,1} & X_{3,2} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \overset{d}{=} \begin{pmatrix} X_{\pi(1),\pi(1)} & X_{\pi(1),\pi(2)} & \cdots \\ X_{\pi(2),\pi(1)} & X_{\pi(2),\pi(2)} & \cdots \\ X_{\pi(3),\pi(1)} & X_{\pi(3),\pi(2)} & \cdots \\ \vdots & \vdots & \ddots \end{pmatrix} \tag{47}$$

In the case $X_{i,j} \in \{0,1\}$, let $X^n = (X_{i,j})_{i,j \leq n}$.Then X is an exchangeable graph if for all $n \in \mathbb{N}$ and isomorphic graphs G, G' on $[n]$, $\mathbb{P}\{X^n = G\} = \mathbb{P}\{X^n = G'\}$.

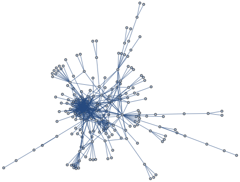
$\mathbb{P}\Big\{ X^{10} =$

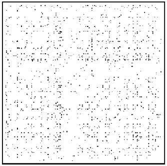


$\Big\} = \mathbb{P}\Big\{ X^{10} =$

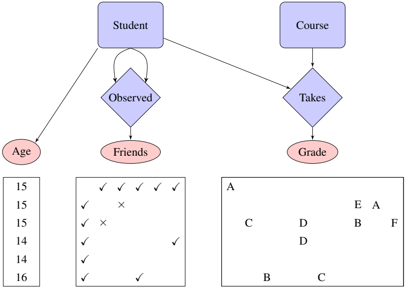


$\Big\}$





- Links between websites
- Proteins that interact
- Products that customers have purchased
- Relational databases



Are the following graphs exchangeable?

Example

$$X_{i,j} := 0 \qquad \text{for } i < j \in \mathbb{N}. \tag{48}$$

Yes.

Example

$$X_{i,j} := 1 \qquad \text{for } i < j \in \mathbb{N}. \tag{49}$$

Yes.

Is the following graph exchangeable?

Example

$$X_{i,j} := (j - i) \mod 2, \qquad \text{for } i < j \in \mathbb{N}. \tag{50}$$

$$\mathbb{P}\left\{ X^3 = 1 - 2 - 3 \right\} \stackrel{?}{=} \mathbb{P}\left\{ X^3 = 1 - 3 - 2 \right\}$$

No.

Is the following graph exchangeable?

Example

Consider the graph with vertex set $\mathbb{N}$ such that for every pair of vertices $i, j \in \mathbb{N}$, we include the edge $\{i, j\}$ indepenently with probability $p \in [0, 1]$.

The adjacency matrix X is such that

$$X_{i,j} \overset{iid}{\sim} \text{Bernoulli}(p) \qquad \text{for } i < j \in \mathbb{N}. \tag{51}$$

Let G be a graph on $[n]$. Then $\mathbb{P}\{X^n = G\} = 2^{-\binom{n}{2}^2}$.

Yes.

The resulting graph is a so-called “Erdös-Rényi graph”.

Is the following graph exchangeable?

Example

Let $X_{1,2} \sim \mathrm{Bernoulli}(1/2)$.
Otherwise, let

$$X_{i,j} := X_{1,2} \qquad \text{for } i < j \in \mathbb{N}. \tag{52}$$

Yes.

Is the following graph exchangeable?

Example

Let $Y_1, Y_2, \ldots$ be a Pölya urn. Let $\phi : \mathbb{N}^2 \to \mathbb{N}$ be a bijection.
Let

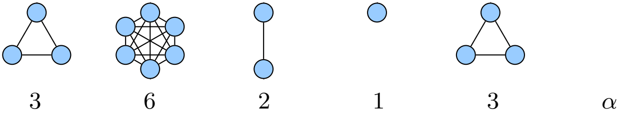
$$X_{i,j} := Y_{\phi(i,j)} \qquad \text{for } i < j \in \mathbb{N}. \tag{53}$$

Yes.

Is the following graph exchangeable?

Example

Consider the graph built one vertex at a time, adding a vertex to a clique with probability proportional to the size of the clique, and creating a new (singleton) clique with probability proportional to a constant $\alpha > 0$.



Yes.

The process is just an graph version of the **Chinese restaurant process** and is very closely related to the **Infinite Relational Model** of Kemp et al. (2008).

Is the following graph exchangeable?

Example

Let $N = (N_1, N_2, \dots)$ be an i.i.d. sequence Gaussian vectors in $\mathbb{R}^D$.
Let $\langle \cdot, \cdot \rangle$ be the dot product.
Let $\mathrm{sigmoid} : \mathbb{R} \to [0, 1]$ be a squashing function.

$$X_{i,j} \mid N \overset{ind}{\sim} \mathrm{Bernoulli}(\mathrm{sigmoid}(\langle N_i, N_j \rangle)) \qquad \text{for } i < j \in \mathbb{N}. \tag{54}$$

Yes.

This model is related to matrix factorization techniques, as well as the **eigenmodel** (Hoff 2008).

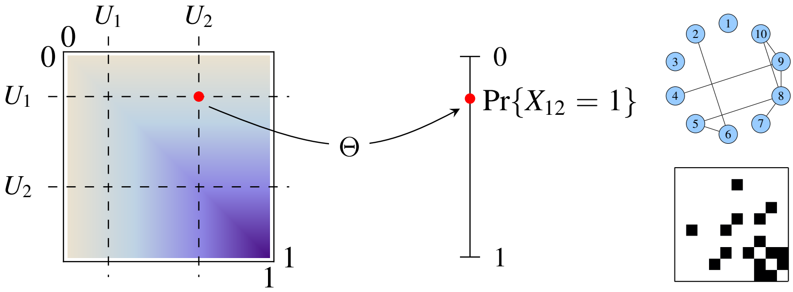
Let $U_1, U_2, \ldots$ be i.i.d. uniform random variables in $[0,1]$.

Definition (Θ-random graph)

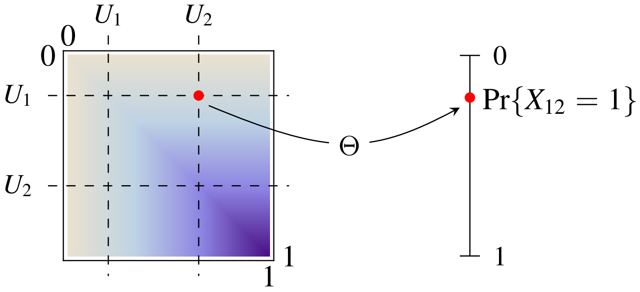
Let $\Theta : [0,1]^2 \to [0,1]$ be a symmetric measurable function, and let

$$X_{i,j} := 1 \text{ with probability } \Theta(U_i, U_j) \tag{55}$$

independently for every $i < j \in \mathbb{N}$. By a **Θ-random graph** we mean an array with the same distribution as X.



Let $\mathcal{W}$ be the space of symmetric measurable functions from $[0,1]^2$ to $[0,1]$.
Such functions Θ are called "graphons".



Let G be a graph on $[n] := \{1, \ldots, n\}$.

$$\mathbb{P}\{X^n = G \mid U_1, \ldots, U_n] = \prod_{i,j} \Theta(U_i, U_j)^{G_{i,j}} (1 - \Theta(U_i, U_j))^{1-G_{i,j}} \text{ a.s.} \tag{56}$$

Taking expectations,

$$\mathbb{P}\{X^n = G\} = \int_{[0,1]^n} \prod_{i,j} \Theta(u_i, u_j)^{G_{i,j}} (1 - \Theta(u_i, u_j))^{1-G_{i,j}} \, \mathrm{d}u_1 \cdots \mathrm{d}u_n \tag{57}$$

Theorem (Aldous, Hoover)

Let $X = (X_{i,j})_{i,j\in\mathbb{N}}$ be the adjacency matrix of an undirected graph on $\mathbb{N}$. The following are equivalent:

1. *X is jointly exchangeable.*
2. *X is conditionally Θ-random, given a random graphon Θ.*

Example 1 - empty graph

Let $\Theta(u, v) = 0$.

Example 2 - complete graph

Let $\Theta(u, v) = 1$.

Example 3 - Erdos-Renyi graph

For $p \in [0, 1]$, let $\Theta(u, v) = p$.

Example

Let $Y_1, Y_2, \ldots$ be a Pölya urn. Let $\phi : \mathbb{N}^2 \to \mathbb{N}$ be a bijection.
Let

$$X_{i,j} := Y_{\phi(i,j)} \qquad \text{for } i < j \in \mathbb{N}. \tag{58}$$

What's Θ?

Let $p \sim \mathrm{Uniform}$. Let $\Theta(u, v) = p$.

Example

Let $X_{1,2} \sim \mathrm{Bernoulli}(1/2)$.
Otherwise, let

$$X_{i,j} := X_{1,2} \qquad \text{for } i < j \in \mathbb{N}. \tag{59}$$

What's Θ?

Let $p \sim \mathrm{Bernoulli}(1/2)$. Let $\Theta(u, v) = p$.

Example

Let $N = (N_1, N_2, \dots)$ be an i.i.d. sequence Gaussian vectors in $\mathbb{R}^D$.
Let $\langle \cdot, \cdot \rangle$ be the dot product.
Let $\mathrm{sigmoid} : \mathbb{R} \to [0, 1]$ be a squashing function.

$$X_{i,j} \mid N \overset{ind}{\sim} \mathrm{Bernoulli}(\mathrm{sigmoid}(\langle N_i, N_j \rangle)) \qquad \text{for } i < j \in \mathbb{N}. \tag{60}$$

What's Θ?

Let $g : [0, 1] \to \mathbb{R}^d$ be such that $g(U) \sim \mathcal{N}_D(0, I_D)$ when $U \sim \mathrm{Uniform}$.

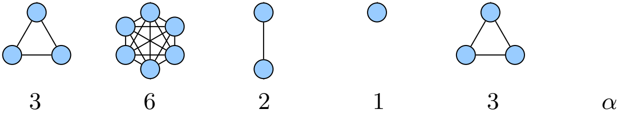
Let $\Theta(u, v) = \mathrm{sigmoid}(\langle g(u), g(v) \rangle)$.
Implicitly, we've been dealing with graphons "defined" on the probability space $([0, 1], \mathcal{B}_{[0,1]}, \mathrm{Uniform})$.

Consider instead a graphon Θ' defined on $(\mathbb{R}^D, \mathcal{B}_{\mathbb{R}^D}, \mathcal{N}_D(0, I_D))$ given by
$\Theta'(n, m) = \mathrm{sigmoid}(\langle n, m \rangle)$.

GRAPHONS UNDERLYING EXCHANGEABLE GRAPHS

Example

Consider the graph built one vertex at a time, adding a vertex to a clique with probability proportional to the size of the clique, and creating a new (singleton) clique with probability proportional to a constant $\alpha > 0$.



What's Θ?

Let $p_1 \geq p_2 \geq \cdots$ be a draw from Poisson-Dirichlet distribution (limiting table size proportions in CRP), considered as a random distribution on $\mathbb{N}$.
Let $(\mathbb{N}, \mathcal{B}_{\mathbb{N}}, p)$ be the random probability space.
Consider a graphon Θ' on this random probability space, given by

$\Theta'(n, m) = 1$ if $n = m$, $= 0$ otherwise.

On $([0,1], \mathcal{B}_{[0,1]}, \mathrm{Uniform})$?

$\Theta(u, v) = \Theta'(g(u), g(v))$ where $g(u) = \sup\{n \in \mathbb{N} : u \leq p_n\}$.

Example

$$X_{i,j} := (j - i) \mod 2, \qquad \text{for } i < j \in \mathbb{N}. \tag{61}$$

What's Θ?

X is not exchangeable!

Theorem (Aldous, Hoover)

Let $X = (X_{i,j})_{i,j\in\mathbb{N}}$ be the adjacency matrix of an undirected graph on $\mathbb{N}$. The following are equivalent:

1. *X is jointly exchangeable.*
2. *X is conditionally Θ-random, given a random graphon Θ.*

Exchangeable:

$$(X_{i,j})_{i,j\in\mathbb{N}} \overset{d}{=} (X_{\pi(i),\pi(j)})_{i,j\in\mathbb{N}}, \qquad \text{for } \pi \in S_\infty \tag{62}$$

Conditionally Θ-random: $\exists$ random graphon Θ s.t., for all finite graphs G on $[n]$,

$$\mathbb{P}[X^n = G \mid \Theta] = \int_{[0,1]^n} \prod_{i,j} \Theta(u_i, u_j)^{G_{i,j}} (1 - \Theta(u_i, u_j))^{1-G_{i,j}}\, \mathrm{d}u_1 \cdots \mathrm{d}u_n \quad \text{a.s.} \tag{63}$$

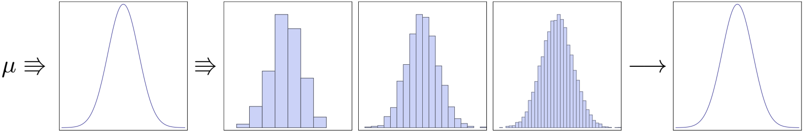
Taking expectations, $\exists$ distribution μ on graphons s.t. for all finite graphs G on $[n]$,

$$\mathbb{P}[X^n = G] = \int \int_{[0,1]^n} \prod_{(i,j)\in G} \theta(u_i, u_j) \prod_{(i,j)\notin G} (1 - \theta(u_i, u_j))\, \mathrm{d}u_1 \cdots \mathrm{d}u_n \mu(\mathrm{d}\theta)$$

Θ AS THE LIMITING EMPIRICAL GRAPHON

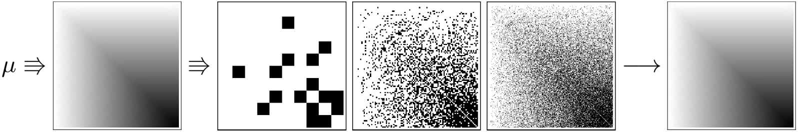
Exchangeable sequences

Recall that if $Y = (Y_1, Y_2, \dots)$ is an exchangeable sequence then $\hat{P}_n \to \nu$ a.s.



Exchangeable graphs

Let $X = (X_{i,j})_{i,j \in \mathbb{N}}$ is an exchangeable graph.



You can recover the graphon Θ underlying a graph by sampling larger and larger graphs. Suitably permuted, they converge in L^1 with probability one.

Θ IS NOT UNIQUE (AS A FUNCTION)

Exchangeable sequences and the uniqueness of $\mu = \mathbb{P}[\nu]$

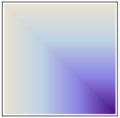
If $X_1, X_2, \ldots$ is an exchangeable sequence, there is a UNIQUE μ s.t.

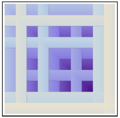
$$\mathbb{P}\{X_1 \in A_1, \ldots, X_n \in A_n\} = \int \prod_{i=1}^{n} v(A_i)\, \mu(\mathrm{d}v) \tag{64}$$

Exchangeable graphs and the uniqueness of $\mu = \mathbb{P}[\Theta]$

Let X be a Θ-random graph.
Let $T : [0,1] \to [0,1]$ be a measure preserving transformation, and define

$$\Theta^T(x,y) := \Theta(T(x), T(y)). \tag{65}$$





X is Θ^T-random too! Θ^T and Θ induce the same distribution on graphs.

Theorem (Hoover)

The measurable function f underlying an exchangeable array is unique up to a measure preserving transformation.

We have now...

1. Reviewed exchangeability for graphs
   (i.e., symmetric $\{0, 1\}$-valued arrays representing adjacency).
2. Presented Aldous-Hoover representation theorem in this special case.
   Graph is exchangeable if and only if conditionally Θ-random
3. Discussed how to interpret Aldous-Hoover theorem.
   Meaning of Θ-random. Θ as the limiting empirical graphon. Θ-nonuniqueness.
4. ~~Shown that probabilistic symmetries lead to statistical models.~~

Tutorial Outline

1. ~~Exchangeable sequences.~~
2. Exchangeable ~~graphs~~ and arrays.

EXCHANGEABLE ARRAYS

Let $X = (X_{i,j})_{i,j\in\mathbb{N}}$ be an infinite array of random variables.
No longer assuming $\{0,1\}$-values or symmetry of X.
E.g., adjacency matrix for a directed graph, or matrix of user-movie ratings.

Definition (jointly exchangeable array)

We say that X is **jointly exchangeable** when

$$X \overset{\text{d}}{=} (X_{\pi(i),\pi(j)})_{i,j\in\mathbb{N}} \tag{66}$$

for every permutation $\pi : \mathbb{N} \to \mathbb{N}$.

E.g., undirected graph, directed graph. Rows and columns are indexing "same set".

Definition (separately exchangeable array)

We say that X is **separately exchangeable** (aka row-column exchangeable) when

$$X \overset{\text{d}}{=} (X_{\pi(i),\pi'(j)})_{i,j\in\mathbb{N}} \tag{67}$$

for every pair of permutations $\pi, \pi' : \mathbb{N} \to \mathbb{N}$.

E.g., user-movie ratings. Rows and columns indexing different sets.

Example

Elements $X_{i,j}$ i.i.d.
Separately exchangeable? Jointly exchangeable?

Example

$X_{i,j} = 1$ if $i = j$ and $X_{i,j} = 0$ otherwise.
Separately exchangeable? Jointly exchangeable?

Example

Let $U_{i,j}$ be i.i.d. Uniform random variables.
Let $X_{i,j} = f(U_{00}, U_{i,0}, U_{0,j}, U_{i,j})$ for a suitable function f.
Separately exchangeable? Jointly exchangeable?

Example

Let U_i be i.i.d. Uniform random variables.
Let $X_{i,j} = g(U_i, U_j)$ for a suitable function g.
Separately exchangeable? Jointly exchangeable?

Definition (Aldous-Hoover, separately exchangeable)

Let $U_{i,j}$ be i.i.d. Uniform random variables.
An infinite array X is separately exchangeable if and only if

$$X \stackrel{\mathrm{d}}{=} (f(U_{00}, U_{i,0}, U_{0,j}, U_{i,j}))_{i,j \in \mathbb{N}} \tag{68}$$

for some measurable function f.

Definition (Aldous-Hoover, jointly exchangeable)

Let $U_{\{i,j\}}$ be i.i.d. Uniform random variables. That is $U_{\{i,j\}} = U_{\{j,i\}}$.

An infinite array X is separately exchangeable if and only if

$$X \stackrel{\mathrm{d}}{=} (f(U_{\{0,0\}}, U_{\{i,0\}}, U_{\{0,j\}}, U_{\{i,j\}}))_{i,j \in \mathbb{N}} \tag{69}$$

for some measurable function f.

Example

Elements $X_{i,j}$ i.i.d.
Separately exchangeable: $f(a, b, c, d) = g(d)$.

Example

$X_{i,j} = 1$ if $i = j$ and $X_{i,j} = 0$ otherwise.
Jointly exchangeable: $f(a, b, c, d) = 1(b == c)$.

Example

Let X be Θ-random.
Jointly exchangeable: $f(a, b, c, d) = 1(d < \Theta(b, c))$.

Let $Y_1, Y_2, \ldots$ be an exchangeable sequence. Define $X_{i,j} = Y_j$.

What does Aldous-Hoover tell us about Y?

Definition (de Finetti in Aldous-Hoover form)

Let U_i be i.i.d. Uniform random variables.
An infinite sequence $Y_1, Y_2, \ldots$ is exchangeable if and only if

$$(Y_i)_{i \in \mathbb{N}} \overset{\mathrm{d}}{=} (f(U_0, U_i))_{i \in \mathbb{N}} \tag{70}$$

for some measurable function f.

Exchangeability and statistics: Not just Bayesian!

Exchangeable observation, version 1

Let $(f(U_{0,0}, U_{i,0}, U_{0,j}, U_{i,j}))_{i,j \in \mathbb{N}}$ be a (partially-observed) exchangeable array, where f is unknown and $(U_{i,j})_{i,j \in \mathbb{N}}$ i.i.d. Uniform.

Define $F(b, c, d) = f(U_{0,0}, b, c, d)$. F is a random measurable function.

Exchangeable observation, version 2

Let $(F(U_{i,0}, U_{0,j}, U_{i,j}))_{i,j \in \mathbb{N}}$ and $F \sim \mu$,
where μ is unknown and $(U_{i,j})_{i,j \in \mathbb{N}}$ i.i.d. Uniform.

Problem 1: Even observing entire array X reflects only one realization F from μ.
Solution: Dissociated arrays (F non-random). These are the ergodic measures.
Problem 2: Random functions of the form $F(b, c, d) = G(b, c)$ are "dense".
Solution: Move to simple arrays, i.e., last parameter is not used.

Simple dissociated array observation

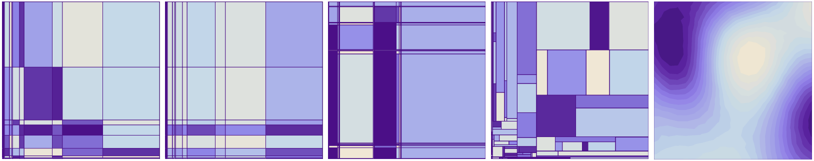
Let U_i, V_j be i.i.d. Uniform random variables.
Let $(F(U_i, V_j))_{i,j \in \mathbb{N}}$ be a simple dissociated array, unknown F.

Bayesian approach

Let F be a random measurable function with some prior.

Then $X_1, X_2, \ldots$ is an exchangeable array.

In joint work with Orbanz [OR13], we show that many nonparametric models of graphs/networks can be recast as prior distributions on random functions F.



1. Infinite Relational Model (IRM) of Kemp et al. (2008) based on Chinese restaurant process (Dirichlet process).
2. IRM where the interaction probabilities are also an exchangeable array. [OR]
3. Infinite Feature Relational Model (IFRM) of Miller et al. (2010) based on Indian buffet process (Beta process).
4. Mondrian process-based relational model of R. and Teh (2009).
5. Gaussian-process-based relational model. Lloyd et al. NIPS 2012 show how many factorization models fit into this framework.

We have...

1. Reviewed exchangeability for sequences, graphs, and arrays
2. Presented de Finetti and Aldous-Hoover representation theorems.
   X exchangeable if and only if $X \overset{d}{=} (F(...))$ for a random F.
3. Discussed how to interpret de Finetti and Aldous-Hoover.
   ν / Θ / F is the limiting empirical distribution/graphon/array. ν unique. Θ and F only unique up to a m.p.t.
4. Shown how probabilistic symmetries lead to statistical models.

More reading

1. Orbanz and Roy. *Bayesian Models of Graphs, Arrays and Other Exchangeable Random Structures*.
   Preprint available at `http://danroy.org`
2. Aldous. *Representations for partially exchangeable arrays of random variables.*
3. Kallenberg. *Probabilistic symmetries and invariance principles.*