The Aldous–Hoover representation theorem and applications to modeling relational data

James Lloyd

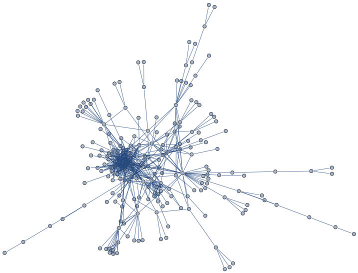
University of Cambridge

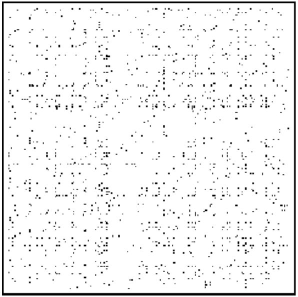
January 2013

Collaborators
Daniel M. Roy (Cambridge)
Peter Orbanz (Columbia)
Zoubin Ghahramani (Cambridge)

Relational data: definition

Anything measured at more than one type of 'object'



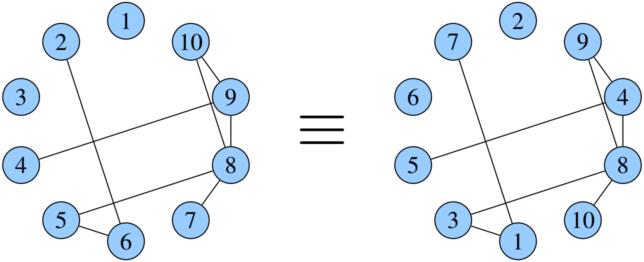


In full generality, anything that can be stored in a relational database

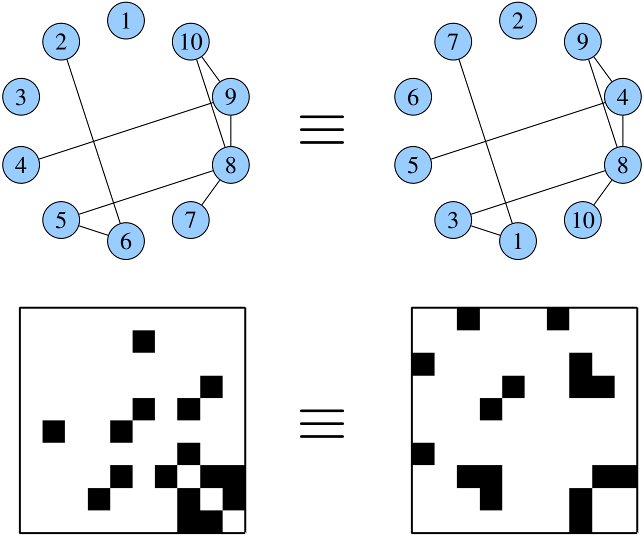
How can we model such data?

- Interested in generative modeling of such data for e.g.,
  - Discovery of latent structure e.g., groups of proteins with similar functions in protein-protein interactomes
  - Prediction of missing data e.g., movie recommendation, friend suggestions
- Relational data typically encoded in arrays. How do reasonable assumptions about the data translate to the array representation
- We make a weak assumption and demonstrate the implied structure for arrays
  - Implied structure allows for classification of many models
  - Also inspires a simple Bayesian nonparametric model with good empirical performance

Exchangeability for relational data



EXCHANGEABILITY FOR CORRESPONDING ARRAYS



EXCHANGEABILITY CAN BE CHARACTERISED

Definition

An array $X = (X_{ij})_{i,j \in \mathbb{N}}$ is called an *exchangeable array* if

$$(X_{ij}) \overset{\mathrm{d}}{=} (X_{\pi(i)\pi(j)}) \qquad \text{for every } \pi \in \mathbb{S}_\infty \ .$$

Theorem (Aldous, Hoover)

A random 2-array (X_{ij}) is exchangeable if and only if there is a random (measurable) function $F : [0,1]^3 \to \mathcal{X}$ such that

$$(X_{ij}) \overset{d}{=} (F(U_i, U_j, U_{ij})).$$

for every collection $(U_i)_{i \in \mathbb{N}}$ and $(U_{ij})_{i \leq j \in \mathbb{N}}$ of i.i.d. Uniform$[0,1]$ *random variables, where $U_{ji} = U_{ij}$ for $j < i \in \mathbb{N}$.*

An arbitrarily good approximation

This representation can be simplified

Call an array (X_{ij}), *simple* if it admits a representation

$$(X_{ij}) \stackrel{d}{=} (\Theta(U_i, U_j))$$

Let $\mathcal{L}(Y)$ be the law (distribution) of a random variable Y and define $\chi_m X := (X_{ij};\ i, j \leq m)$.

Theorem (Kallenberg)

Let X be a d-dimensional exchangeable array in a Borel space $\mathcal{X}$. Then there exist some simple exchangeable arrays $X_1, X_2, \ldots$ such that $\mathcal{L}(\chi_m X_n)$ and $\mathcal{L}(\chi_m X)$ are mutually absolutely continuous for all $m, n \in \mathbb{N}$ and the associated Radon–Nikodym derivatives converge uniformly to 1 as $n \to \infty$ for fixed m.

THIS INSPIRES A BAYESIAN NONPARAMETRIC MODEL

We decompose the function F into two functions $\Theta : [0,1]^2 \to \mathcal{W}$ and $H : [0,1] \times \mathcal{W} \to \mathcal{X}$ for a suitable space $\mathcal{W}$, such that

$$(X_{ij}) \stackrel{d}{=} (F(U_i, U_j, U_{ij})) = (H(U_{ij}, \Theta(U_i, U_j))) \; .$$

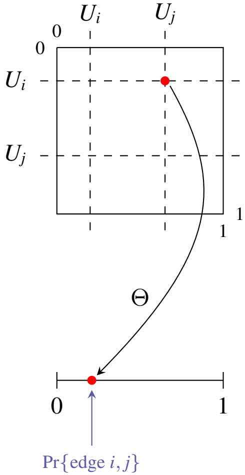
Inspiring the following generative model

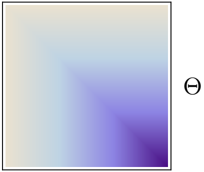
$$\begin{aligned} \Theta &\sim \mathcal{GP}(0, \kappa) \\ U_1, U_2, \ldots &\sim_{\text{iid}} \text{Uniform}[0,1] \\ X_{ij} \,|\, W_{ij} &\sim P[\,.\,|W_{ij}] \end{aligned}$$

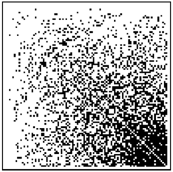
where $W_{ij} = \Theta(U_i, U_j)$.

$$\Theta : [0,1]^2 \longrightarrow [0,1] \text{ measurable and symmetric} \qquad U_1, U_2, \ldots \sim_{\text{iid}} \text{Uniform}[0,1]$$

$$\Pr\{\text{edge } i,j\} = \Theta(U_i, U_j)$$







Many models fit this pattern

Graph data			
Random function model	Θ	$\sim$	$\mathcal{GP}(0, \kappa)$
Latent class	W_{ij}	$=$	$\Lambda_{U_i U_j}$ where $U_i \in \{1, \ldots, K\}$
IRM	W_{ij}	$=$	$\Lambda_{U_i U_j}$ where $U_i \in \{1, \ldots, \infty\}$
Latent distance	W_{ij}	$=$	$-\lvert U_i - U_j \rvert$
Eigenmodel	W_{ij}	$=$	$U_i' \Lambda U_j$
LFRM	W_{ij}	$=$	$U_i' \Lambda U_j$ where $U_i \in \{0, 1\}^\infty$
ILA	W_{ij}	$=$	$\sum_d \mathbb{I}_{U_{id}} \mathbb{I}_{U_{jd}} \Lambda^{(d)}_{U_{id} U_{jd}}$ where $U_i \in \{0, \ldots, \infty\}^\infty$
SMGB	Θ	$\sim$	$\mathcal{GP}(0, \kappa_1 \otimes \kappa_2)$

Real-valued array data			
Random function model	Θ	$\sim$	$\mathcal{GP}(0, \kappa)$
Mondrian process based	Θ	=	piece-wise constant random function
PMF	W_{ij}	$=$	$U_i' V_j$
GPLVM	Θ	$\sim$	$\mathcal{GP}(0, \kappa \otimes \delta)$

A CORRESPONDENCE RESULT

Proposition

A matrix factorization model defined as

$$W_{ij} = U_i' \Lambda V_j \qquad \Lambda_{ij} \sim_{iid} \mathcal{N}(0, 1)$$

is equivalent to

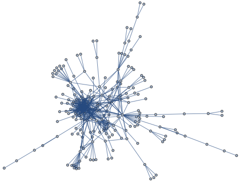
$$W_{ij} = \Theta\left(U_i, V_j\right) \qquad \Theta \sim \mathcal{GP}\left(0, L_U \otimes L_V\right)$$

where $L_U(U_{i_1}, U_{i_2}) = U_{i_1}' U_{i_2}$ *and similarly for* L_V.

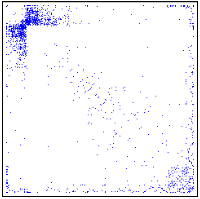
A simpler overview of model space

	W_{ij}	κ	$U_i, V_j \sim$.
Random function model	$\phi(U_i, V_j)'\Lambda$	$\kappa_{U\times V}$	Gaussian
SMGB, InfTucker	$\phi(U_i)'\Lambda\phi(V_j)$	$\kappa_U \otimes \kappa_V$	Laplace
GPLVM	$\phi(U_i)'\Lambda$	$\kappa_U \otimes \delta_V$	Gaussian
Eigenmodel	$U_i'\Lambda V_j$	$L_U \otimes L_V$	Gaussian
Linear relational GP	$U_i'\Lambda V_j$	$L_U \otimes L_V$	Gaussian
PCA, PMF	$U_i'\Lambda$	$L_U \otimes \delta_V$	Gaussian
Latent distance	$-\|U_i - U_j\|$	0	Gaussian
Mondrian process based	Decision tree	*	Uniform
Latent class	$\Lambda_{U_i U_j}$	$\delta_{U\times U}$	Multinomial
IRM	$\Lambda_{U_i V_j}$	$\delta_{U\times V}$	CRP
IHRM	$\Lambda_{U_i V_j}$	$\delta_{U\times V}$	CRP
BMF	$U_i'\Lambda V_j$	$L_U \otimes L_V$	IBP
LFRM	$U_i'\Lambda U_j$	$L_U \otimes L_U$	IBP
ILA	$\sum_d \mathbb{I}_{U_{id}} \mathbb{I}_{U_{jd}} \Lambda^{(d)}_{U_{id} U_{jd}}$	*	CRP + IBP

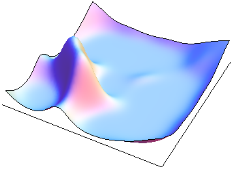
POSTERIOR



A protein interactome



Adjacency matrix sorted by MAP embedding



MAP Θ

ONGOING RESEARCH / IDEAS

- Modeling multiple arrays e.g., joint modelling of social network and ‘like’ data
  - Corollaries of Aldous–Hoover suggest representations for such data
  - Many unanswered questions about generating good models
- Trying new priors on functions
  - Many priors on functions for sequential data that could have utility for relational data
  - e.g., Analogous versions of k-means, mixture of Gaussians?
- Trying new priors on latent variables
  - CRP + IBP prior in ILA could be more broadly applicable

Corollary

Let $(X_{ij})_{i,j\in\mathbb{N}}$ and $(C_i)_{i\in\mathbb{N}}$ be random variables in $\mathcal{X}$ and $\mathcal{X}'$ respectively. Then the following are equivalent:

i. $(X_{ij}, C_i) \overset{d}{=} (X_{\pi(i)\pi(j)}, C_{\pi(i)})$ *for every* $\pi \in \mathbb{S}_\infty$.

ii. *There are random (measurable) functions* $F : [0,1]^3 \to \mathcal{X}$ *and* $G : [0,1] \to \mathcal{X}'$ *such that*

$$(X_{ij}, C_i) \overset{d}{=} (F(U_i, U_j, U_{ij}), G(U_i)), \tag{1}$$

for every collection $(U_i)_{i\in\mathbb{N}}$ *and* $(U_{ij})_{i\leq j\in\mathbb{N}}$ *of i.i.d.* $\mathrm{Uniform}[0,1]$ *random variables, where* $U_{ji} = U_{ij}$ *for* $j < i \in \mathbb{N}$.

Consider rating data (X_{ij}) with users i and movies j, with side information in the form of covariates for both users, C_i, and movies, D_j, and a social network (S_{ik}) over users i, k.

Corollary

The following are equivalent

i. $(X_{ij}, C_i, D_j, S_{ik}) \overset{d}{=} (X_{\pi(i)\pi'(j)}, C_{\pi(i)}, D_{\pi'(j)}, S_{\pi(i)\pi(k)})$ *for every* $\pi, \pi' \in \mathbb{S}_\infty$.

ii. *There exist random functions* F, G, H, I *such that*

$$(X_{ij}, C_i, D_j, S_{ik}) \overset{d}{=} (F(U_i, V_j, W_{ij}), G(U_i), H(V_j), I(U_i, U_k, U_{ik})) \qquad (2)$$

for every collection $(U_i)_{i\in\mathbb{N}}, (V_j)_{j\in\mathbb{N}}, (W_{ij})_{i,j\in\mathbb{N}}$ *and* $(U_{ik})_{i\leq k\in\mathbb{N}}$ *of i.i.d.* Uniform$[0, 1]$ *random variables, where* $U_{ki} = U_{ik}$ *for* $k < i \in \mathbb{N}$.

Data

- A friend of friends network collected from last.FM (S_{ik})
- A user $\times$ genre matrix: $X_{ij} = 1$ iff user i has listened to genre j

Cold start task

- Want to predict entire rows of X_{ij} i.e., recommendations for new users
- Consider jointly modelling the array

Preliminary numerical results promising

Insert a table and some comparisons

Multiple arrays: many open questions

- Which designs of model will effectively model multiple arrays without having to 'balance' or compromise?
  - Flat clustering models seem especially inappropriate e.g., IRM
  - Multiple clustering models seem well suited
  - How does this transfer to GP case - in particular, prior on length scales
- Is generative modelling appropriate, or can we find more efficient models of conditional densities?
  - What are appropriate representations for conditional densities?

POTENTIAL FUTURE RESEARCH - 1-ARRAY -> 2-ARRAY

e.g., Mixture of basis functions (motivate via Mondrian)
Relational *k*-means
Must be something interesting

Table

APPENDIX: RFM POSTERIOR

Pictures

APPENDIX: RFM INFERENCE

Words and maths