SPARSE THREE-PARAMETER RESTRICTED INDIAN BUFFET PROCESS FOR UNDERSTANDING INTERNATIONAL TRADE

Melanie F. Pradier[1]
Fernando Perez-Cruz[1,2]

[1]University Carlos III in Madrid
[2]Stevens Institute of Technology
melanie@tsc.uc3m.es

Viktor Stojkoski[3]
Zoran Utkovski[3,4]
Ljupco Kocarev[3,5]

[3]Macedonian Academy of Sciences and Arts
[4]University Goce Delcev in Stip
[5]University of California in San Diego

INTRODUCTION

- **Aim**: Explore high-dimensional count data.
  a) Increase model interpretability.
  b) Find structured solutions in latent space.
- **Contribution**: A Bayesian non-parametric Poisson factorization model that gives easy-to-interpret and structured solutions.
- **Key Idea**: Force sparsity in the features and improve prior flexibility to be consistent with reality, by combining the stable-beta process with the restricted Indian Buffet Process.

THEORETICAL BACKGROUND

Indian-Buffet Process (Ghahramani et.al, 2006)

- Stochastic process defining a probability distribution over equivalent classes of binary matrices. We denote: $\mathbf{Z} \sim \mathrm{IBP}(\alpha)$.
- It corresponds to the limit when $K \to \infty$ of parametric model:

$$\pi_k \sim \mathrm{Beta}(\alpha/K, 1),$$
$$z_{nk} \sim \mathrm{Bernoulli}(\pi_k) \quad (1)$$

- It can also be constructed based on its underlying De Finetti's representation, i.e., as a mixture of Bernoulli processes directed by a beta process:

$$\mu \sim \mathrm{BP}(1, \alpha, H) \quad (2)$$
$$\mathbf{Z}_{n\cdot} \sim \mathrm{BeP}(\mu) \quad (3)$$

where $\mu = \sum_k \pi_k \delta_{\theta_k}$ is the directing measure, and H is the probability base measure (Thibaux et.al, 2007).

- Disadvantage: Mass parameter α couples both a priori number of ones per row J_n and total number of active features K^+.

$$J_n \sim \mathrm{Poisson}(\alpha) \quad (4)$$
$$K^+ \sim \mathrm{Poisson}(\alpha \sum_{n=1}^{N} \left(\frac{1}{n}\right)) \quad (5)$$

SPARSE 3-PARAMETER RESTRICTED IBP (S3R-IBP)

- Combine strengths of three-parameter IBP and restricted IBP:

$$\mu \sim \mathrm{SBP}(1, \alpha, H) \quad (9)$$
$$\mathbf{Z}_{n\cdot} \sim \text{R-BeP}(\mu, f) \quad (10)$$

We denote this flexible prior as $\mathbf{Z} \sim \text{S3R-IBP}(\alpha, c, \sigma, f)$.

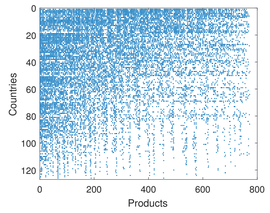
- Let $\mathbf{X} \in \mathbb{N}^{N \times D}$, N samples, and D dimensions.
- We build a structured infinite latent feature model for count data:

$$x_{nd} \sim \mathrm{Poisson}(\mathbf{Z}_{n\cdot}\mathbf{B}_{\cdot d}), \quad (11)$$
$$B_{kd} \sim \mathrm{Gamma}(\alpha_B, \frac{\mu_B}{\alpha_B}), \quad (12)$$
$$\mathbf{Z} \sim \text{3R-IBP}(\alpha, c, \sigma, f) \quad (13)$$

where α_B and μ_B are the shape and mean of the prior Gamma distribution.

- Available parameters:
  - mass parameter α
  - concentration parameter $c > -\sigma$
  - stability component $\sigma \in [0, 1)$
  - marginal prior f for J_n
- Features are made sparse by choosing $\alpha_B < 1$.

Motivation: Why some countries are wealthier than others?
Theory of Economic Complexity: Capabilities are "intangible assets which drive the development, wealth and competitiveness of a country" (Cristelli et.al, 2013).



- Triangular structure
- Diversified countries producing exclusive products
- Non-diversified countries producing standard products

Three-parameter IBP (Teh et.al, 2009)

- More flexible distribution for stick weights (power-law behaviors).
- In the De Finetti's representation, it uses a Stable-beta process (SBP).
- Culinary Metaphor:
  - Customer 1 tries $\mathrm{Poisson}(\alpha)$ dishes.
  - Customer n tries:

$$p(z_{nk} = 1 | \mathbf{Z}_{\neg n}) = \frac{m_k - \sigma}{n + c - 1} \quad (6)$$
$$p(J_{new}) \sim \mathrm{Poisson}\left(\alpha \frac{\Gamma(1+c)\Gamma(n+c+\sigma-1)}{\Gamma(n+c)\Gamma(c+\sigma)}\right) \quad (7)$$

- Disadvantage: Number of ones per row J_n still Poisson-distributed.

Restricted IBP (Doshi-Velez et.al, 2015)

- Non-exchangeable, with arbitrary marginal prior f over J_n
- In the De Finetti's representation, it uses *restricted* Bernoulli processes:

$$\text{R-BeP}(\mathbf{Z}_{n\cdot}; \mu, f) = f(J_n) \cdot \frac{\prod_{k=1}^{\infty} \pi_k^{z_{nk}}(1-\pi_k^{1-z_{nk}}) 1(\sum_K z_{nk} = J_n)}{\sum_{z' \in \mathcal{Z}} \prod_k \pi_k^{z'_k}(1-\pi_k)^{(1-z'_k)} 1(\sum_K z'_k = J_n)} \quad (8)$$

- Disadvantage: Stick weights cannot follow power-law behaviors.

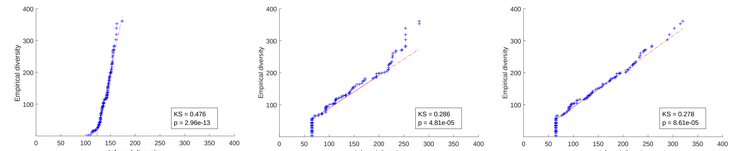
Inference Scheme

- Model conditionally conjugate: auxiliary variables $x'_{nd,1}, \ldots, x'_{nd,K}$ such that $x_{nd} = \sum_{k=1}^{K} x'_{nd,k}$, and $x'_{nd,k} \sim \mathrm{Poisson}(Z_{nk}B_{kd})$
- For each iteration, do:
  1: Sample each element of matrix $\mathbf{Z}$ using inclusion probabilities (Aires, 1999).
  2: Sample latent measure $\boldsymbol{\pi}$ using Metropolis-Hasting within Gibbs (Doshi-Velez et.al, 2015). Use posterior distribution from the standard IBP as proposal $\mathcal{Q}$.

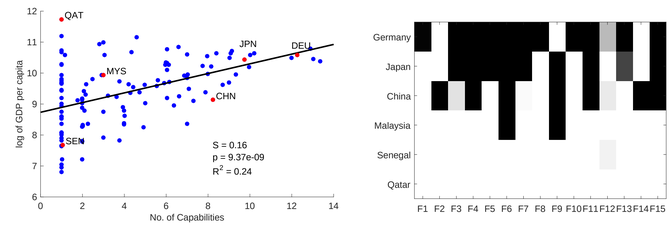
$$\mathcal{Q}(\pi_k | \mathbf{Z}_{\cdot k}) \propto \mathrm{Beta}(\frac{\alpha}{K} + \sum_{n=1}^{N} z_{nk} - \sigma,\ 1 + N - \sum_{n=1}^{N} z_{nk} + c + \sigma) \quad (14)$$
$$\mathcal{Q}(\pi_{k_{new}} | \mathbf{Z}_{\cdot k}) \propto \alpha \frac{\Gamma(1+c)}{\Gamma(1-\sigma)\Gamma(c+\sigma)} \pi_{k_{new}}^{-\sigma-1}(1-\pi_{k_{new}})^{c+N+\sigma-1} \quad (15)$$

  Sample from eq. 2 using Adaptive Rejection Metropolis Sampling (Martino et.al, 2015).
  Let us call $D_{J_n}^K$ the denominator in eq. 8: $D_{J_n}^K = (1-\pi_K)D_{J_n}^{K-1} + \pi_K D_{J_n-1}^{K-1}$.
  3: Sample each element of $\mathbf{B}$ and $\mathbf{X}'$ from their conditional distributions.
  4: Sample hyperparameter α according to (Escobar et.al, 1995).

RESULTS



(a) Baseline (Hidalgo et.al, 2009) (b) S-IBP (c) Our model S3R-IBP



Id	$\bar{m}_k$	Top-5 products with sorted highest weights (B_{kd}) associated	Repr. countries ($\bar{J}_n$)
F1	18.27	Miscellaneous Animal Oils (0.78), Bovine and Equine Entrails (0.72), Bovine meat (0.68), Preserved Milk (0.63), Equine (0.62)	Paraguay (2.00)
F2	21.39	Synthetic Woven Fabrics (0.74), Non-retail Synthetic Yarn (0.60), Woven Fabric of less than 85% Discontinuous Synthetic Fibres (0.60), Woven Fabrics of More Than 85% Discontinuous Synthetic Fiber (0.58), Yarn of Less Than 85% Synthetic Fibers (0.53)	United Arab Emirates (2.82)
F3	14.87	Parts of Metalworking Machine Tools (0.74), Interchangeable Tool Parts (0.72), Polishing Stones (0.69), Tool Holders (0.66), Miscellaneous Metalworking Machine-Tools (0.54)	Israel (5.97)
F4	18.67	Aldehyde, Ketone and Quinone-Function Compounds (0.68), Glycosides and Vaccines (0.67), Medicaments (0.65), Inorganic Esters (0.64), Cyclic Alcohols (0.62)	Ireland (4.34)
F5	11.04	Synthetic Rubber (0.87), Acrylic Polymers (0.85), Silicones (0.76), Miscellaneous Polymerization Products (0.71), Tinned Sheets (0.65)	North Korea (3.99)
F6	21.95	Measuring Controlling Instruments (0.61), Mathematical Calculation Instruments (0.59), Miscellaneous Electrical Instruments (0.57), Miscellaneous Heating and Cooling Equipment (0.51), Parts of Office Machines (0.49)	Malaysia (3.00)
F7	31.14	Vehicles Parts and Accessories (0.59), Cars (0.58), Iron Wire (0.53), Trucks and Vans (0.53), Air Pumps and Compressors (0.50)	Belarus (4.20)
F8	33.00	Improved Wood (0.71), Mineral Wool (0.62), Central Heating Equipment (0.62), Aluminium Structures (0.62), Harvesting Machines (0.60)	Belarus (4.20)
F9	16.53	Miscellaneous Electrical Machinery (0.76), Vehicles Stereos (0.72), Miscellaneous Data Processing Equipment (0.64), Video and Sound Recorders (0.57), Calculating Machines (0.55)	Malaysia (3.00)
F10	45.93	Baked Goods (0.67), Metal Containers (0.62), Miscellaneous Edibles (0.59), Miscellaneous Articles of Paper (0.59), Miscellaneous Organic Surfactants (0.58)	Costa Rica (2.06)
F11	33.23	Miscellaneous Articles of Iron (0.65), Carpentry Wood (0.61), Miscellaneous Manufactured Wood Articles (0.60), Sawn Wood Less Than 5mm Thick (0.56), Electric Current (0.51)	Russia (2.93)
F12	38.67	Vegetables (0.60), Fruit or Vegetable Juices (0.54), Miscellaneous Fruit (0.50), Frozen Vegetables (0.48), Apples (0.47)	Peru (2.00)
F13	23.29	Miscellaneous Pumps (0.51), Ash and Residues (0.45), Chemical Wood Pulp of sulphite (0.44), Rolls of Paper (0.43), Worked Nickel (0.43)	Russia (2.93)
F14	46.11	Synthetic Knitted Undergarments (0.76), Miscellaneous Feminine Outerwear (0.74), Miscellaneous Knitted Outerwear (0.73), Men's Shirts (0.70), Blouses (0.67)	Sri Lanka (2.00)
F15	32.12	Miscellaneous Rotating Electric Plant Parts (0.66), Control Instruments of Gas or Liquid (0.58), Valves (0.57), Miscellaneous Rubber (0.56), Miscellaneous Articles of Plastic (0.55)	Philippines (4.01)

We run a 2nd layer of the S3R-IBP model which grouped capabilities in three groups. Blue ≡ "developing"; Black ≡ "developed", green ≡ "mixed".

CONCLUSIONS

1. BNP model for data exploration in high-dim data
2. **interpretable** and **structured** solutions.
3. Analysis of productive structure of world economies.

WORK IN PROGRESS

- **Improve mixing**, coverage of the posterior (split and merge moves).
- **Time-dependent extension** with Markovian activation of features and smooth variation of capabilities.