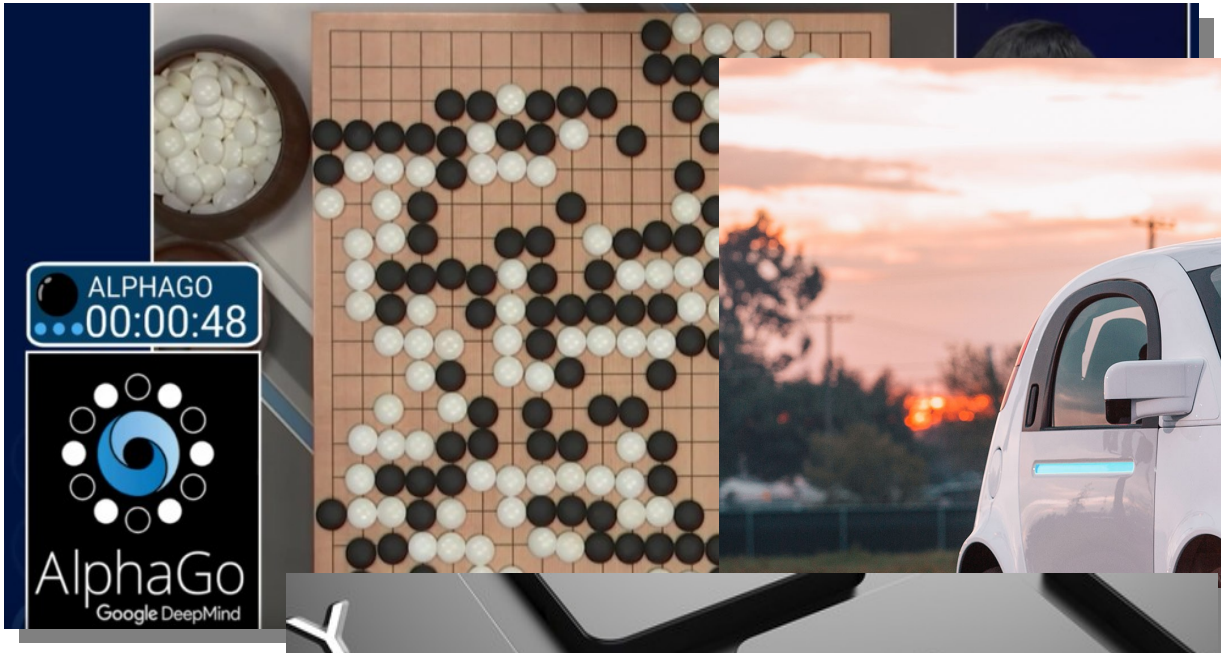
A Bayesian Non-parametric Approach to Understand World Economies

Melanie F. Pradier, Viktor Stojkoski, Zoran Utkovski,
Ljupco Kocarev, and Fernando Perez-Cruz

Friday 17, 2017

- High-dimensional count data.
- Focus on **Data Exploration**.



On top of accuracy...

- Safety
- Fairness
- Understanding

We need **Interpretability**

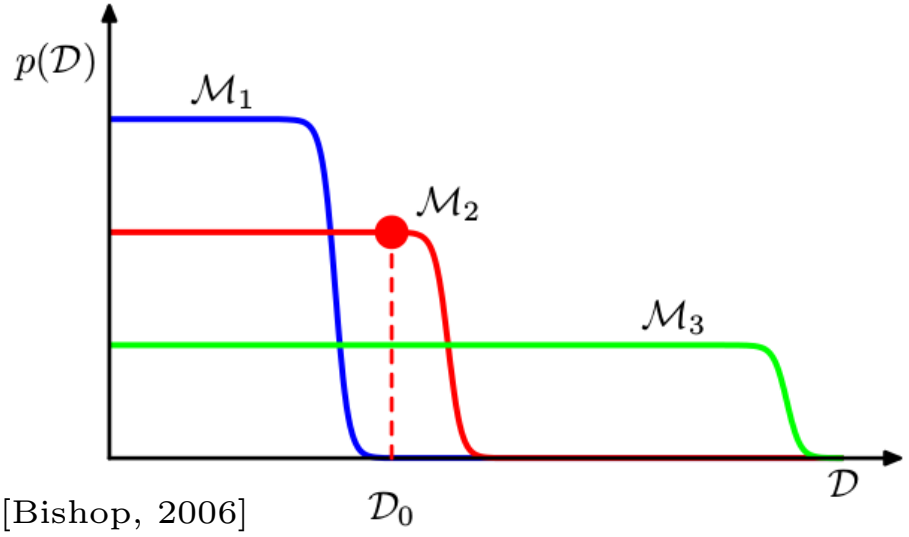
Interpretability

[F. Doshi-Velez, B. Kim, *Towards A Rigorous Science of Interpretable Machine Learning*]

- Understandable for humans (Doshi-Velez, 2017)
- "Right to explanation" (European Legislation, 2018)

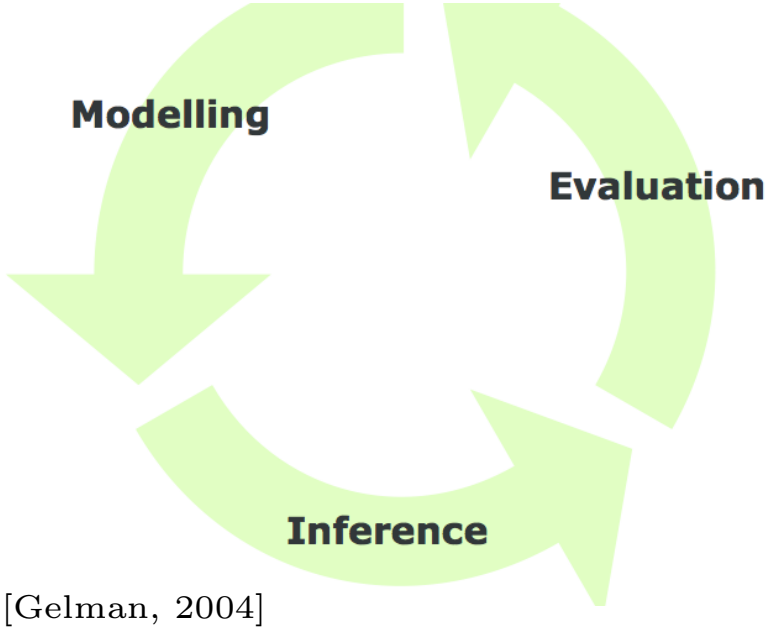
Focus: Data Exploration

- Adequate model assumptions
- Constrainted solution spaces



[Bishop, 2006]

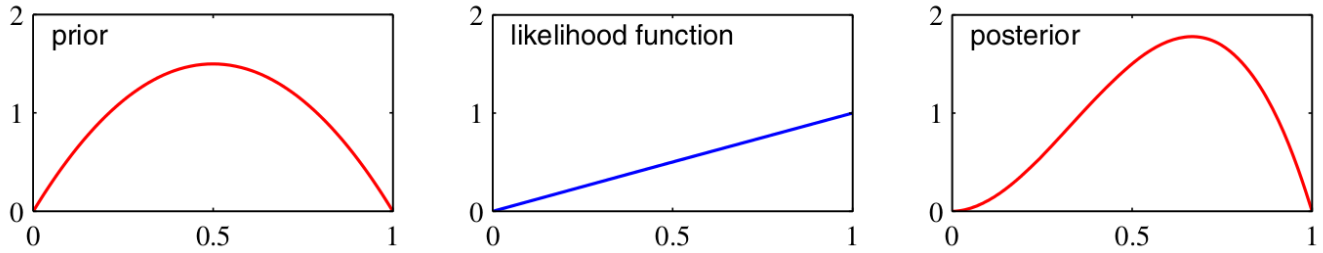
Multidisciplinary Research



[Gelman, 2004]

Bayesian Non-Parametric Models

- Bayesian: Combine Prior Knowledge with Data Evidence



[Bishop, 2006]

- Non-parametric
  - actually... really large parametric model
  - number of latent variables grows with data

During my PhD...

- Marathon modeling
- Biomarker discovery
- **Economic complexity**

Outline

Motivation: Wealth of Nations

What makes some countries wealthier than others?

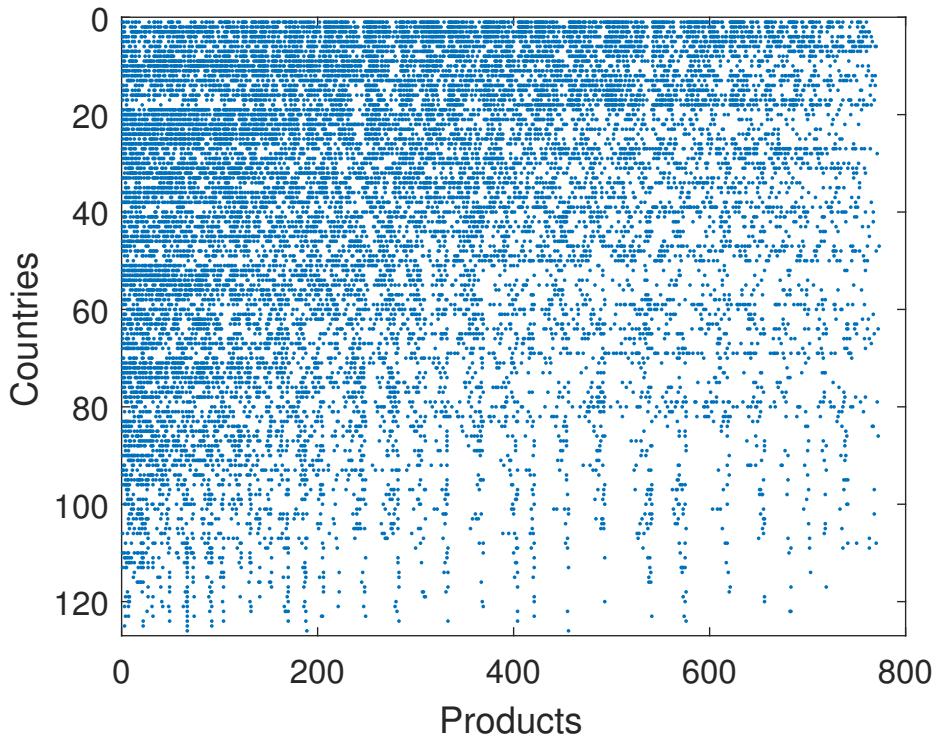
Classical view

- Division of labor (Ricardo, 1817; A. Smith, 1976)
- Specialization leads to economic efficiency
- Export portfolios

  $\rightarrow$ block-structure

Motivation: Wealth of Nations

The reality:



Properties:

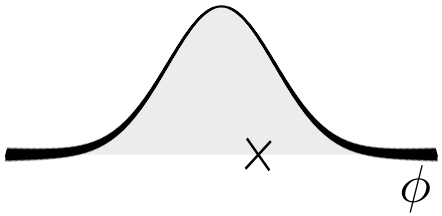
1. Triangularity
2. $D \gg N$

Our Contribution

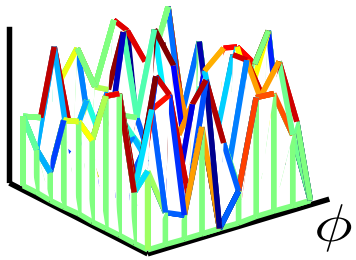
Develop an Infinite Poisson-Gamma Model

- Flexible prior
- Feature sparsity

Indian Buffet Process (Ghahramani et.al, 2006)



$$\xrightarrow[x \sim \mathcal{N}(\mu_0, \sigma_0^2)]{} \quad x = 0.356$$

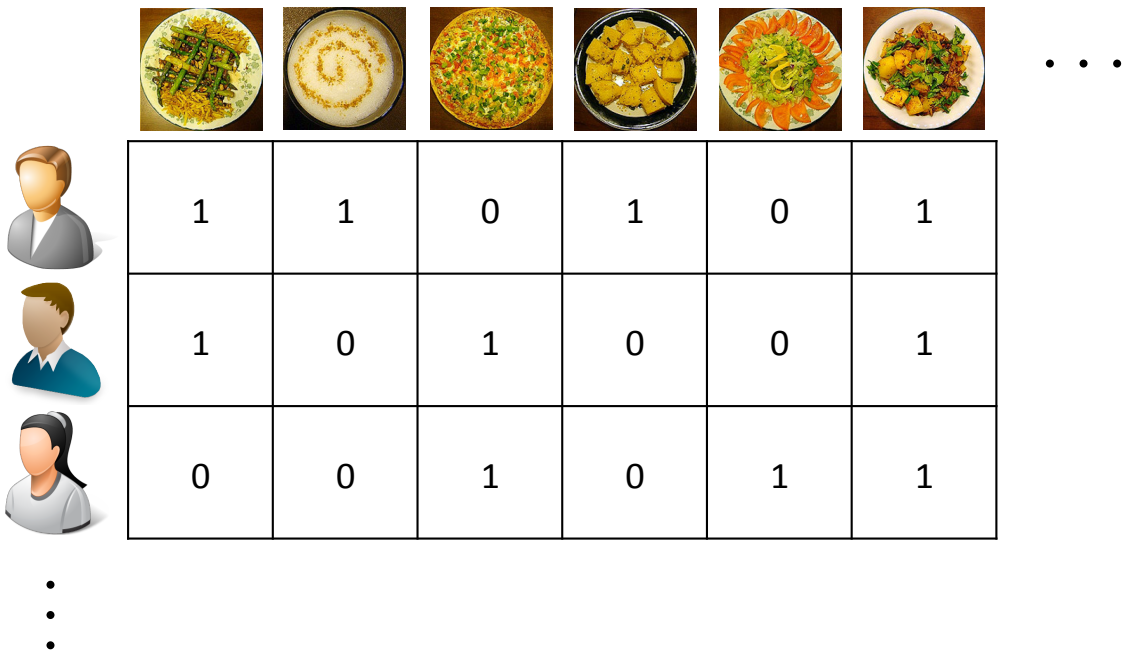


$$\xrightarrow[Z \sim \text{IBP}(\alpha)]{} \quad Z = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \vdots$$

- IBP: distribution over binary matrices $Z_{N \times K}$
- Model chooses number of hidden features, $K \to \infty$

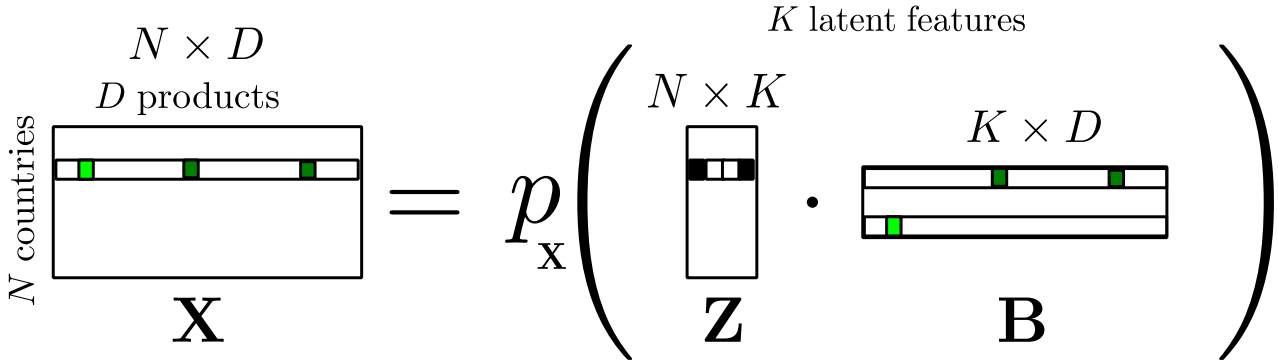
Indian Buffet Process

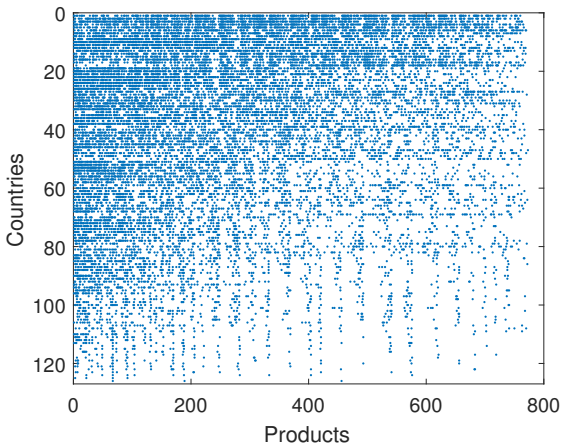
(Slides from F.J.R. Ruiz)



1	1	0	1	0	1
1	0	1	0	0	1
0	0	1	0	1	1

Our Approach





Generative Model

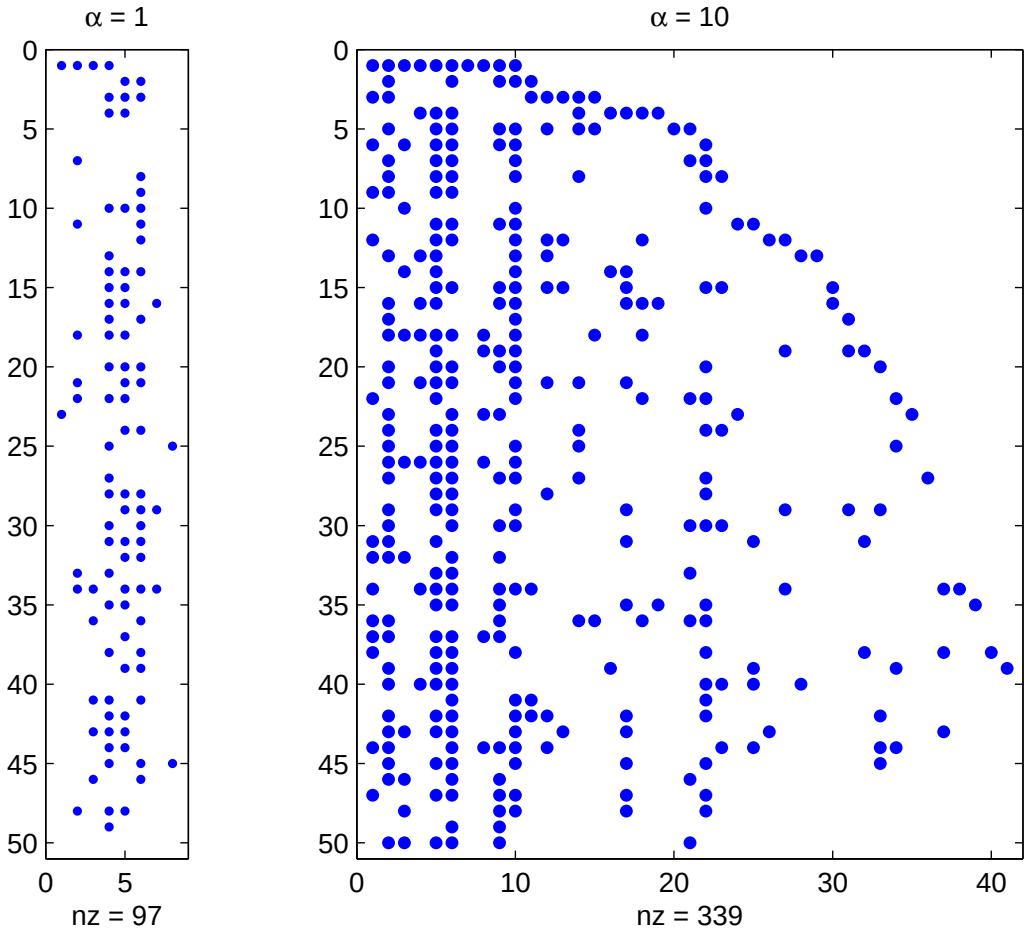
$$x_{nd} \sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\right) \quad (1)$$

$$B_{kd} \sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right) \quad (2)$$

$$\mathbf{Z}_{n\bullet} \sim \text{IBP}(\alpha) \quad (3)$$

A limitation of the IBP

- **Disadvantage**: Mass parameter α couples both J_n and K^+



Beyond the standard IBP

Three-parameter IBP (Teh et.al, 2007)

- More flexible distribution for feature weights

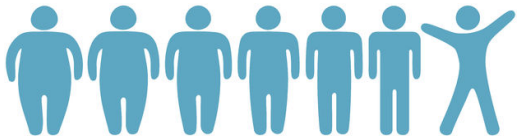
$$\mathbf{Z}_{n\bullet} \sim \text{BeP}(\mu) \quad (4)$$
$$\mu \sim \text{SBP}(1, \alpha, H, c, \sigma) \quad (5)$$

$$p(J_{new}) \sim \text{Poisson}\left(\boldsymbol{\alpha}\frac{\Gamma(1+\mathbf{c})\Gamma(n+\mathbf{c}+\boldsymbol{\sigma}-1)}{\Gamma(n+\mathbf{c})\Gamma(\mathbf{c}+\boldsymbol{\sigma})}\right)$$

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

- Arbitrary prior f over J_n

$$\mathbf{Z}_{n\bullet} \sim \text{R-BeP}(\mu, f) \quad (6)$$
$$\mu \sim \text{BP}(1, \alpha, H) \quad (7)$$

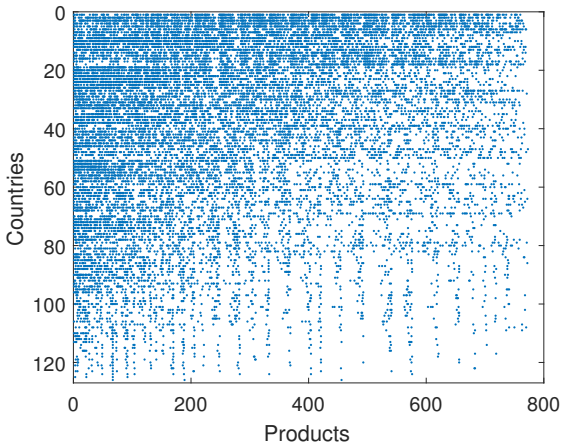


- Combination of both
- Flexible prior

Our Approach

$$\underset{N \times D}{\mathbf{X}} = p_{\mathbf{X}}\left(\underset{N \times K}{\mathbf{Z}} \cdot \underset{K \times D}{\mathbf{B}} \right)$$

D products; N countries; K latent features



Generative Model

$$x_{nd} \sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\right) \quad (8)$$

$$B_{kd} \sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right) \quad (9)$$

$$\mathbf{Z}_{n\bullet} \sim \text{3R-IBP}(\alpha, c, \sigma, f) \quad (10)$$

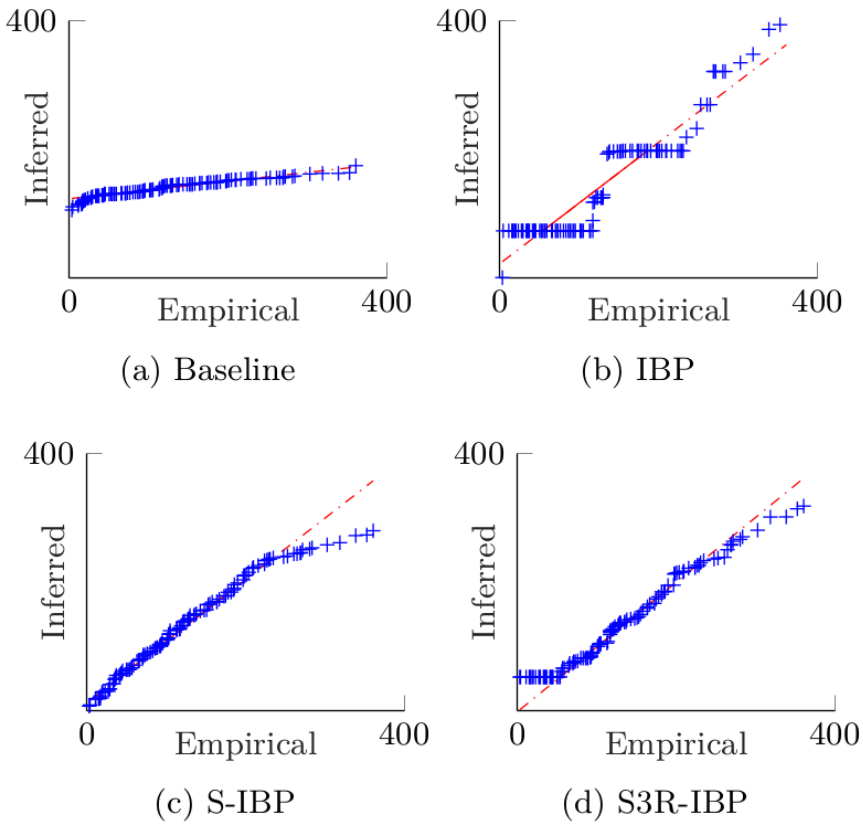
Results

Metric	IBP	S-IBP	S3R-IBP
Log Perplexity	$\mathbf{2.91 \pm 0.17}$	15.29 ± 1.84	3.03 ± 0.04
Coherence	-149.36 ± 7.56	-178.44 ± 4.50	$\mathbf{-140.51 \pm 2.73}$
Lift	120.36 ± 3.66	121.77 ± 0.71	$\mathbf{125.28 \pm 0.50}$

(a) SITC database ($N = 126$, $D = 744$)

Metric	IBP	S-IBP	S3R-IBP
Log Perplexity	$\mathbf{2.85 \pm 0.06}$	8.75 ± 0.95	2.88 ± 0.07
Coherence	-148.91 ± 10.57	-168.39 ± 13.16	$\mathbf{-134.51 \pm 4.4}$
Lift	152.12 ± 6.85	157.38 ± 2.20	$\mathbf{161.80 \pm 0.50}$

(b) HS database ($N = 123$, $D = 4890$)



(a) Baseline (b) IBP

(c) S-IBP (d) S3R-IBP

Results
Interpretability

F0: Bias	F1: Agriculture	F2: Clothing I	F3: Farming	F4: Clothing II
Non-Coniferous Worked Wood	Vegetables	Synthetic Knitted Undergarments	Misc. Animal Oils	Synthetic Woven Fabrics
Bran and Other Cereals Residues	Fruit or Vegetable Juices	Misc. Feminine Outerwear	Bovine and Equine Entrails	Non-retail Synthetic Yarn
Misc. Non-Iron Waste	Misc. Fruit	Misc. Knitted Outerwear	Bovine meat	Woven Fabric $<$ 85% Discontinuous Synthetic Fibres

F5: Electronics I	F6: Processed Materials	F7: Electronics II	F8: Materials I	F9: Machinery I
Misc. Electrical Machinery	Baked Goods	Measuring Controlling Instruments	Misc. Articles of Iron	Misc. Rotating Electric Plant Parts
Vehicles Stereos	Metal Containers	Mathematical Calculation Instruments	Carpentry Wood	Control Instruments of Gas or Liquid
Misc. Data Processing Equipment	Misc. Edibles	Misc. Electrical Instruments	Misc. Manufactured Wood Articles	Valves

F10: Materials II	F11: Automobile	F12: Chemicals I	F13: Chemicals II	F14: Machinery II	F15: Miscellaneous
Improved Wood	Vehicles Parts - Accessories	Synthetic Rubber	Aldehyde, Ketone	Parts of Metalworking Machine Tools	Misc. Pumps
Mineral Wool	Cars	Acrylic Polymers	Glycosides, Vaccines	Interchangeable Tool Parts	Ash and Residues
Central Heating Equipment	Iron Wire	Silicones	Medicaments	Polishing Stones	Chemical Wood Pulp of sulphite

Results
Interpretability

Top Products (decay 30%)	B_{kd}
Bovine	0.49
Miscellaneous Refrigeration Equipment	0.43
Radioactive Chemicals	0.41
Blocks of Iron and Steel	0.41
Rape Seeds	0.40
Animal meat, misc	0.39
Refined Sugars	0.38
Miscellaneous Tire Parts	0.38
Leather Accessories	0.38
Liquor	0.38
Bovine meat	0.38
Embroidery	0.37
Unmilled Barley	0.37
Dried Vegetables	0.36
Textile Fabrics Clothing Accessories	0.36
Horse Meat	0.35
Iron Bars and Rods	0.35
Analog Navigation Devices	0.35

(a) SVD

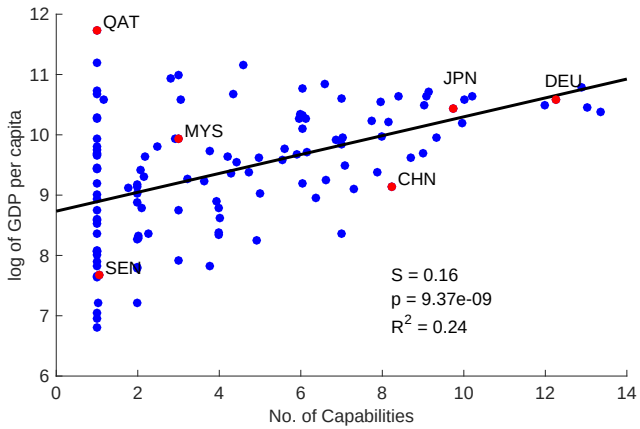
Top Products (decay 30%)	B_{kd}
Miscellaneous Animal Oils	0.78
Bovine and Equine Entrails	0.72
Bovine meat	0.68
Preserved Milk	0.63
Equine	0.62
Butter	0.58
Misc. Animal Origin Materials	0.57
Glues	0.56

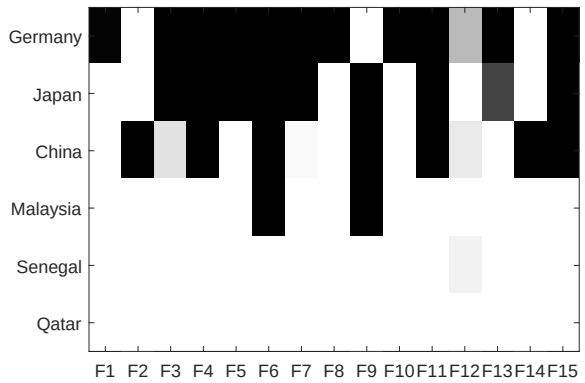
(b) S3R-IBP

Results
Interpretability

Countries in Capability Space

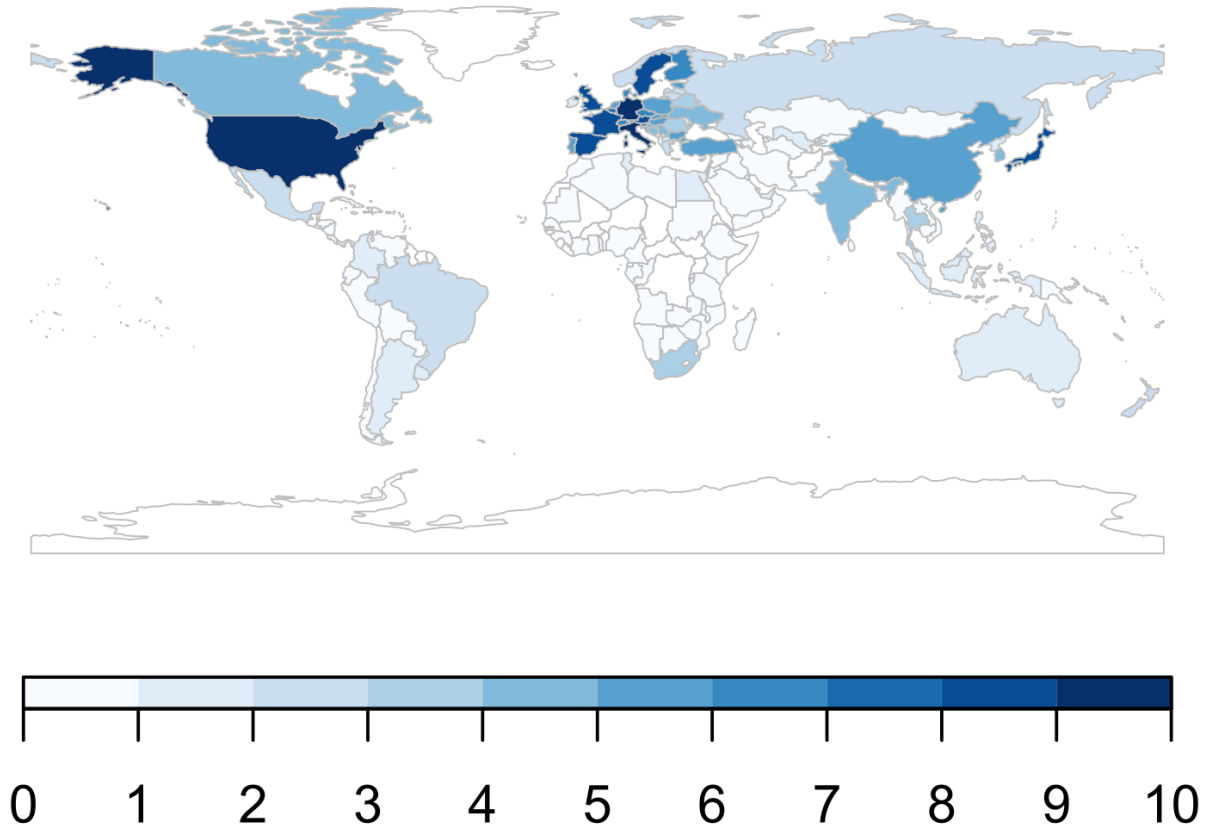
- France = Belgium + Industrial Machinery
- Germany - Chemical = Austria
- Malaysia (Electronics) + Clothing → Phillipines
- Phillipines + Basic Processing → Indonesia, Vietnam





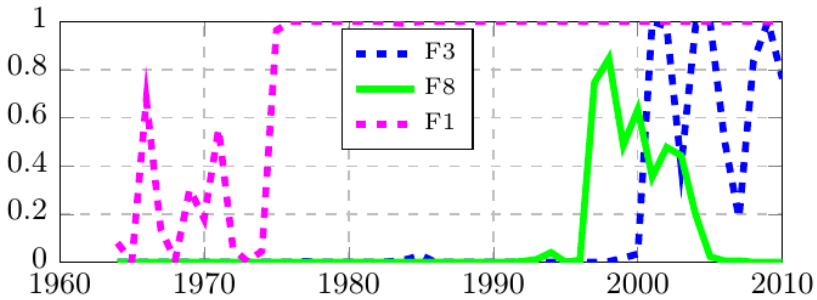
Deep S3R-IBP: using a 2nd layer

1. "Simple" and "advanced" capabilities
2. Countries divided in two big groups: "quiescence" trap.

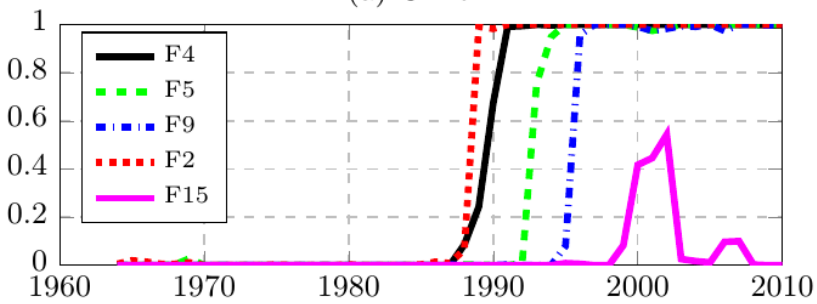


Temporal Dynamics

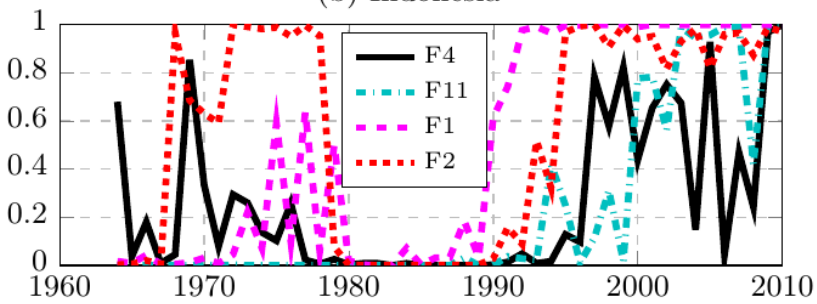
	Capabilities
F0	Bias
F1	Agriculture
F2	Clothing I
F3	Farming
F4	Clothing II
F5	Electronics I
F6	Processed Materials
F7	Electronics II
F8	Materials I
F9	Machinery I
F10	Materials II
F11	Automobile
F12	Chemicals I
F13	Chemicals II
F14	Machinery II
F15	Miscellaneous



(a) Chile



(b) Indonesia



(c) Egypt

Conclusion

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

Future works

- **Time-dependent extension** with Markovian activation of features and smooth variation of capabilities.

Thank you for listening! Any question?

melanie@tsc.uc3m.es

Sources and References


C. Bishop: *Pattern Recognition and Machine Learning,* 2006.


K. P. Murphy: *Machine Learning: a Probabilistic Perspective*, 2012.


D. J.C. MacKay: *Information Theory, Inference, and Learning Algorithms*, 2003.


Z. Ghahramani & C. E. Rasmussen, Slides for *Machine Learning Course* at Cambridge University.


S. J. Gershman, D.M. Blei: A tutorial on Bayesian nonparametric models, 2012.


Y.W. Teh: Slides for *Probabilistic and Bayesian Machine Learning*, UC3M, 2010.


M. N. Schmidt & M. Morup: *Advanced Topics in Machine Learning*, MLSS, DTU, 2013.


D. B. Dunson: *Nonparametric Bayes Applications to Biostatistics*, 2010.