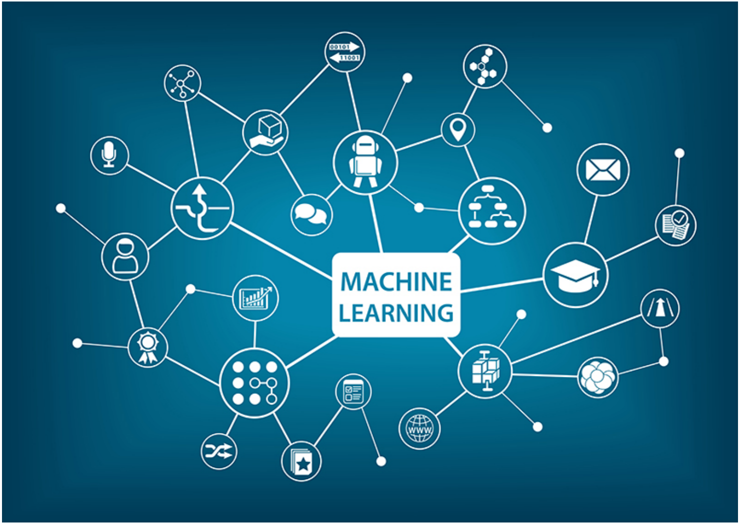
Bayesian Nonparametric Models for Data Exploration

Melanie F. Pradier

Friday 15th September, 2017

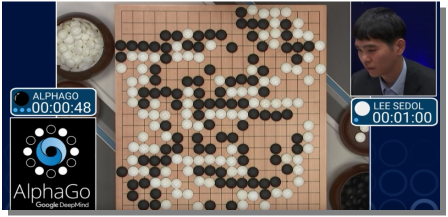
Outline

Motivation

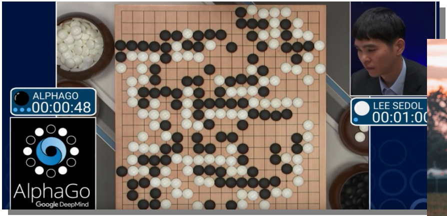


Motivation

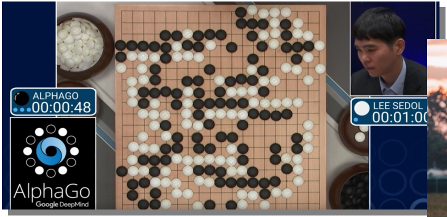
Motivation

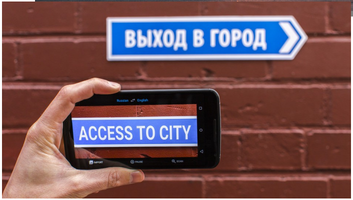


Motivation



Motivation

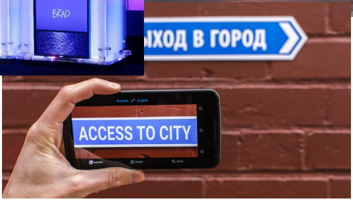

ALPHAGO
00:00:48
LEE SEDOL
AlphaGo
Google DeepMind


ВЫХОД В ГОРОД
ACCESS TO CITY

Motivation


ALPHAGO
00:00:48
LEE SEDOL
00:01:0
AlphaGo
Google DeepMind

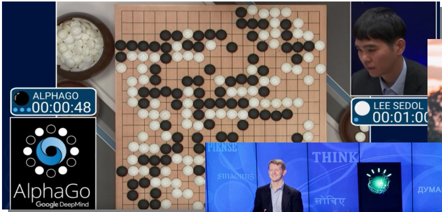

THINK
ДУМАЙ
सोचिए
$0
$1,200
$0
KEN
WATSON
BRAD

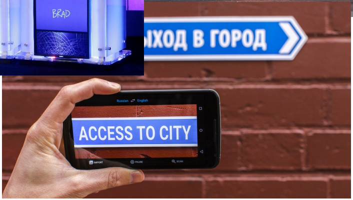

ACCESS TO CITY

Motivation

Data Exploitation Age

Motivation





Data Exploitation Age

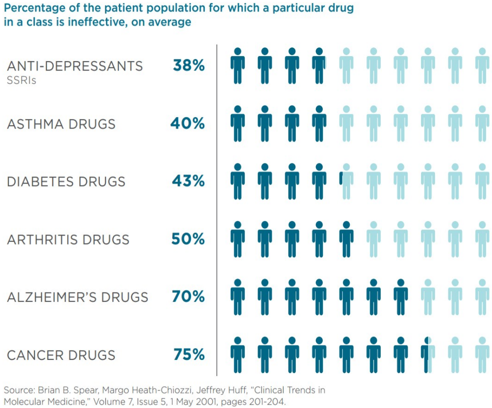
· · · but are we making the outmost out of data?

Motivation

An example: personalized medicine

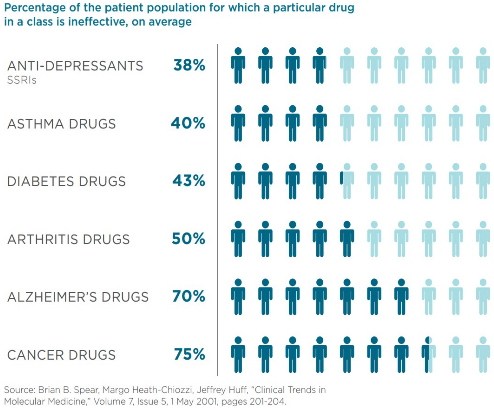
Motivation

An example: personalized medicine



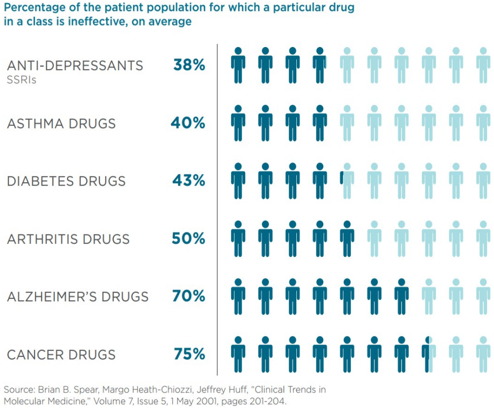
Motivation

An example: personalized medicine



Motivation

An example: personalized medicine

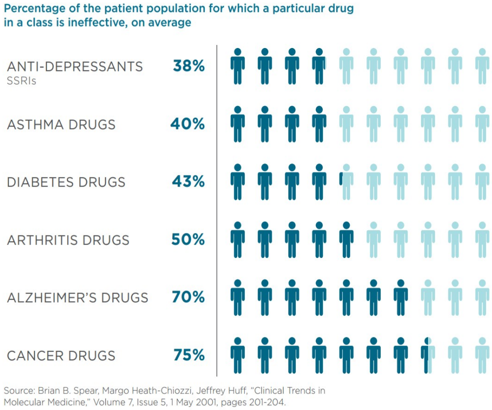


Challenges

- Complexity

Motivation

An example: personalized medicine

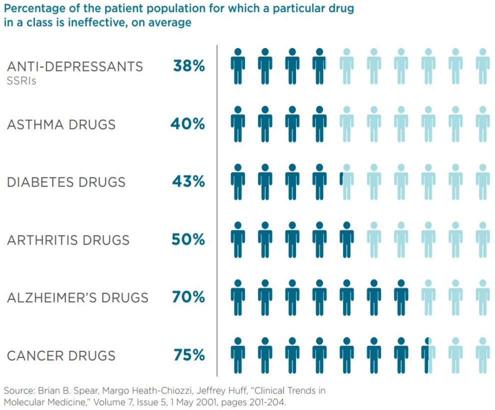


Challenges

- Complexity
- Missing data

Motivation

An example: personalized medicine

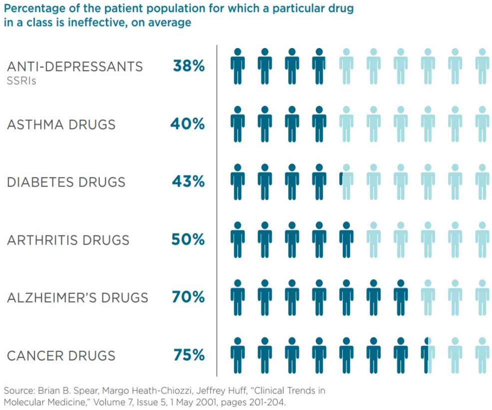


Challenges

- Complexity
- Missing data
- *Small data within big data*
- ...

Motivation

An example: personalized medicine

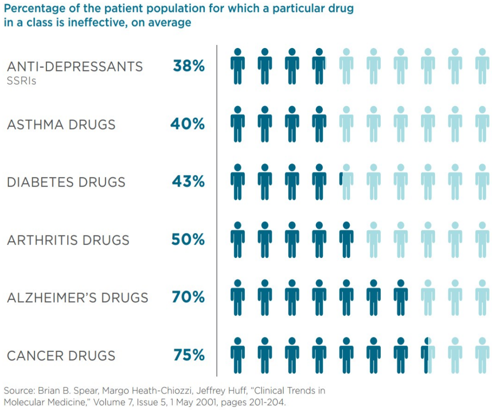


Challenges

- Complexity
- Missing data
- *Small data within big data*
- ...
- Research focus
  $\rightarrow$ data exploration

Motivation

An example: personalized medicine



Challenges

- Complexity
- Missing data
- *Small data within big data*
- ...
- Research focus
  → data exploration

2018 EU General Data Protection Regulation

"right to explanation"

(Goodman et.al. 2016)

Motivation

Focus: data exploration

In this thesis ...

1. How does aging impact our athletic performance? (Ch. 3)
2. What are the underlying mechanisms of cancer? (Ch. 4 & 5)
3. Which factors make countries wealthier than others? (Ch. 6)

Motivation

Focus: data exploration

In this thesis ...

1. How does aging impact our athletic performance? (Ch. 3)
2. What are the underlying mechanisms of cancer? (Ch. 4 & 5)
3. Which factors make countries wealthier than others? (Ch. 6)

Main goal

- Knowledge discovery
- Hypothesis generation

Motivation

Focus: data exploration

In this thesis . . .

1. How does aging impact our athletic performance? (Ch. 3)
2. What are the underlying mechanisms of cancer? (Ch. 4 & 5)
3. Which factors make countries wealthier than others? (Ch. 6)

Main goal

- Knowledge discovery
- Hypothesis generation

Our Approach

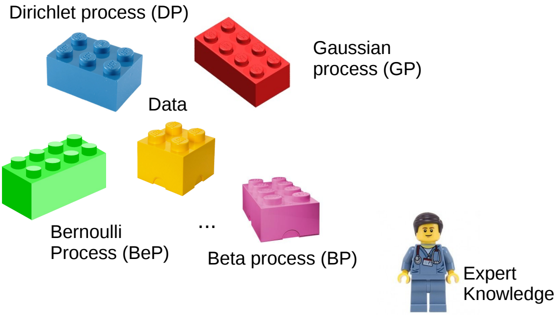
Bayesian nonparametrics

Why Bayesian nonparametrics?

- Bayesian: to handle uncertainty
- Nonparametric: to adapt model complexity depending on input data (hypothesis generation)

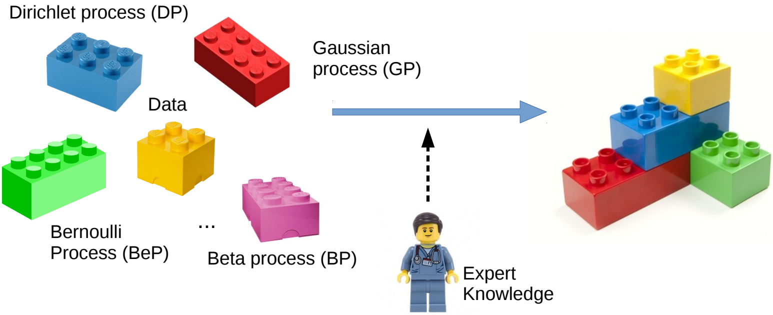
Why Bayesian nonparametrics?

- Bayesian: to handle uncertainty
- Nonparametric: to adapt model complexity depending on input data (hypothesis generation)



Why Bayesian nonparametrics?

- Bayesian: to handle uncertainty
- Nonparametric: to adapt model complexity depending on input data (hypothesis generation)



Contributions

Goal: build useful BNP models for specific data exploration tasks.

Contributions

Goal: build useful BNP models for specific data exploration tasks.

- Novel applications
- Make things work with real data (modeling, inference, validation)
- Interpretability, sharing across observations, replicability
- Open-source software and databases

Contributions

Goal: build useful BNP models for specific data exploration tasks.

Atom-dependent DP mixture model

- estimates density in stratified data
- suitable for fairness requirements
- link to mixture of experts

Contributions

Goal: build useful BNP models for specific data exploration tasks.

Atom-dependent DP mixture model

- estimates density in stratified data
- suitable for fairness requirements
- link to mixture of experts

Case-control IBP feature model

- infers latent features in heterogeneous structured data
- suitable to separate global and group-specific effects
- combined with statistical testing

Contributions

Goal: build useful BNP models for specific data exploration tasks.

Atom-dependent DP mixture model

- estimates density in stratified data
- suitable for fairness requirements
- link to mixture of experts

Case-control IBP feature model

- infers latent features in heterogeneous structured data
- suitable to separate global and group-specific effects
- combined with statistical testing

Poisson factor analysis (PFA) models

→ flexible feature models for count data

1. Hierarchical PFA:
   - deals with stratified data
2. Three-parameter Restricted PFA:
   - imposes structured sparsity in latent space
3. Dynamic PFA:
   - allows for time-varying activation of latent factors

Contributions

Goal: build useful BNP models for specific data exploration tasks.

Atom-dependent DP mixture model

- estimates density in stratified data
- suitable for fairness requirements
- link to mixture of experts
- Application: marathon

Case-control IBP feature model

- infers latent features in heterogeneous structured data
- suitable to separate global and group-specific effects
- combined with statistical testing
- Application: clinical trial

Poisson factor analysis (PFA) models

$\rightarrow$ flexible feature models for count data

1. Hierarchical PFA:
   - deals with stratified data
2. Three-parameter Restricted PFA:
   - imposes structured sparsity in latent space
3. Dynamic PFA:
   - allows for time-varying activation of latent factors

- Application: international trade

Outline

1. Introduction
2. Bayesian nonparametrics
3. ADDP mixture model for marathon model
4. C-IBP feature model for clinical trials
5. PFA models for international trade
6. Conclusions

Bayesian nonparametrics (BNPs)

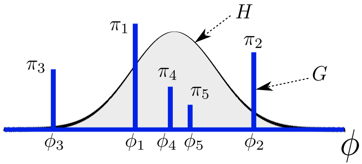
- Bayesian framework for **model selection**
- Nonparametric: number of parameters grows with the amount of data:
  - Prior over **infinite-dimensional** parameter space
  - Only a **finite subset** of parameters is used for any finite dataset

Bayesian nonparametrics (BNPs)

- Bayesian framework for **model selection**
- Nonparametric: number of parameters grows with the amount of data:
  - Prior over **infinite-dimensional** parameter space
  - Only a **finite subset** of parameters is used for any finite dataset
- Rely on stochastic processes:
  - Dirichlet process
  - Beta process
  - Gaussian process
  - ...

Dirichlet process (DP)

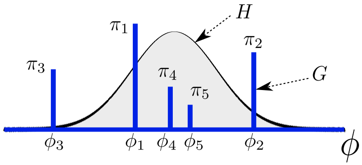
$$G \sim \mathrm{DP}(\alpha, H)$$



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

Dirichlet process (DP)

$$G \sim \mathrm{DP}(\alpha, H)$$

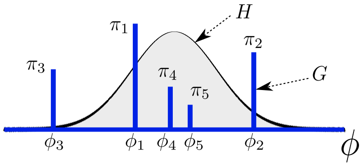


$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

- central block for infinite mixture models

Dirichlet process (DP)

$$G \sim \mathrm{DP}(\alpha, H)$$



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

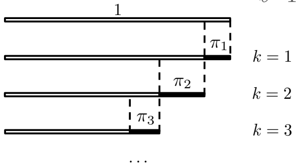
- central block for infinite mixture models

Stick-breaking representation

(Ishwaran et.al, 2001)

For $k = 1, \cdots, \infty$

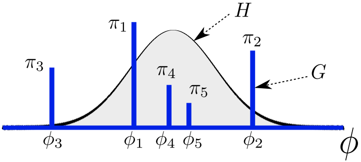
$$v_k \sim \mathrm{Beta}(\alpha, 1),\ \pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell)$$



$$\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$$

Dirichlet process (DP)

$$G \sim \mathrm{DP}(\alpha, H)$$



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k}$$

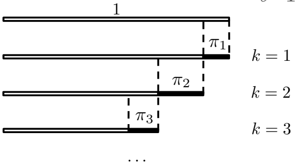
- central block for infinite mixture models

Stick-breaking representation

(Ishwaran et.al, 2001)

For $k = 1, \cdots, \infty$

$$v_k \sim \mathrm{Beta}(\alpha, 1),\ \pi_k = v_k \prod_{\ell=1}^{k-1} (1 - v_\ell)$$



$$\boldsymbol{\pi} \sim \mathrm{GEM}(\alpha)$$

For $k = 1, \cdots, \infty$

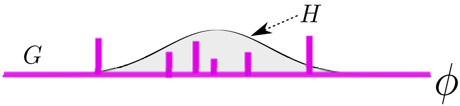
$$\phi_k \sim H$$

Indian buffet process (IBP)

- central block for infinite latent feature models

Indian buffet process (IBP)

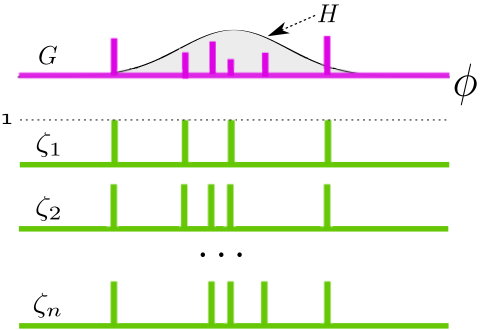
- central block for infinite latent feature models
- hierarchy of a Beta process (BP) with multiple Bernoulli processes (BeP)



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{BP}(c, \alpha, H)$$

Indian buffet process (IBP)

- central block for infinite latent feature models
- hierarchy of a Beta process (BP) with multiple Bernoulli processes (BeP)



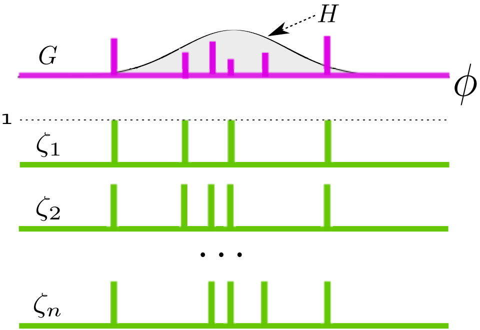
$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{BP}(c, \alpha, H)$$

For $n = 1, \cdots, \infty$

$$\zeta_n = \sum_{k=1}^{\infty} z_{nk} \delta_{\phi_k} \sim \mathrm{BeP}(G)$$

Indian buffet process (IBP)

- central block for infinite latent feature models
- hierarchy of a Beta process (BP) with multiple Bernoulli processes (BeP)



$$G = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_k} \sim \mathrm{BP}(c, \alpha, H)$$

For $n = 1, \cdots, \infty$

$$\zeta_n = \sum_{k=1}^{\infty} z_{nk} \delta_{\phi_k} \sim \mathrm{BeP}(G)$$

$$\mathbf{Z} \sim \mathrm{IBP}(\alpha)$$

Outline

Motivation

1. What is the impact of age and gender on runners performance?
2. Can we compare different runners in a fair manner?
   - entry requirements
   - rewards

Motivation

1. What is the impact of age and gender on runners performance?
2. Can we compare different runners in a fair manner?
   - entry requirements
   - rewards

Our Approach

- dependent density estimation model
  - delivers scientific knowledge in sport sciences
  - constitutes a fair age-gender grading system
  - relies on dependent Dirichlet process

Dependent Dirichlet process (DDP)

(MacEachern,2000)

J: number of groups

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_{jk}}$$

Dependent Dirichlet process (DDP)

(MacEachern,2000)

J: number of groups

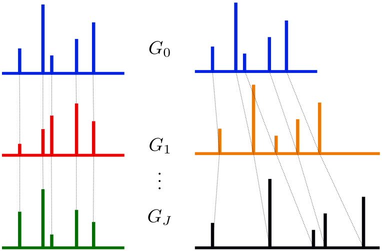
$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_{jk}}$$

- hierarchical DP (Teh et.al, 2005)

$$G_j = \sum_{k=1}^{\infty} \pi_{jk} \delta_{\phi_k}$$

- single-p DDP (MacEachern, 2000)

$$G_j = \sum_{k=1}^{\infty} \pi_k \delta_{\phi_{jk}}$$



hierarchical DP

$$G_0 \sim \mathrm{DP}(\alpha, \mathrm{H})$$
$$G_j \sim \mathrm{DP}(\gamma, \mathrm{G}_0)$$

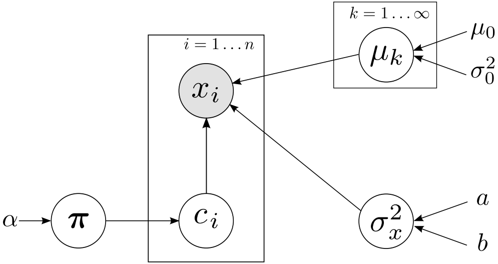
single-p DDP

$$G_0 \sim \mathrm{DP}(\alpha, \mathrm{H})$$
$$G_j = \mathrm{T}_j[G_0]$$

Atom-dependent DP mixture model

Generative model

$x_i \equiv$ marathon finishing time for runner i



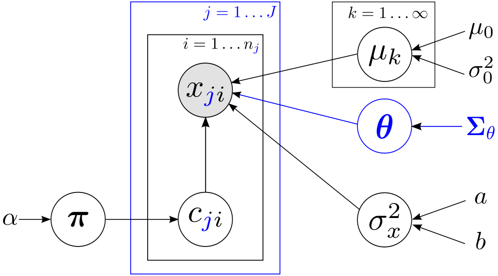
$$\boldsymbol{\pi}|\alpha \sim \mathrm{GEM}(\alpha)$$
$$c_i|\boldsymbol{\pi} \sim \mathrm{Cat}(\boldsymbol{\pi})$$
$$\mu_k \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$\sigma_x^2 \sim \mathcal{IG}\left(a, b\right)$$

$$x_i|\text{other vars} \sim \mathcal{N}\left(x_i|\mu_{c_i}, \sigma_x^2\right)$$

Atom-dependent DP mixture model

Generative model

$x_{ji} \equiv$ marathon finishing time for runner i in age group j



$$\boldsymbol{\pi}|\alpha \sim \mathrm{GEM}(\alpha)$$
$$c_{ji}|\boldsymbol{\pi} \sim \mathrm{Cat}(\boldsymbol{\pi})$$
$$\mu_k \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$
$$\sigma_x^2 \sim \mathcal{IG}\left(a, b\right)$$
$$\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right)$$

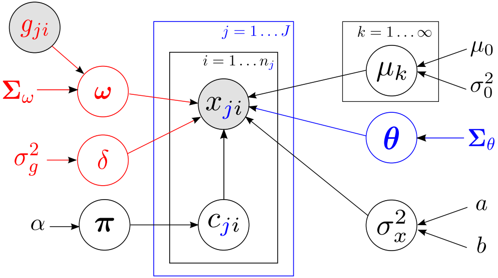
$$x_{ji}|\text{other vars} \sim \mathcal{N}\left(x_{ji}|\mu_{c_{ji}} + \theta_j, \sigma_x^2\right)$$

$$(\boldsymbol{\Sigma}_\theta)_{\ell q} = \sigma_\theta^2 \exp\left(-\frac{(\ell - q)^2}{2\nu^2}\right) + \kappa\delta(\ell - q)$$

Atom-dependent DP mixture model

Generative model

$x_{ji} \equiv$ marathon finishing time for runner i in age group j $\quad$ $g_{ji} \equiv$ gender



$$\boldsymbol{\pi}|\alpha \sim \mathrm{GEM}(\alpha)$$

$$c_{ji}|\boldsymbol{\pi} \sim \mathrm{Cat}(\boldsymbol{\pi})$$

$$\mu_k \sim \mathcal{N}\left(\mu_0, \sigma_0^2\right)$$

$$\sigma_x^2 \sim \mathcal{IG}\left(a, b\right)$$

$$\boldsymbol{\theta} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\theta\right)$$

$$\delta \sim \mathcal{N}\left(\mathbf{0}, \sigma_\omega^2\right)$$

$$\boldsymbol{\omega} \sim \mathcal{N}\left(\mathbf{0}, \boldsymbol{\Sigma}_\omega\right)$$

$$x_{ji}|\text{other vars} \sim \mathcal{N}\left(x_{ji}|\mu_{c_{ji}} + \theta_j + \mathbb{1}[g_{ji} = 1](\delta + \omega_j), \sigma_x^2\right)$$

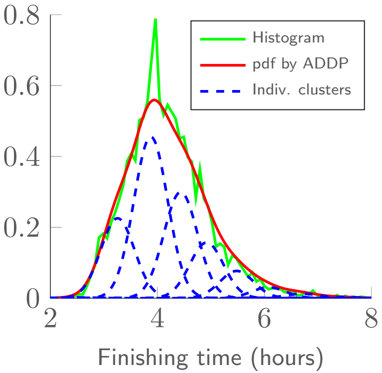
$$(\boldsymbol{\Sigma}_\theta)_{\ell q} = \sigma_\theta^2 \exp\left(-\frac{(\ell - q)^2}{2\nu^2}\right) + \kappa\delta(\ell - q)$$

Results

Impact of age

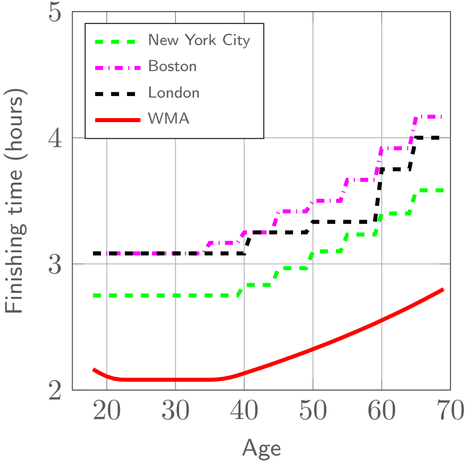
- MCMC approach
- conditional conjugacy
- block Gibbs sampler

- $1/4$ M runners

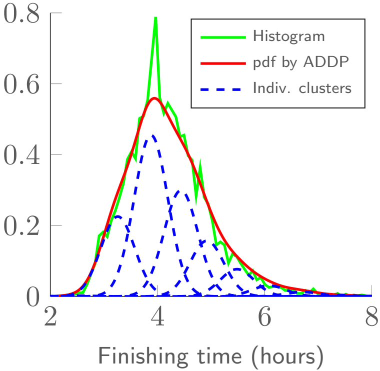


Results

Impact of age

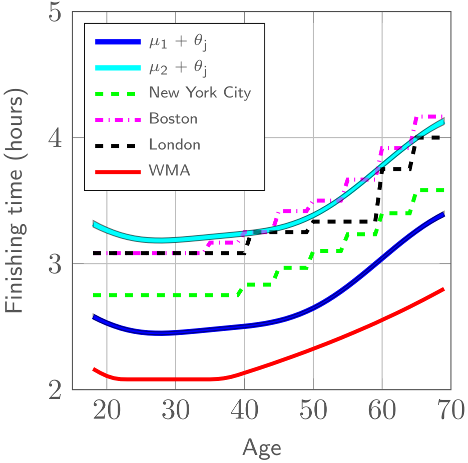


- MCMC approach
- conditional conjugacy
- block Gibbs sampler
- $1/4$ M runners

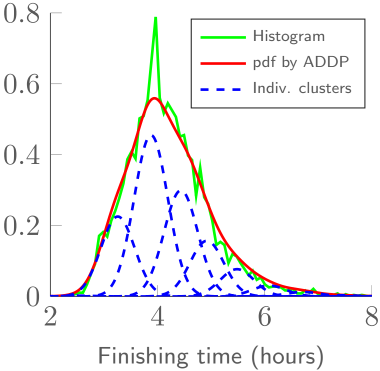


Results

Impact of age

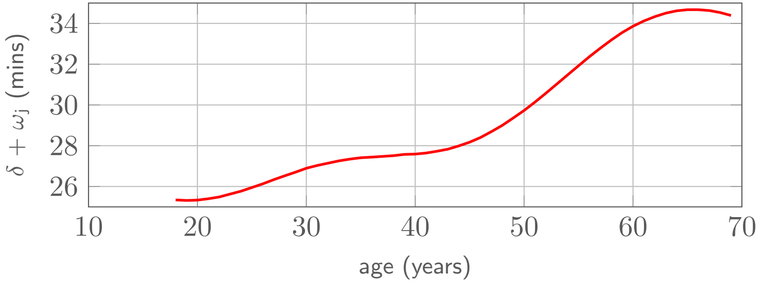


- MCMC approach
- conditional conjugacy
- block Gibbs sampler
- 1/4 M runners



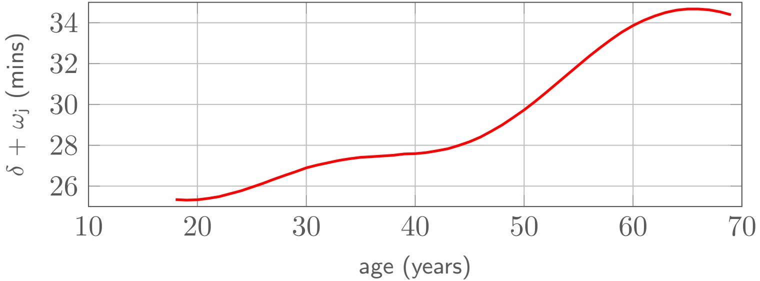
Results

Impact of gender



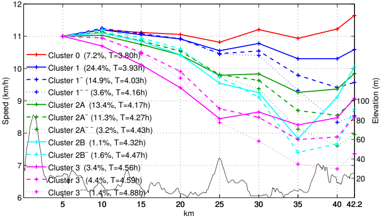
Results

Impact of gender



Other Results

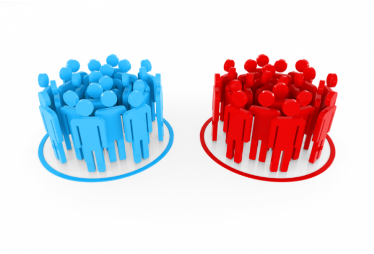
- **Speed-dependent** cluster means
- Link to **mixture of experts**
- Analysis of **running patterns**
- **Prediction** of finishing time



Outline

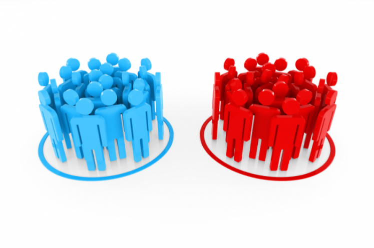
Motivation: biomarker discovery in clinical trials

Def: "any variable that can be used as an indicator of a particular disease state".



Motivation: biomarker discovery in clinical trials

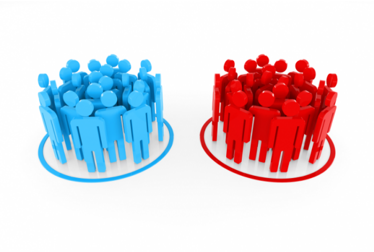
Def: "any variable that can be used as an indicator of a particular disease state".



We want to discover:

Motivation: biomarker discovery in clinical trials

Def: "any variable that can be used as an indicator of a particular disease state".

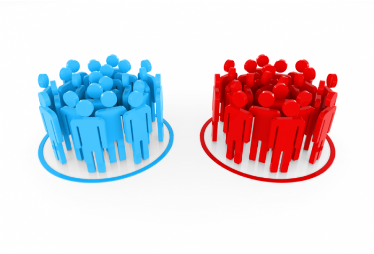


We want to discover:

1. Indicators of disease progression: prognostic biomarkers

Motivation: biomarker discovery in clinical trials

Def: "any variable that can be used as an indicator of a particular disease state".

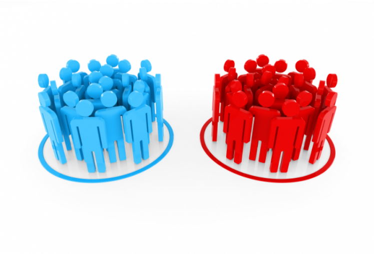


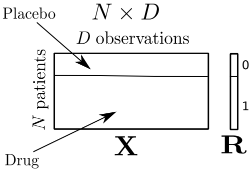
We want to discover:

1. Indicators of disease progression: prognostic biomarkers
2. Indicators of (positive) drug response: predictive biomarkers

Motivation: biomarker discovery in clinical trials

Def: "any variable that can be used as an indicator of a particular disease state".



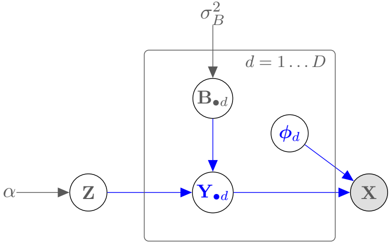


We want to discover:

1. Indicators of disease progression: prognostic biomarkers
2. Indicators of (positive) drug response: predictive biomarkers

General latent feature model (GLFM)
(Valera et.al, 2017)

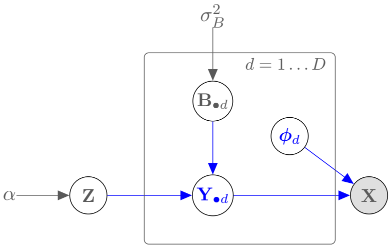
Latent feature model for heterogeneous datasets



General latent feature model (GLFM)

(Valera et.al, 2017)

Latent feature model for heterogeneous datasets



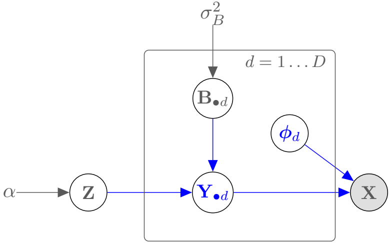
- Link functions T_d depend on type of data for each dimension d

$$
\begin{aligned}
x_{nd} &= T_d(y_{nd}; \boldsymbol{\phi_d}) \\
y_{nd}|\mathbf{Z}, \mathbf{B} &\sim \mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}, \sigma_y^2) \\
B_{kd} &\sim \mathcal{N}(0, \sigma_B^2) \\
\mathbf{Z} &\sim \mathrm{IBP}(\alpha)
\end{aligned}
$$

General latent feature model (GLFM)

(Valera et.al, 2017)

Latent feature model for heterogeneous datasets



- Link functions T_d depend on type of data for each dimension d

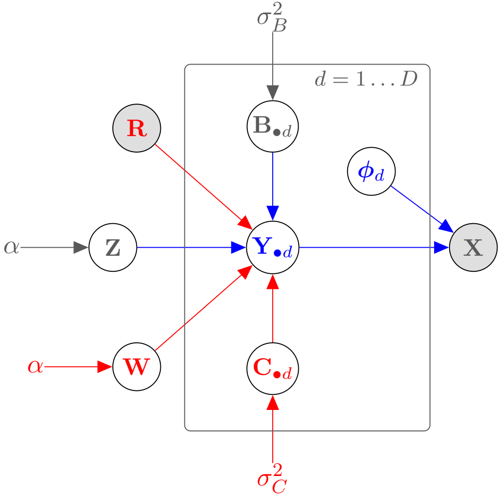
$$
\begin{aligned}
x_{nd} &= T_d(y_{nd}; \boldsymbol{\phi_d}) \\
y_{nd}|\mathbf{Z}, \mathbf{B} &\sim \mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}, \sigma_y^2) \\
B_{kd} &\sim \mathcal{N}(0, \sigma_B^2) \\
\mathbf{Z} &\sim \mathrm{IBP}(\alpha)
\end{aligned}
$$

Our contribution to GLFM project

- Open-source python code
- Simulations for data exploration

https://github.com/ivaleraM/GLFM

Our contribution: Case-control IBP (C-IBP)



R_n: drug indicator por patient n

$$x_{nd} = T_d(y_{nd}; \boldsymbol{\phi_d})$$
$$y_{nd}|\mathbf{Z}, \mathbf{W}, \mathbf{B}, \mathbf{C}, \mathbf{R} \sim$$
$$\mathcal{N}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d} + \mathbb{1}[\mathbf{R}_n = 1]\mathbf{W}_{n\bullet}\mathbf{C}_{\bullet d}, \sigma_y^2)$$
$$B_{kd} \sim \mathcal{N}(0, \sigma_B^2)$$
$$\mathbf{Z} \sim \text{IBP}(\alpha)$$
$$C_{kd} \sim \mathcal{N}(0, \sigma_C^2)$$
$$\mathbf{W} \sim \text{IBP}(\alpha)$$

- **Inference**: MCMC approach with accelerated Gibbs sampling
- **Biomarker discovery**: statistical multiple hypothesis testing

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

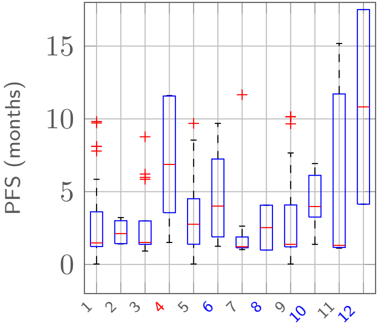
Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01

Results: subpopulations

GPC3 Antibody Treatment against Liver Cancer (J. Hepatology. 2016 Apr, Abou-Alfa et.al.)

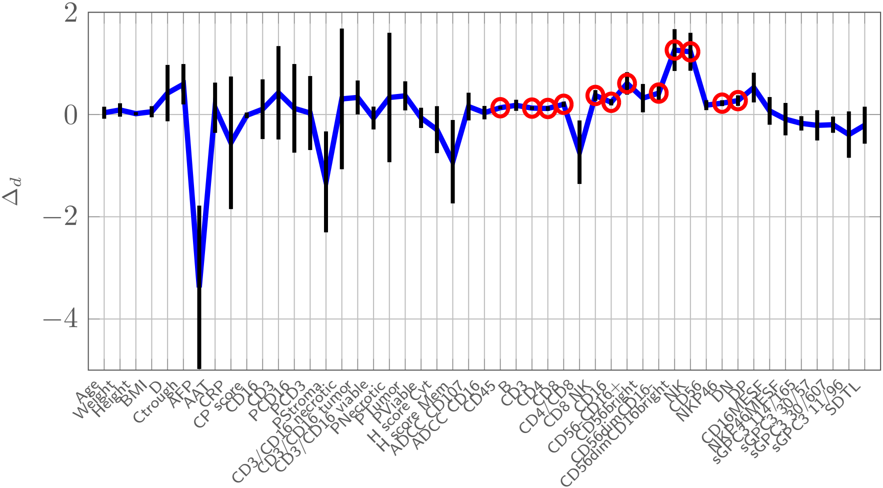
- 180 patients: 60 took a placebo, 120 took the drug
- PFS: Progression Free Survival

Sub-population	Drug Identifier	F1	F2	F3	Size (number of patients)	Mean PFS (months)	Median PFS (months)
1.	0	0	0	0	33.37	3.06	1.65
2.	0	0	1	0	4.07	2.29	2.24
3.	0	1	0	0	17.84	2.72	1.81
4.	0	1	1	0	4.72	7.05	7.18
5.	1	0	0	0	51.52	3.22	2.55
6.	1	0	0	1	16.77	4.17	3.65
7.	1	0	1	0	8.38	1.74	1.33
8.	1	0	1	1	2.07	2.69	2.65
9.	1	1	0	0	29.88	3.36	2.03
10.	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1.94	10.04	10.01



Results: biomarker discovery

Treatment-specific feature F3



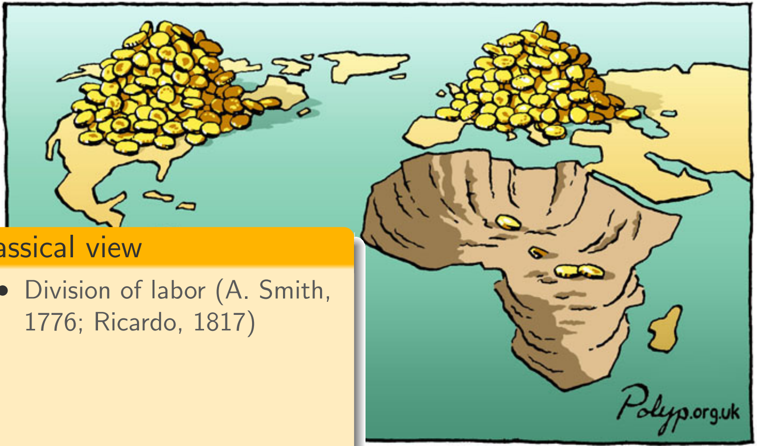
Outline

Motivation: wealth of nations

What makes some countries wealthier than others?

Motivation: wealth of nations

What makes some countries wealthier than others?



Classical view

- Division of labor (A. Smith, 1776; Ricardo, 1817)

Motivation: wealth of nations

What makes some countries wealthier than others?

Classical view

- Division of labor (A. Smith, 1776; Ricardo, 1817)
- Specialization leads to economic efficiency

Motivation: wealth of nations

What makes some countries wealthier than others?

Classical view

- Division of labor (A. Smith, 1776; Ricardo, 1817)
- Specialization leads to economic efficiency
- Export portfolios

Motivation: wealth of nations

What makes some countries wealthier than others?

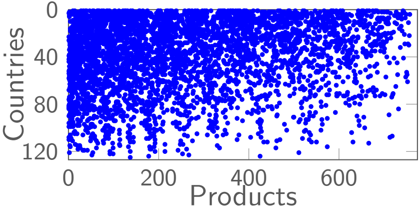
Classical view

- Division of labor (A. Smith, 1776; Ricardo, 1817)
- Specialization leads to economic efficiency
- Export portfolios

$\rightarrow$ block-structure

Motivation: wealth of nations

The reality:

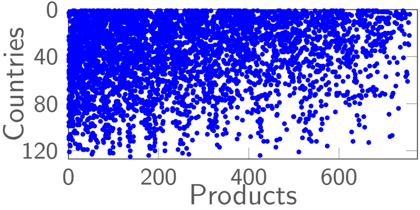


$$\mathrm{RCA}_{nd} = \frac{E_{nd} / \sum_p E_{nd}}{\sum_n E_{nd} / \sum_{n,d} E_{nd}}$$

$$x_{nd} = \begin{cases} 1, & \text{if } \mathrm{RCA}_{nd} \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

Motivation: wealth of nations

The reality:



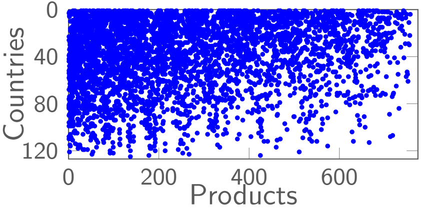
$$\mathrm{RCA}_{nd} = \frac{E_{nd}/\sum_p E_{nd}}{\sum_n E_{nd}/\sum_{n,d} E_{nd}}$$

$$x_{nd} = \begin{cases} 1, & \text{if } \mathrm{RCA}_{nd} \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

Properties:

Motivation: wealth of nations

The reality:



$$\mathrm{RCA}_{nd} = \frac{E_{nd}/\sum_p E_{nd}}{\sum_n E_{nd}/\sum_{n,d} E_{nd}}$$

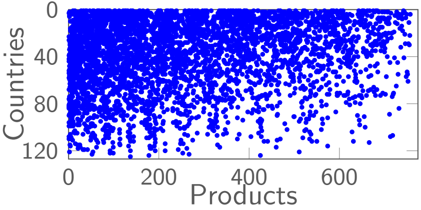
$$x_{nd} = \begin{cases} 1, & \text{if } \mathrm{RCA}_{nd} \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

Properties:

1. Triangularity

Motivation: wealth of nations

The reality:



$$\mathrm{RCA}_{nd} = \frac{E_{nd}/\sum_p E_{nd}}{\sum_n E_{nd}/\sum_{n,d} E_{nd}}$$

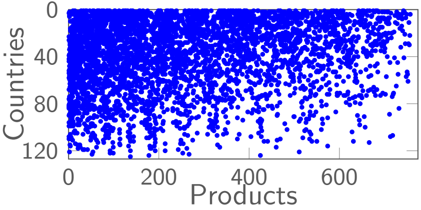
$$x_{nd} = \begin{cases} 1, & \text{if } \mathrm{RCA}_{nd} \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

Properties:

1. Triangularity
2. $D \gg N$

Motivation: wealth of nations

The reality:



$$\mathrm{RCA}_{nd} = \frac{E_{nd}/\sum_p E_{nd}}{\sum_n E_{nd}/\sum_{n,d} E_{nd}}$$

$$x_{nd} = \begin{cases} 1, & \text{if } \mathrm{RCA}_{nd} \geq 1 \\ 0, & \text{otherwise} \end{cases}$$

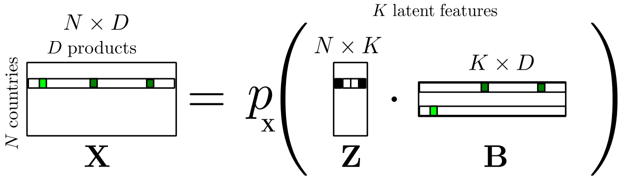
Properties:

1. Triangularity
2. $D \gg N$

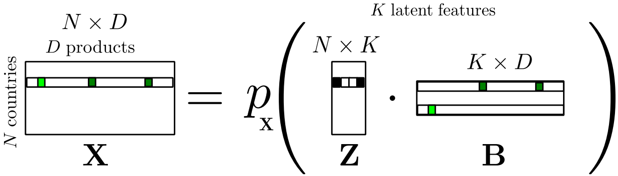
Our Approach

1. Develop an infinite Poisson factor analysis model ...
   - flexible prior
   - feature sparsity
2. Design a time-varying extension

Bernoulli process Poisson factor analysis (BeP-PFA)



Bernoulli process Poisson factor analysis (BeP-PFA)

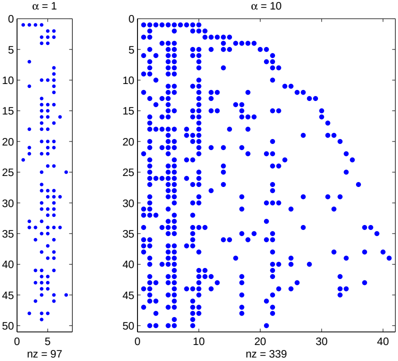


Generative Model

$$
\begin{aligned}
x_{nd} &\sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\right) \\
B_{kd} &\sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right) \\
\mathbf{Z} &\sim \text{IBP}(\alpha)
\end{aligned}
$$

Limitation of the IBP

- Number of ones per row $J_n \propto \text{Poisson}(\alpha)$
- Number of non-empty features $K \propto \text{Poisson}(\alpha \sum_{j=1}^{N} \frac{1}{j})$
- 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

$$\mathbf{Z}_{n\bullet} \sim \mathrm{BeP}(\mu) \quad (5.1)$$

$$\mu \sim \mathrm{SBP}(1, \alpha, H, c, \sigma) \quad (5.2)$$

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

Beyond the standard IBP

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

- More flexible distribution for feature weights

$$\mathbf{Z}_{n\bullet} \sim \mathrm{BeP}(\mu) \tag{5.1}$$
$$\mu \sim \mathrm{SBP}(1, \alpha, H, c, \sigma) \tag{5.2}$$

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

Beyond the standard IBP

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

- More flexible distribution for feature weights

$$\mathbf{Z}_{n\bullet} \sim \mathrm{BeP}(\mu) \tag{5.1}$$
$$\mu \sim \mathrm{SBP}(1, \alpha, H, c, \sigma) \tag{5.2}$$

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

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

- Arbitrary prior f over J_n

$$\mathbf{Z}_{n\bullet} \sim \text{R-BeP}(\mu, f) \tag{5.3}$$
$$\mu \sim \mathrm{BP}(1, \alpha, H) \tag{5.4}$$

Beyond the standard IBP

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

- More flexible distribution for feature weights

$$\mathbf{Z}_{n\bullet} \sim \mathrm{BeP}(\mu) \quad (5.1)$$
$$\mu \sim \mathrm{SBP}(1, \alpha, H, c, \sigma) \quad (5.2)$$

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

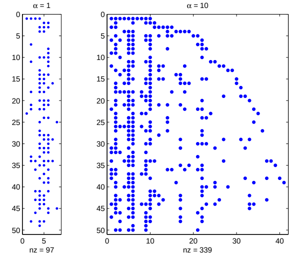
Restricted IBP (Doshi-Velez et.al, 2015)

- Arbitrary prior f over J_n

$$\mathbf{Z}_{n\bullet} \sim \mathrm{R\text{-}BeP}(\mu, f) \quad (5.3)$$
$$\mu \sim \mathrm{BP}(1, \alpha, H) \quad (5.4)$$

- Combination of both
- Flexible prior

Our contributions

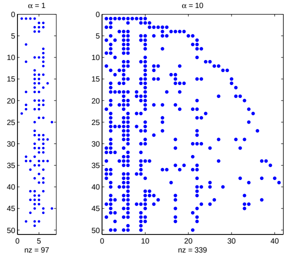


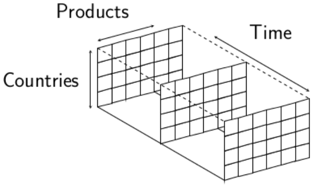
3RBeP-PFA for static scenario

$$
\begin{aligned}
x_{nd} &\sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\right) \\
B_{kd} &\sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right) \\
\mathbf{Z} &\sim \text{3R-IBP}(\alpha, c, \sigma, f)
\end{aligned}
$$

- **Inference**: aux. vars + dynamic programming (Doshi-Velez et.al, 2015)

Our contributions





3RBeP-PFA for static scenario

$$x_{nd} \sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d}\right)$$
$$B_{kd} \sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right)$$
$$\mathbf{Z} \sim \text{3R-IBP}(\alpha, c, \sigma, f)$$

- **Inference**: aux. vars + dynamic programming (Doshi-Velez et.al, 2015)

dBeP-PFA for dynamic scenario

$$x_{nd}^{(t)} \sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}^{(t)}\mathbf{B}_{\bullet d}\right)$$
$$B_{kd} \sim \text{Gamma}\left(\alpha_B, \frac{\mu_B}{\alpha_B}\right)$$
$$\mathbf{Z}_{n\bullet}^{(\bullet)} \sim \text{mIBP}(\alpha, \gamma, \delta)$$

- **Inference**: forward-filtering backward-sampling (Gael et.al, 2009)

Results in static scenario

Quantitative analysis: accuracy Vs interpretability

Metric	PMF	NNMF	BeP-PFA	sBeP-PFA	3RBeP-PFA
Log Perplexity	1.68 ± 0.01	1.61 ± 0.01	$\mathbf{1.59 \pm 0.04}$	3.26 ± 0.17	1.62 ± 0.01
Coherence	-264.60 ± 4.74	-263.27 ± 7.45	-149.36 ± 7.56	-178.44 ± 4.50	$\mathbf{-140.51 \pm 2.73}$

(a) 2010 SITC database ($N = 126$, $D = 744$)

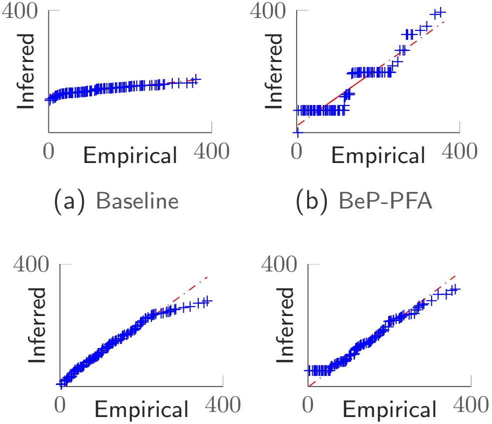
Metric	PMF	NNMF	BeP-PFA	sBeP-PFA	3RBeP-PFA
Log Perplexity	1.48 ± 0.01	$\mathbf{1.47 \pm 0.01}$	1.58 ± 0.01	2.56 ± 0.12	1.57 ± 0.02
Coherence	-264.73 ± 3.11	-264.67 ± 6.22	-148.91 ± 10.57	-168.39 ± 13.16	$\mathbf{-134.51 \pm 4.43}$

(b) 2010 HS database ($N = 123$, $D = 4890$)

- PMF: Probabilistic matrix factorization (Mnih et.al, 2008)
- NNMF: Non-negative matrix factorization (Schmidt et.al, 2009)
- BeP-PFA: Bernoulli process Poisson factor analysis
- sBeP-PFA: sparse Bernoulli process Poisson factor analysis
- 3RBeP-PFA: Three-parameter Restricted Bernoulli process Poisson factor analysis

Results in static scenario

Capturing input sparsity structure



(a) Baseline (b) BeP-PFA

(c) sBeP-PFA (d) 3RBeP-PFA

Results in static scenario

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
Unwrought Lead	Frozen Vegetables	Men's Shirts	Preserved Milk	Woven Fabrics > 85% Discontinuous Synthetic Fiber
Bones, Ivory and Horns	Apples	Blouses	Equine	Yarn < 85% Synthetic Fibers

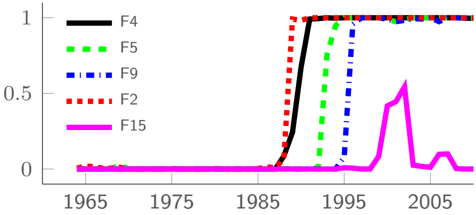
F5: Electronics I	F6: Processed Materials	F7: Electronics II	F8: Materials I	F9: Machinery I
Misc. Electrical Machinery	Baked Goods	Measuring Controlling Instruments	Misc. Articles of Iron	Misc. Rotating Electric Plant Parts
Vehicles Stereos	Metal Containers	Mathematical Calculation Instruments	Carpentry Wood	Control Instruments of Gas or Liquid
Misc. Data Processing Equipment	Misc. Edibles	Misc. Electrical Instruments	Misc. Manufactured Wood Articles	Valves
Video and Sound Recorders	Misc. Articles of Paper	Misc. Heating and Cooling Equipment	Sawn Wood Less Than 5mm Thick	Misc. Rubber
Calculating Machines	Misc. Organic Surfactants	Parts of Office Machines	Electric Current	Misc. Articles of Plastic

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
Aluminium Structures	Trucks - Vans	Misc. Polymer. Products	Inorganic Esters	Tool Holders	Rolls of Paper
Harvesting Machines	Air Pumps - Compressors	Tinned Sheets	Cyclic Alcohols	Misc. Metalworking Machine-Tools	Worked Nickel

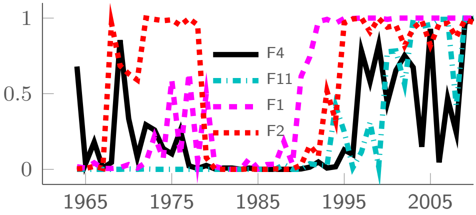
Temporal Dynamics

	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

Indonesia

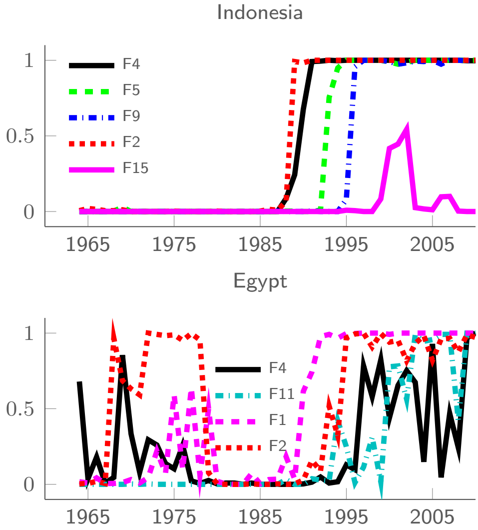


Egypt



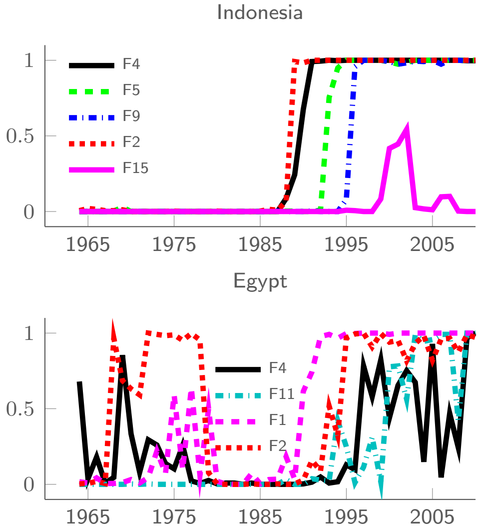
Temporal Dynamics

	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



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



Model extension: Dynamic PFA

Model extension

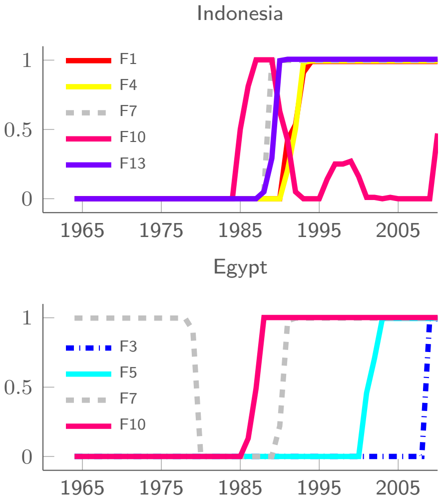
$$x_{nd}^{(t)} \sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}^{(t)} \mathbf{B}_{\bullet d}\right)$$

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

$$\mathbf{Z}_{n\bullet}^{(\bullet)} \sim \text{mIBP}(\alpha, \gamma, \delta)$$

mIBP: markov Indian buffet process

(Gael et.al, 2009)



Outline

Conclusions

BNPs

- useful BNP models for specific data exploration tasks
  - Fair density estimation model
  - Structured general latent feature model (global and group-specific factors)
  - Flexible Poisson factor analysis models in static/dynamic scenarios

Conclusions

BNPs

- useful BNP models for specific data exploration tasks
  - Fair density estimation model
  - Structured general latent feature model (global and group-specific factors)
  - Flexible Poisson factor analysis models in static/dynamic scenarios

Sports science

- age-gender curves
- fair grading system
- running patterns over time

Conclusions

BNPs

- useful BNP models for specific data exploration tasks
  - Fair density estimation model
  - Structured general latent feature model (global and group-specific factors)
  - Flexible Poisson factor analysis models in static/dynamic scenarios

Sports science

- age-gender curves
- fair grading system
- running patterns over time

Cancer research

- subpopulation learning
- biomarker discovery
- clinico-genetic associations

Conclusions

BNPs

- useful BNP models for specific data exploration tasks
  - Fair density estimation model
  - Structured general latent feature model (global and group-specific factors)
  - Flexible Poisson factor analysis models in static/dynamic scenarios

Sports science

- age-gender curves
- fair grading system
- running patterns over time

Cancer research

- subpopulation learning
- biomarker discovery
- clinico-genetic associations

Economics

- meaningful features
- evolution of countries over time
- transition model

Future Work

1. Modeling
   - encode complex prior knowledge
   - generalized ADDP: multiple-input/output, other applications
   - atom-dependent latent feature model
   - ...

Future Work

1. Modeling
   - encode complex prior knowledge
   - generalized ADDP: multiple-input/output, other applications
   - atom-dependent latent feature model
   - ...
2. Inference
   - scale algorithms (e.g., black-box variational inference)
   - better exploration (e.g., split-merge moves)

Future Work

1. Modeling
   - encode complex prior knowledge
   - generalized ADDP: multiple-input/output, other applications
   - atom-dependent latent feature model
   - ...
2. Inference
   - scale algorithms (e.g., black-box variational inference)
   - better exploration (e.g., split-merge moves)
3. Validation
   - new "data exploration" metrics
   - how to quantify model utility?

Thank you for listening!

Any questions?

Journal Publications

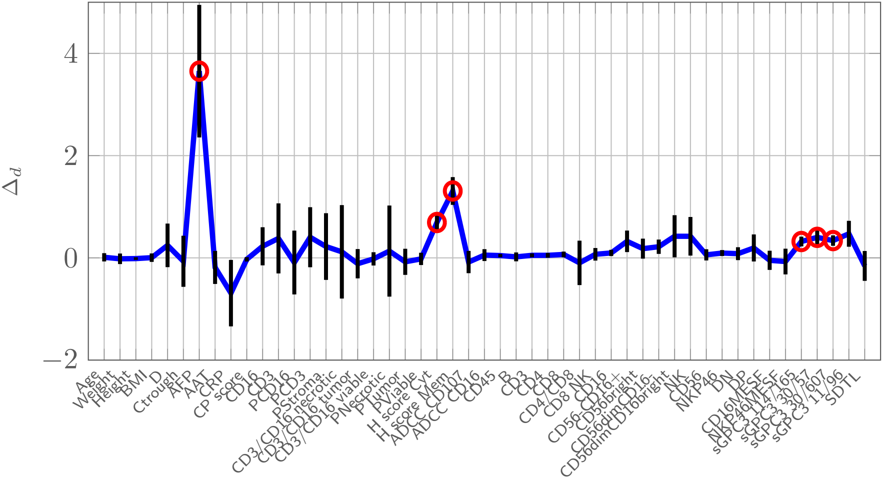
- Melanie F. Pradier, Francisco J. R. Ruiz, and Fernando Perez-Cruz, **"Prior design for dependent Dirichlet processes: An application to marathon modeling,"** *PLoS ONE*, vol. 11, no. 1, pp. e0147402, Jan. 2016, doi:10.1371/journal.pone.0147402.
- Melanie F. Pradier, Bernhard Reis, Lori Jukofsky, Francesca Milletti, Toshihiko Ohtomo, Fernando Perez-Cruz, and Oscar Puig, **"Indian Buffet process identifies NK cell biomarkers as predictors of response to Codrituzumab in patients with advanced hepatocellular carcinoma.,"** *Submitted to BMC Cancer*, September 2017.
- Isabel Valera, Melanie F. Pradier, and Zoubin Ghaharamani, **"General latent feature model for heterogeneous datasets,"** *Submitted to Journal of Machine Learning Research*, June 2017, arXiv:1706.03779.
- Melanie F. Pradier, Pablo M. Olmos, and Fernando Perez-Cruz, **"Entropy-constrained scalar quantization with a lossy-compressed bit,"** *Entropy*, vol. 18, no. 12, pp. 449, 2016, doi:10.3390/e18120449.

Workshop Publications

- Isabel Valera, Melanie F. Pradier, and Zoubin Ghahramani, **"General latent feature modeling for data exploration tasks,"** *Workshop on Human Iinterpretability in Machine Learning at Neural Information Processing Systems*, 2017, arXiv:1707.08352.
- Melanie F. Pradier, Theofanis Karaletsos, Stefan Stark, Julia E. Vogt, Gunnar Ratsch, and Fernando Perez-Cruz, **"Bayesian Poisson factorization for genetic associations with clinical features in cancer,"** in *Machine Learning for Healthcare Workshop in Neural Information Processing Systems*, 2015.
- Melanie F. Pradier and Fernando Perez-Cruz, **"Infinite mixture of global Gaussian processes,"** in *Bayesian Non-parametric: the Next Generation Workshop in Neural Information Processing Systems*, 2015.
- Melanie F. Pradier, Stefan Stark, Stephanie Hyland, Julia E. Vogt, and Gunnar Ratsch, **"Large-scale sentence clustering from electronic health records for genetic associations in cancer,"** in *Machine Learning for Computational Biology Workshop in Neural Information Processing Systems*, 2015.
- Melanie F. Pradier, Pablo G. Moreno, Francisco J. R. Ruiz, Isabel Valera, Harold Molina-Bulla, and Fernando Perez-Cruz, **"Map/reduce uncollapsed Gibbs sampling for Bayesian nonparametric models,"** in *Software Engineering for Machine Learning Workshop in Neural Information Processing Systems*, 2014.

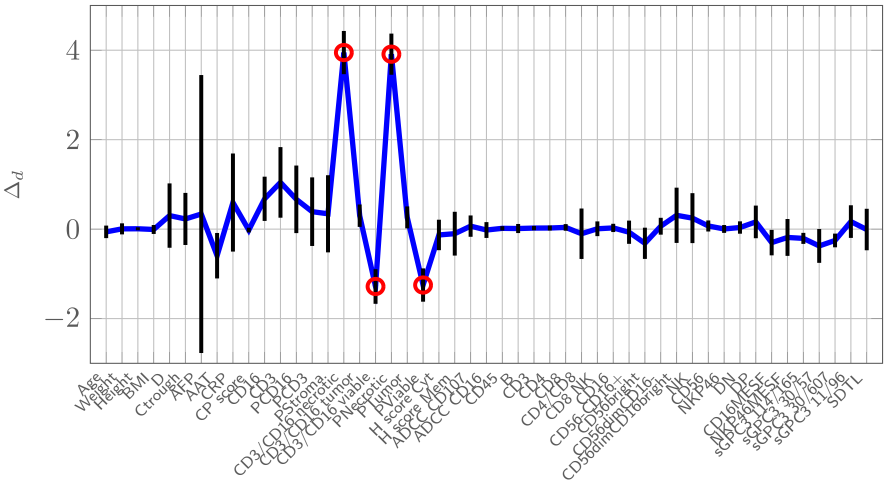
Results: biomarker discovery

Global feature F1

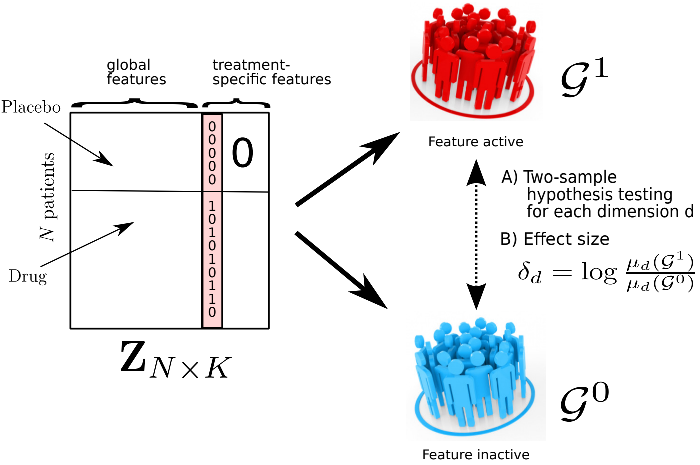


Results: biomarker discovery

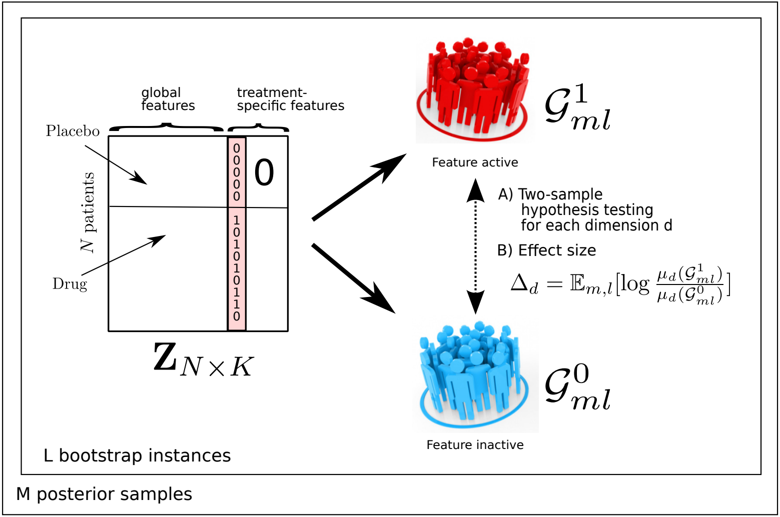
Global feature F2



Statistical procedure for biomarker discovery



Statistical procedure for biomarker discovery



Appendix: Inference in PFA models

- Markov Chain Monte Carlo approach.
- Conditional conjugacy using auxiliary variables.

$$x_{nd} = \sum^{K} x^{'}_{nd,k} \quad \text{where} \quad x^{'}_{nd,k} \sim \text{Poisson}(\mathbf{Z}_{n\bullet}\mathbf{B}_{\bullet d})$$

- Truncated approximation of feature weights
- In 3RBeP-PFA, dynamic programming to compute likelihood (Doshi-Velez et.al, 2015)
- In dBeP-PFA, forward-filtering backward-sampling procedure (Gael et.al, 2009)

Appendix: Results for 3RBeP-PFA

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

(c) 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

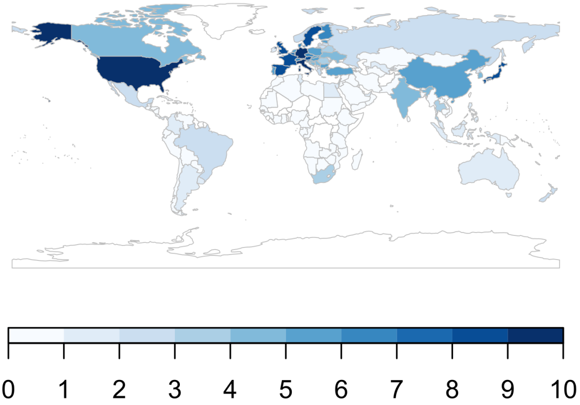
(d) S3R-IBP

Deep 3RBeP-PFA: using a 2nd layer

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

Deep 3RBeP-PFA: using a 2nd layer

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



Appendix: Modeling dBeP-PFA

Dynamic PFA

- T timestamps (years)
- markov IBP to account for temporal dynamics (Gael et.al, 2009)
- Generative model:

$$
\begin{aligned}
x_{nd}^{(t)} &\sim \text{Poisson}\left(\mathbf{Z}_{n\bullet}^{(t)}\,\mathbf{B}_{\bullet d}\right)\\
B_{kd} &\sim \text{Gamma}(\alpha_B, \frac{\mu_B}{\alpha_B})\\
a_k &\sim \text{Beta}(\frac{\alpha}{K}, 1),\\
b_k &\sim \text{Beta}(\gamma, \delta),\\
z_{nk}^{(t)}|a_k, b_k &\sim \text{Bernoulli}\left(a_k^{1-z_{nk}^{(t-1)}}\, b_k^{z_{nk}^{(t-1)}}\right)
\end{aligned}
$$

The transition matrix Q_k for feature k is given by:

$$
Q_k = \begin{pmatrix} 1-a_k & a_k \\ 1-b_k & b_k \end{pmatrix}
$$

Appendix: Inference dBeP-PFA

Inference

- MCMC approach, e.g., Gibbs sampler + slice sampler for the IBP
- K Poisson-distributed auxiliary random variables, i.e., $x_{nd}^{(t)} = \sum_{k=1}^{K} r_{nd,k}^{(t)}$
- Forward Filtering Backward Sampling (FFBS) to approximate $p(\mathbf{Z}|\mathbf{X}, \mathbf{B})$

$$p(\mathbf{X}_{n\bullet}^{(1:t)}, z_{nk}^{(t)}|-) = p(\mathbf{X}_{n\bullet}^{(t)}|z_{nk}^{(t)}, -) \sum_{z_{nk}^{(t-1)}} p(\mathbf{X}_{n\bullet}^{(1:t-1)}, z_{nk}^{(t-1)}|-)p(z_{nk}^{(t)}|z_{nk}^{(t-1)})$$

  - Forward step: compute $p(z_{nk}^{(t)}|\mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B})$
  - Backward step: sample from $p(z_{nk}^{(t)}|z_{nk}^{(t+1)}, \mathbf{X}_{n\bullet}^{(1:t)}, \mathbf{Z}_{n,\neg k}^{(t)}, \mathbf{B})$

Appendix: Results for dBeP-PFA

Id	Top-3 products with highest weights
F0	(bias) crude petroleum, crustaceans, cereals
F1	light fixtures, locksmith hardw., misc. ceramic ornaments
F2	inorganic esters, chemical products, nitrogen compound
F3	iron sheets, iron wire, thin iron sheets
F4	misc. elect. machinery, typewriters, misc. office equipment
F5	soaps, confectionary sugar, baked goods
F6	bovine – equine entrails, bovine meat, misc. prepared meats
F7	knit clothing accessories, linens, leather accessor.
F8	glazes, textiles fabrics for machinery, mineral wool
F9	misc. vegetables, grapes – raisins, misc. fruit
F10	inorganic bases, nitrogenous fertilizers, lubricating petrol. oils
F11	imitation jewellery, embroidery, synth. precious stones
F12	coffee, non-coniferous worked wood, cane sugar
F13	copper ores, chemical wood pulp, misc. non-ferrous ores
F14	pepper, vegetable planting materials, natural rubber
F15	raw cotton, cotton linters, green groundnuts