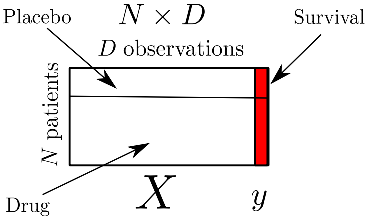
Indian Buffet Process for Biomarker Discovery

Melanie F. Pradier

Dep. of Signal Theory and Communications, University Carlos III in Madrid

Advisor: Fernando Perez-Cruz
Collaborators: Francesca Milletti, Oscar Puig
January 11th, 2015

Problem Formulation



1. Which observations have an impact on survival? (prognostic vars.)
2. Which observations make the drug work? (predictive variables)

Problem Formulation

Challenges

- Noisy/missings
- Uncertainty
- Complexity
- Heterogeneity
- $N << D$

Potential Approaches

X: observations matrix, y: survival, θ: model parameters, W: latent variables

Supervised Methods

$$y = f(X;\theta) + \epsilon, \quad p(y|X,\theta) \tag{1}$$

- Examples: Linear Regression, Lasso (Penalized LR), Gaussian Process, Random Forest, ...
- Problems: Not so easy to interpret, and $N << D$ (suitable for prediction)

Unsupervised Methods

$$(X,y) = f(W;\theta) + \epsilon, \quad p(y,X|W,\theta) \tag{2}$$

- Examples: Dimensionality Reduction, Clustering, **Latent Factors**, ...
- Advantages: Interpretable, flexible (suitable for data exploration)

Our Approach

Challenges

- Noisy/missings
- Uncertainty
- Complexity
- Heterogeneity
- $N << D$

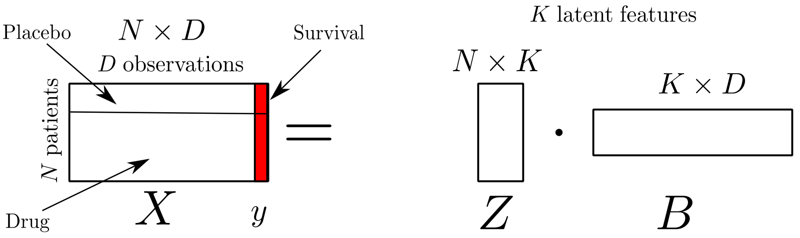
Solutions

- Probabilistic Models
- Bayesian Approach
- Non-parametric
- Generalized
- Sharing Information

- Bayesian: Put a prior over assumptions
- Non-parametric: Model complexity, i.e., number of latent vars., is also infered

In particular...

- We focus on a Latent Factor Model.

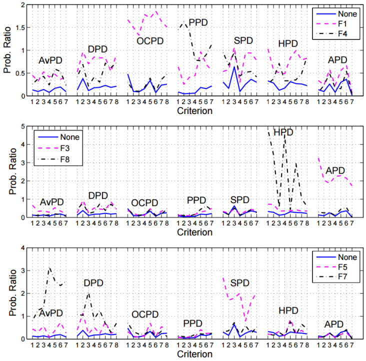


$$Z \sim \text{Indian Buffet Process}(\alpha) \tag{3}$$

- If we know B, patients are independent

Previous Works using the IBP

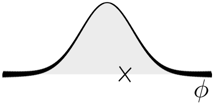
- Identify patients at risk of suicide attempts [F.J.R. Ruiz et.al, NIPS2012].
- Find out latent relationship among psyquiatric disorders [F.J.R. Ruiz et.al, JMLR2014, I. Valera et.al, NC2015].



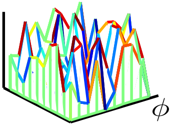
Outline

1. Motivation
2. Indian Buffet Process
3. Results
4. Conclusions

Indian Buffet Process



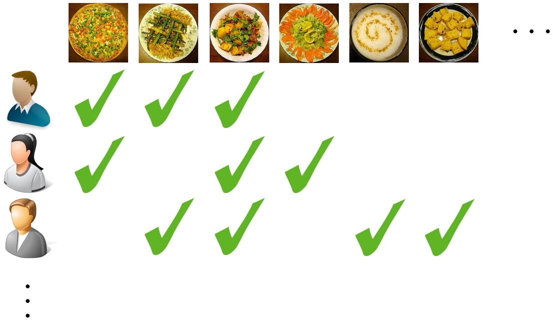
$$\xrightarrow[x \sim \mathcal{N}(\mu_0, \sigma_0^2)]{} \qquad x = 0.356$$



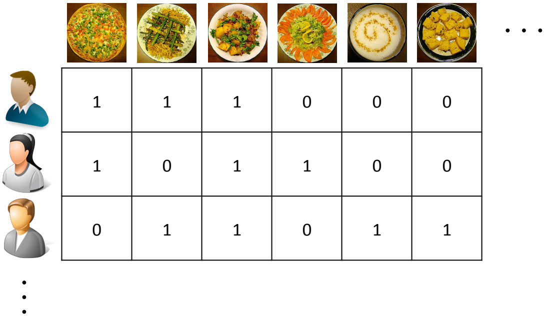
$$\xrightarrow[Z \sim \mathrm{IBP}(\alpha)]{} \qquad Z = \begin{bmatrix} 1 & 0 & 1 & \\ 0 & 1 & 0 & \vdots \\ 1 & 1 & 0 & \\ 1 & 0 & 1 & \end{bmatrix}$$

- IBP: distribution over binary matrices $Z_{N \times K}$
- Model chooses number of hidden features, $K \to \infty$
- Finite N implies finite number of non-zero columns K_+.

Indian Buffet Process
(Slides from F.J.R.Ruiz)

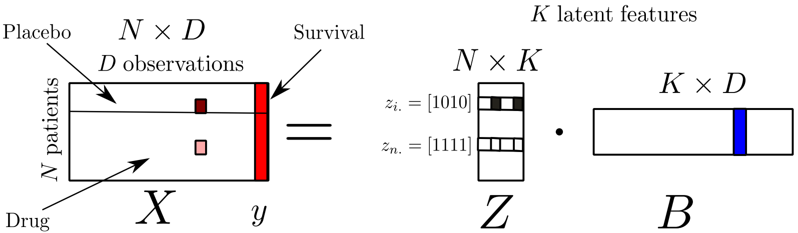


Indian Buffet Process
(Slides from F.J.R.Ruiz)



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

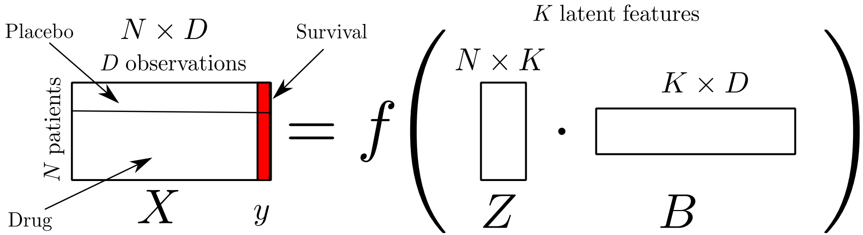
Infinite Latent Feature Model



- ■ $x_{id} = 173\,\mathrm{ml/dL} = 73 + 0 + 100\,\mathrm{ml/dL}$
- □ $x_{nd} = 136\,\mathrm{ml/dL} = 86 + 40 + 60 - 50\,\mathrm{ml/dL}$

Note: Correlation does not imply causality!

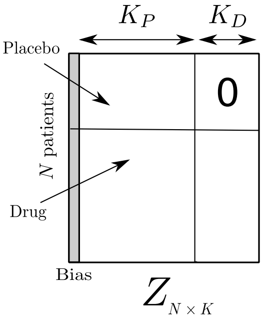
What about heterogeneous data?



- Generalized IBP [I.Valera et.al, 2015]
- Link function f depending on data type

What about $N << D$ problem?

- Placebo patients define background population
- Some extra features only for patients taking the drug
- Shared information between all patients



Methodology

1. Sample from posterior $p(Z|\text{data})$ to identify interesting subpopulations
2. Analysis of feature effect on observations
   - Define patterns of interest G^* and reference G^B
   - Do Bootstrapping L times (to deal with low N)
   - Compute measure of effect size and significance

Outline

1. Motivation
2. Indian Buffet Process
3. Results
4. Conclusions

Database

GC33 Antibody Treatment against Liver Cancer

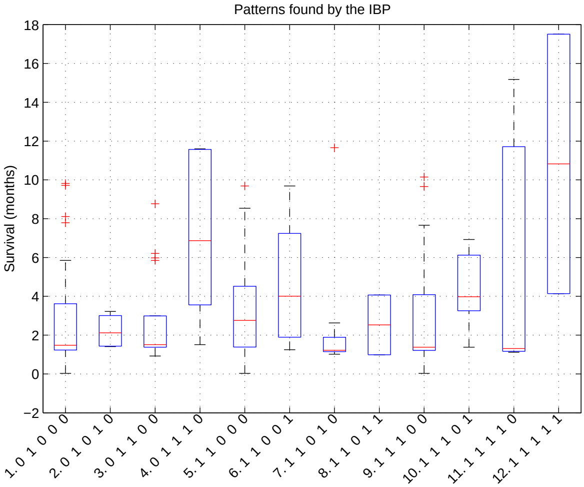
- Clinical trial with N =180 patients
- 60 patients take Placebo, 120 take the drug
- $D = 80$ observations (including demographics, clinical data, and survival)

- Our model infers:
  - 3 features to define whole population
  - 1 extra feature for Drug population

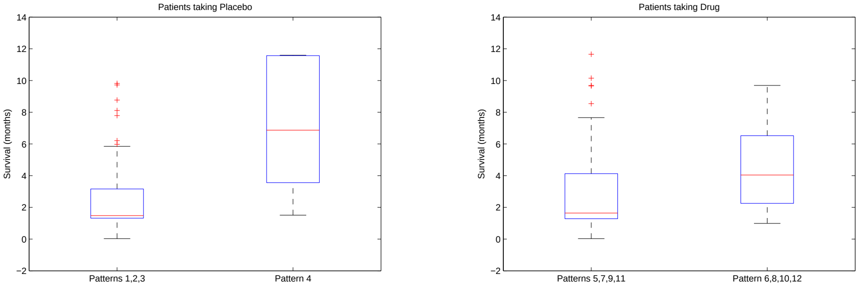
Analysis of Clinical Data

Nr.	Patterns					Occur. (number patients)	Mean TFPD (months)	Median TFPD (months)
	F1	F2	F3	F4	F5			
1.	0	1	0	0	0	33.37	3.06	1.65
2.	0	1	0	1	0	4.07	2.29	2.24
3.	0	1	1	0	0	17.84	2.72	1.81
4.	0	1	1	1	0	4.72	7.05	7.18
5.	1	1	0	0	0	51.52	3.22	2.55
6.	1	1	0	0	1	16.77	4.17	3.65
7.	1	1	0	1	0	8.38	1.74	1.33
8.	1	1	0	1	1	2.07	2.69	2.65
9.	1	1	1	0	0	29.88	3.36	2.03
10.	1	1	1	0	1	4.90	4.44	4.34
11.	1	1	1	1	0	4.53	6.31	5.31
12.	1	1	1	1	1	1.94	10.04	10.01
Total	120.00	180.00	63.82	25.72	25.69	180	3.44	2.04

Different Survival in Subpopulations

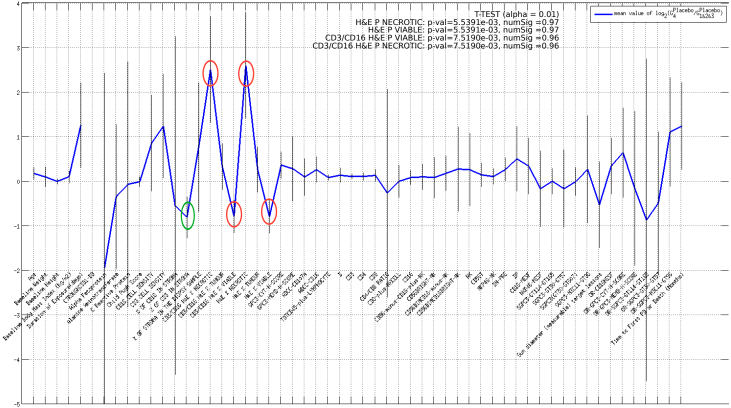


Different Survival in Subpopulations



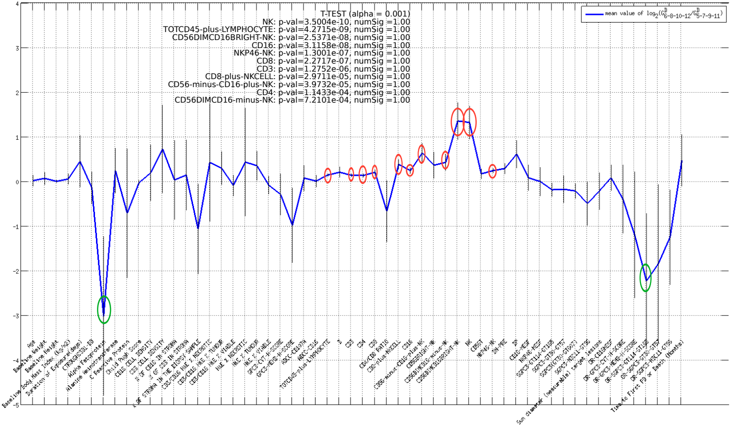
Strong Placebo Vs Normal Placebo

1. Which observations have an impact on survival?



Strong Drug Vs Normal Drug

2. Which observations make the drug work?



T-TEST (alpha = 0.001)
NK: p-val=3.5004e-10, numSig =1.00
TOTCD45-plus-LYMPHOCYTE: p-val=4.2715e-09, numSig =1.00
CD56DIMCD16BRIGHT-NK: p-val=2.5371e-08, numSig =1.00
CD16: p-val=3.1158e-08, numSig =1.00
NKP46-NK: p-val=1.3001e-07, numSig =1.00
CD8: p-val=2.2717e-07, numSig =1.00
CD3: p-val=1.2752e-06, numSig =1.00
CD8-plus-NKCELL: p-val=2.9711e-05, numSig =1.00
CD56-minus-CD16-plus-NK: p-val=3.9732e-05, numSig =1.00
CD4: p-val=1.1433e-04, numSig =1.00
CD56DIMCD16-minus-NK: p-val=7.2101e-04, numSig =1.00

Outline

1. Motivation
2. Indian Buffet Process
3. Results
4. Conclusions

Conclusions

In this talk...

- Bayesian Non-parametrics for Data Exploration
- Indian Buffet Process in Latent Feature Models
- IBP Adaptation for Clinical Trial Problem

In particular...

1. Identification of subpopulations
2. Potential prognostic and predictive variables
3. Ongoing work:
    - Analysis of RNA-seq data ($D = 48.000$)
    - Improve Statistical Test (Maximum Mean Discrepancy)

References

1. S. J. Gershman and D. M. Blei, **A tutorial on Bayesian nonparametric models**, Journal of Mathematical Psychology, vol. 56, no. 1, pp. 1-12, Feb. 2012.
2. T. L. Griffiths and Z. Ghahramani, **The Indian Buffet Process: An Introduction and Review**, J. Mach. Learn. Res., vol. 12, pp. 1185-1224, Jul. 2011.
3. D. Knowles and Z. Ghahramani, **Nonparametric Bayesian Sparse Factor Models with application to gene expression modeling**, The Annals of Applied Statistics, vol. 5, no. 2B, pp. 1534-1552, 2011.
4. F. J. R. Ruiz, I. Valera, C. Blanco, and F. Pérez-Cruz, **Bayesian Nonparametric Modeling of Suicide Attempts**, in Advances in Neural Information Processing Systems 25, F. Pereira, C. J. C. Burges, L. Bottou, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2012, pp. 1853-1861.
5. F. J. R. Ruiz, I. Valera, C. Blanco, and F. Perez-Cruz, **Bayesian Nonparametric Comorbidity Analysis of Psychiatric Disorders**, J. Mach. Learn. Res., vol. 15, no. 1, pp. 1215-1247, Jan. 2014.
6. I. Valera, F. J. R. Ruiz, P. M. Olmos, C. Blanco, and F. Perez-Cruz, **Infinite Continuous Feature Model for Psychiatric Comorbidity Analysis**, Neural Comput, pp. 1-28, Dec. 2015.
7. I. Valera and Z. Ghahramani, **General Table Completion using a Bayesian Nonparametric Model**, in Advances in Neural Information Processing Systems 27, Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, Eds. Curran Associates, Inc., 2014, pp. 981-989.

Acknowledgments

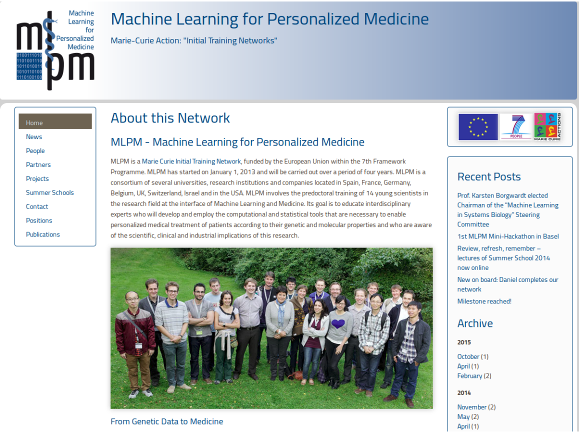
- Fernando Perez-Cruz
- Francesca Miletti
- Oscar Puig
- Fran J.R. Ruiz
- Isabel Valera
- Rätsch Lab at MSKCC
- TSC at UC3M
- Marie-Curie ITN-MLPM

Thank you!

Appendix

Appendix

European Initiative: Marie-Curie ITN-MLPM
Machine Learning for Personalized Medicine



Machine Learning for Personalized Medicine

Marie-Curie Action: "Initial Training Networks"

- Home
- News
- People
- Partners
- Projects
- Summer Schools
- Contact
- Positions
- Publications

About this Network

MLPM - Machine Learning for Personalized Medicine

MLPM is a Marie Curie Initial Training Network, funded by the European Union within the 7th Framework Programme. MLPM has started on January 1, 2013 and will be carried out over a period of four years. MLPM is a consortium of several universities, research institutions and companies located in Spain, France, Germany, Belgium, UK, Switzerland, Israel and in the USA. MLPM involves the predoctoral training of 14 young scientists in the research field at the interface of Machine Learning and Medicine. Its goal is to educate interdisciplinary experts who will develop and employ the computational and statistical tools that are necessary to enable personalized medical treatment of patients according to their genetic and molecular properties and who are aware of the scientific, clinical and industrial implications of this research.

From Genetic Data to Medicine

Recent Posts

- Prof. Karsten Borgwardt elected Chairman of the "Machine Learning in Systems Biology" Steering Committee
- 1st MLPM Mini-Hackathon in Basel
- Review, refresh, remember – lectures of Summer School 2014 now online
- New on board: Daniel completes our network
- Milestone reached!

Archive

2015

October (1)
April (1)
February (2)

2014

November (2)
May (2)
April (1)

- 3.75M €
- 4 years
- 14 students
- 8 countries
- 13 institutions

www.mlpm.eu

European Initiative: Marie-Curie ITN-MLPM

Machine Learning for Personalized Medicine

MAX-PLANCK-GESELLSCHAFT
The University Of Sheffield.
SIEMENS
PHARMATICS
Université de Liège
ULg
Inserm
MINES ParisTech
PRINCIPE FELIPE
CENTRO DE INVESTIGACION
Universidad Carlos III de Madrid
Memorial Sloan Kettering Cancer Center
1884
ETH zürich
IBM
Roche

European Initiative: Marie-Curie ITN-MLPM

Machine Learning for Personalized Medicine

Main Objective

Develop statistical tools and computational methods for personalized medicine.

- Precise Diagnosis
- Specific Treatments

Research Lines

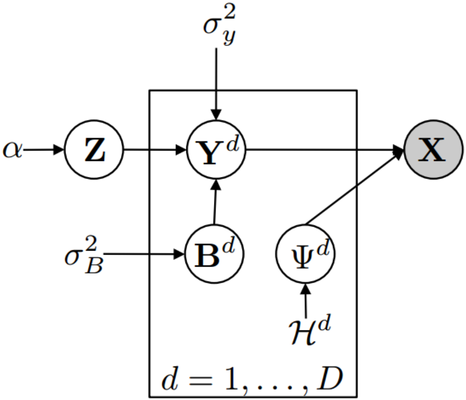
- Biomarker Discovery
- Heterogeneous Data Integration
- Causal Mechanisms of Diseases
- Gene-Environment Interactions

My Focus

Probabilistic Modeling $\longrightarrow$ Bayesian Non-parametric Models

What about heterogeneous data?

- Generalized IBP [I.Valera et.al, 2015]
- Link function f depending on data type



Measures for Effect Size and Significance

Continuous variable d

- Effect Size

$$\beta_d = \frac{1}{L} \sum_{l=1}^{L} \log_2 \left(\frac{\mu_d(\widetilde{G_l^*})}{\mu_d(\widetilde{G_l^B})} \right)$$

- Significance
  - Relative Deviation Metric
  - T-Test

Categorical variable r

- Effect Size

$$\beta_r = \frac{1}{L} \sum_{l=1}^{L} \left(\mu_d(\widetilde{G_l^*}) - \mu_d(\widetilde{G_l^B}) \right)$$

- Significance
  - Binomial Test
  - Fisher Exact Test

Measure of Effect Size

- For continuous variable d:

$$\beta_d = \frac{1}{L}\sum_{l=1}^{L} \log_2 \left(\frac{\mu_d(\widetilde{G_l^*})}{\mu_d(\widetilde{G_l^B})} \right) \qquad (4)$$

- For categorical variable r:

$$\beta_r = \frac{1}{L}\sum_{l=1}^{L} \left(\mu_d(\widetilde{G_l^*}) - \mu_d(\widetilde{G_l^B}) \right) \qquad (5)$$

Measure of Significance
Continuous Variables

For continuous variables, compute:

- Deviation compared to G^* variance

$$\gamma^* = \frac{\left|\mu_d(G^*) - \mu_d(G^B)\right|}{\sigma_d(G^*)} \tag{6}$$

- Deviation compared to G^B variance

$$\gamma^B = \frac{\left|\mu_d(G^*) - \mu_d(G^B)\right|}{\sigma_d(G^B)} \tag{7}$$

- T-test: Standard statistical test to compare two groups of data.

Measure of Significance

Categorical Variables

For categorical variables, compute:

- Distance to Binomial Mean
    - Fit a Binomial distribution to G^B
    - A variable r is considered significant if $\mu_r(G^*)$ is outside confidence interval
- Fisher Exact Test: Standard statistical test for contingency tables.