Indian Buffet Process for Biomarker Discovery

Melanie F. Pradier

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

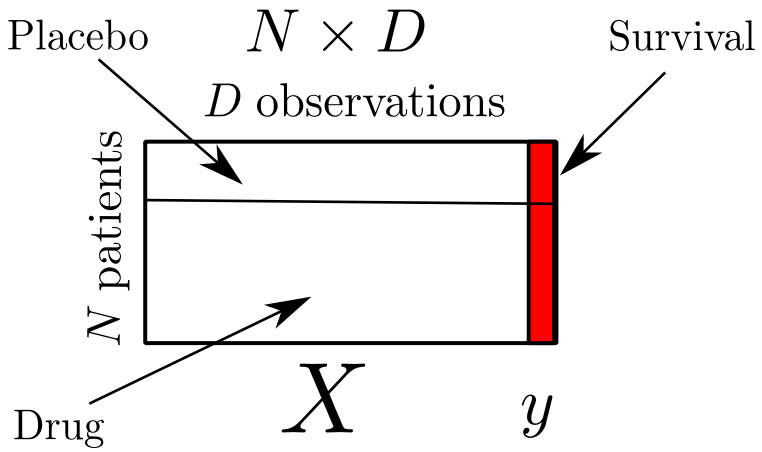
Advisor: Fernando Perez-Cruz
Collaborators: F. Milletti, O. Puig at Roche Innovation Center, New York
S. Stark, S. Hyland, J. Vogt, Gunnar Rätsch at MSKCC, New York

April 6th, 2016

Example 1: Genetic Association Study

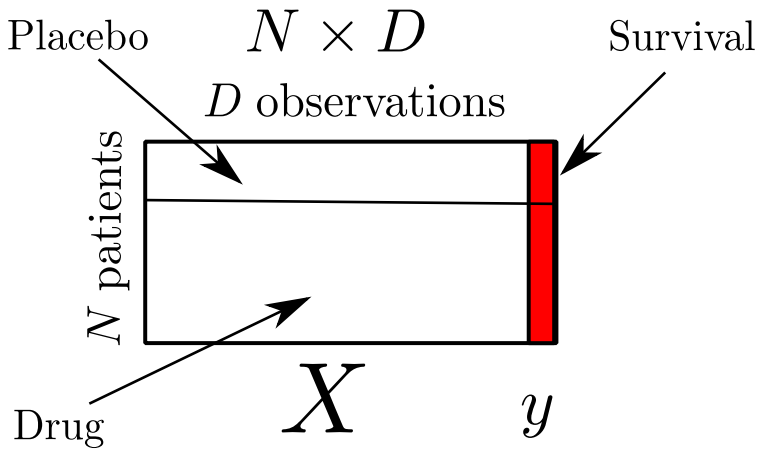
- Objective: Find association between genotype and phenotype.
- We would like a robust method able to deal with epistasis and pleiotropy while taking into account confounders.

Example 2: Clinical Trial



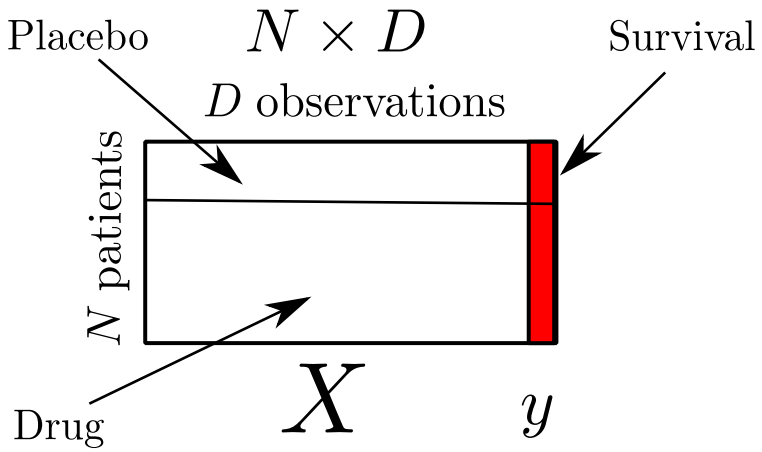
1. Which observations have an impact on survival? (prognostic vars.)
2. Which observations make the drug work? (predictive variables)

Example 2: Clinical Trial



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

Example 2: Clinical Trial



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

Problem Formulation

Aim: Find interesting correlations (data exploration).

Challenges

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

Problem Formulation

Aim: Find interesting correlations (data exploration).

Challenges

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

Potential Approaches

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

Supervised Methods

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

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

Unsupervised Methods

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

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

Potential Approaches

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

Supervised Methods

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

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

Unsupervised Methods

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

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

Potential Approaches

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

Supervised Methods

$$p(y|X,\theta) \tag{1}$$

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

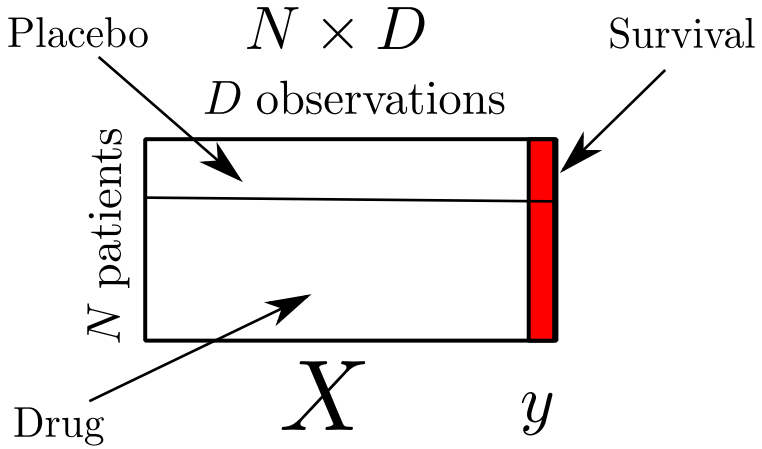
Unsupervised Methods

$$p(y,X|W,\theta) \tag{2}$$

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

Our approach
A probabilistic perspective

- Focus on latent factor models using Bayesian non-parametrics.

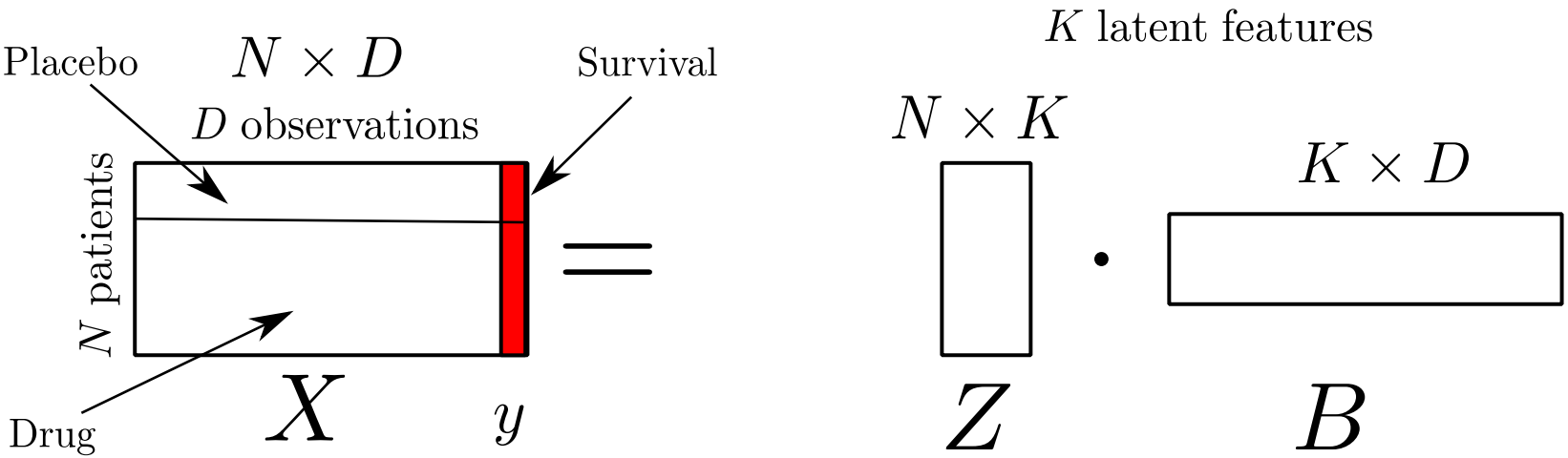


$$Z \sim \text{Indian Buffet Process}(\alpha) \quad (3)$$

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

Our approach
A probabilistic perspective

- Focus on latent factor models using Bayesian non-parametrics.



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

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

Previous Works using the IBP

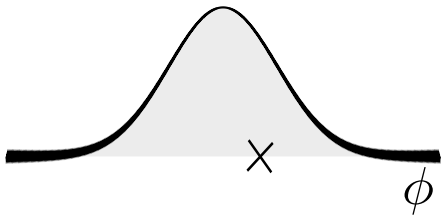
- Identify patients at risk of suicide attempts [F.J.R. Ruiz et.al, NIPS2012].
- Find out latent relationship among psyquiatric disorders [F.J.R. Ruiz et.al, JMLR2014, I. Valera et.al, NC2015].
- Analysis of gene expression data [D. Knowles, and Z. Ghahramani, 2011]
- Discovery of biological interaction networks [H. Son, B. Joseph, 2011]
- Multi-platform Genomics [Ray et. al. 2014]
- Modeling of genetic tumor variations [Chen et. al. 2013, Lee et. al. 2015]

Outline

Indian Buffet Process

- 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_+.

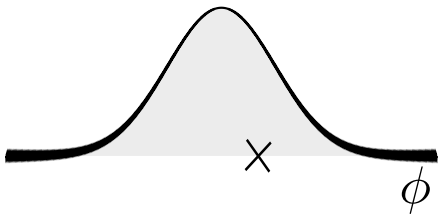
Indian Buffet Process



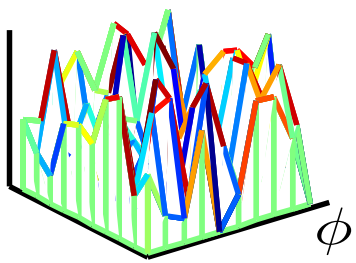
$$x \sim \mathcal{N}(\mu_0, \sigma_0^2) \longrightarrow x = 0.356$$

- 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



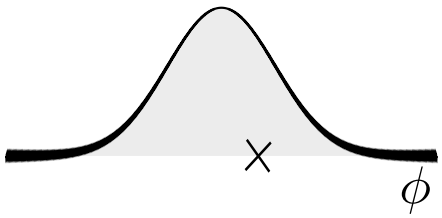
$$x \sim \mathcal{N}(\mu_0, \sigma_0^2) \longrightarrow x = 0.356$$



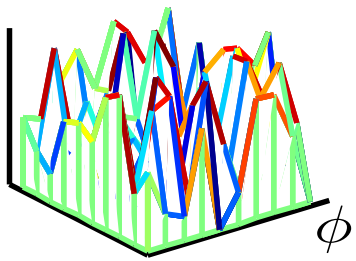
$$Z \sim \mathrm{IBP}(\alpha) \longrightarrow Z = \begin{bmatrix} 1 & 0 & 1 \\ 0 & 1 & 0 \\ 1 & 1 & 0 \\ 1 & 0 & 1 \end{bmatrix} \vdots$$

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

Indian Buffet Process



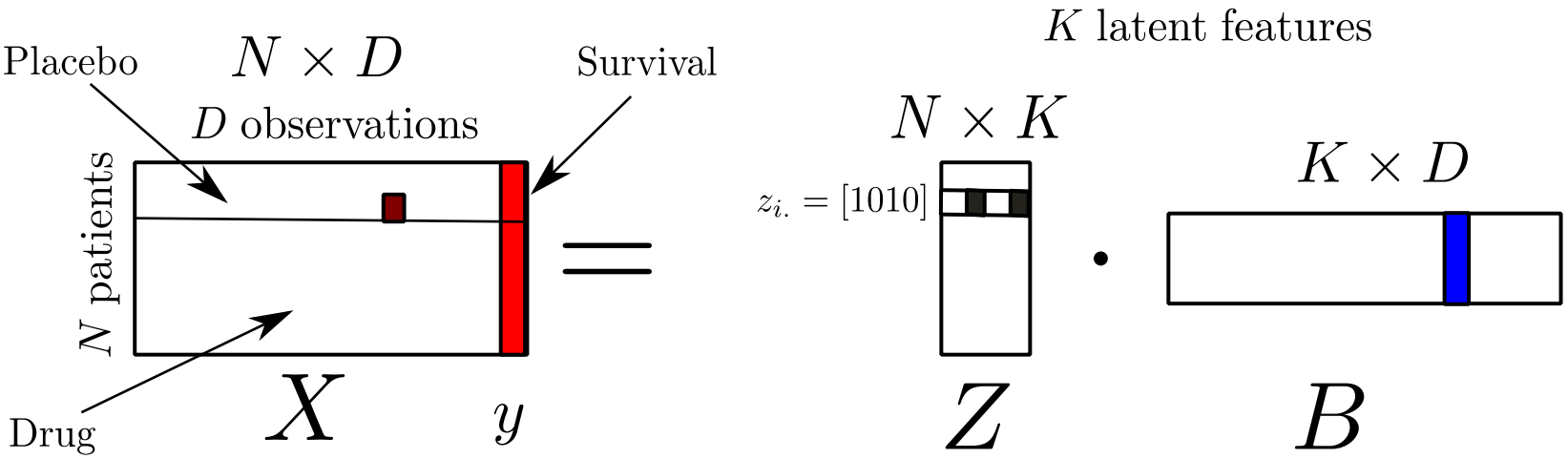
$$x \sim \mathcal{N}(\mu_0, \sigma_0^2) \longrightarrow x = 0.356$$



$$Z \sim \mathrm{IBP}(\alpha) \longrightarrow Z = \begin{bmatrix} 1 & 0 & 1 & \\ 0 & 1 & 0 & \vdots \\ 1 & 1 & 0 & \\ 1 & 0 & 1 & \end{bmatrix}$$

- 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_+.

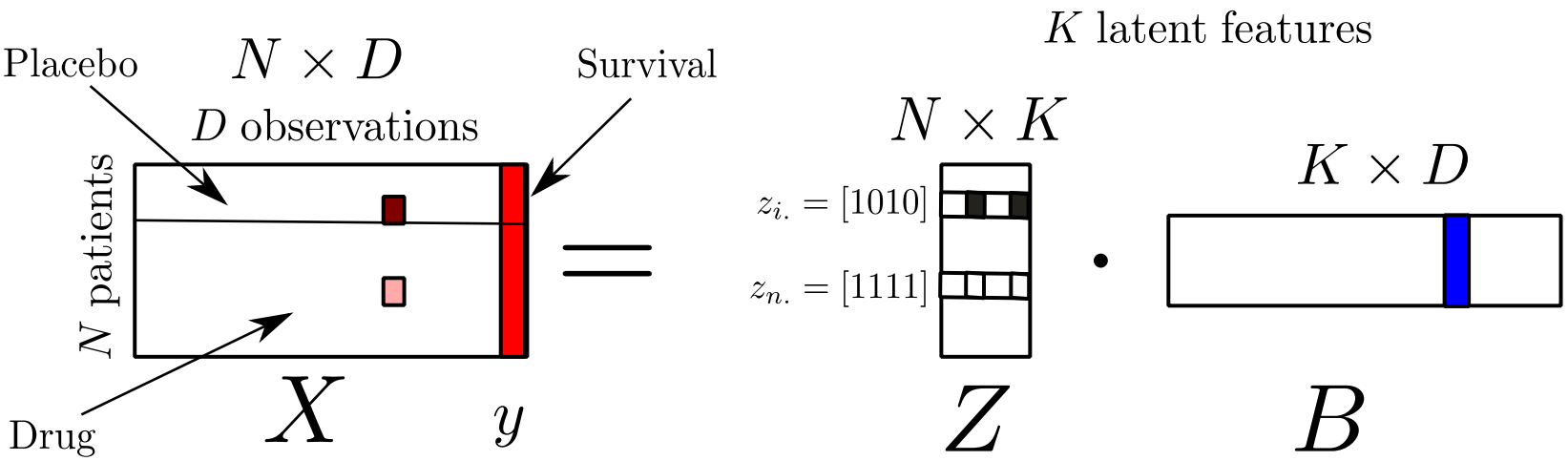
Infinite Latent Feature Model



- $x_{id} = 173\,\mathrm{ml/dL} = 73 + 0 + 100\,\mathrm{ml/dL}$

Note: Correlation does not imply causality!

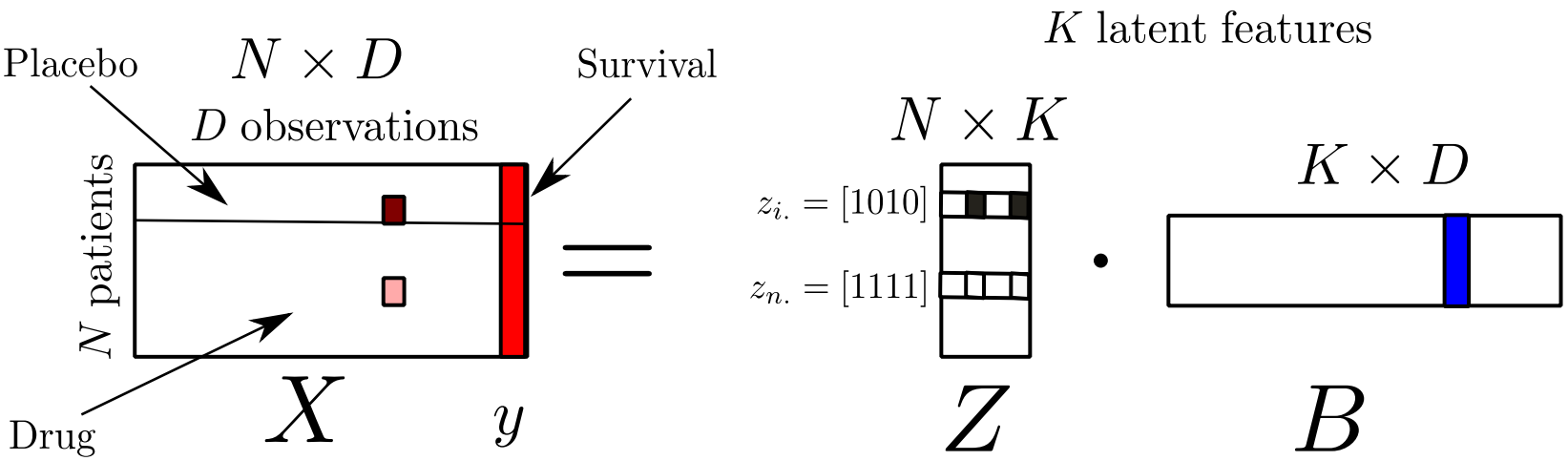
Infinite Latent Feature Model



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

Note: Correlation does not imply causality!

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!

Addressing the challenges

Challenges

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

Our Approach

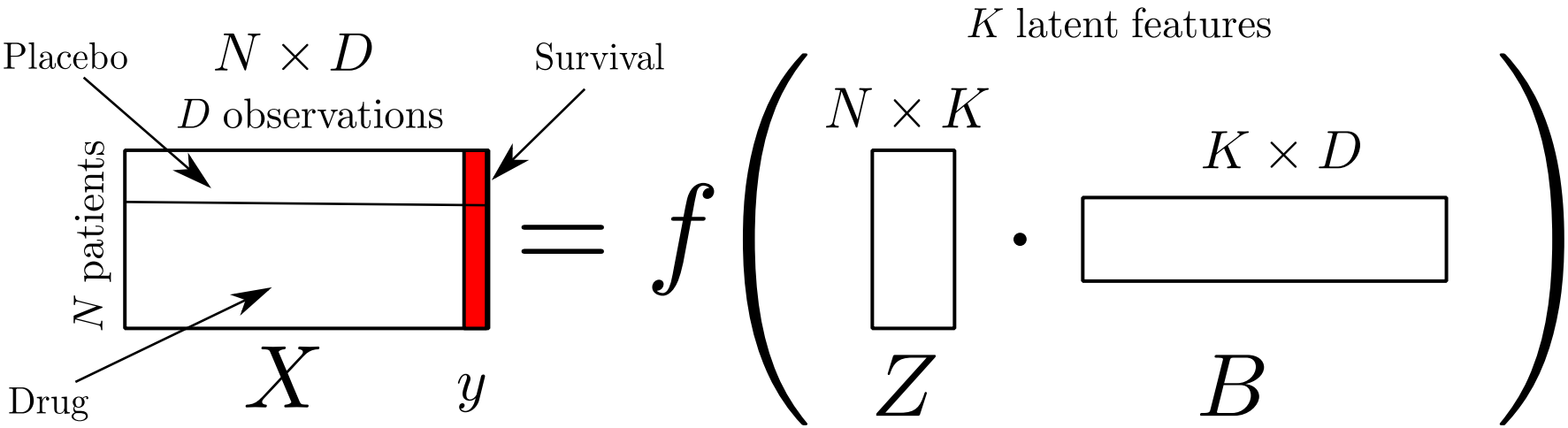
Challenges

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

Our Approach

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

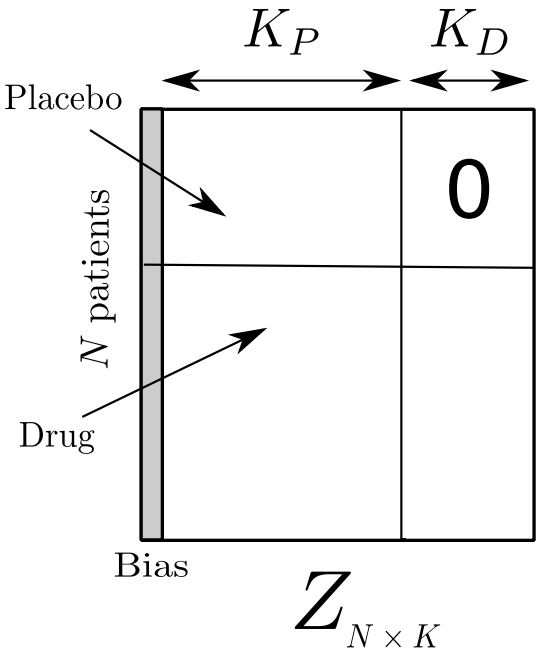
What about heterogeneous data?



- Generalized IBP [I.Valera et.al, 2015]
- Idea: assume auxiliary Gaussian latent variables, and choose link function f depending on data type

How about $N << D$ problem?

1. Shared information between all patients
   - Placebo patients define background population
   - Some extra features only for patients taking the drug
2. Robust method: bootstrapping + soft partitions



Robust 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

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 uses:
  - 3 features to define whole population
  - 1 extra feature for Drug population

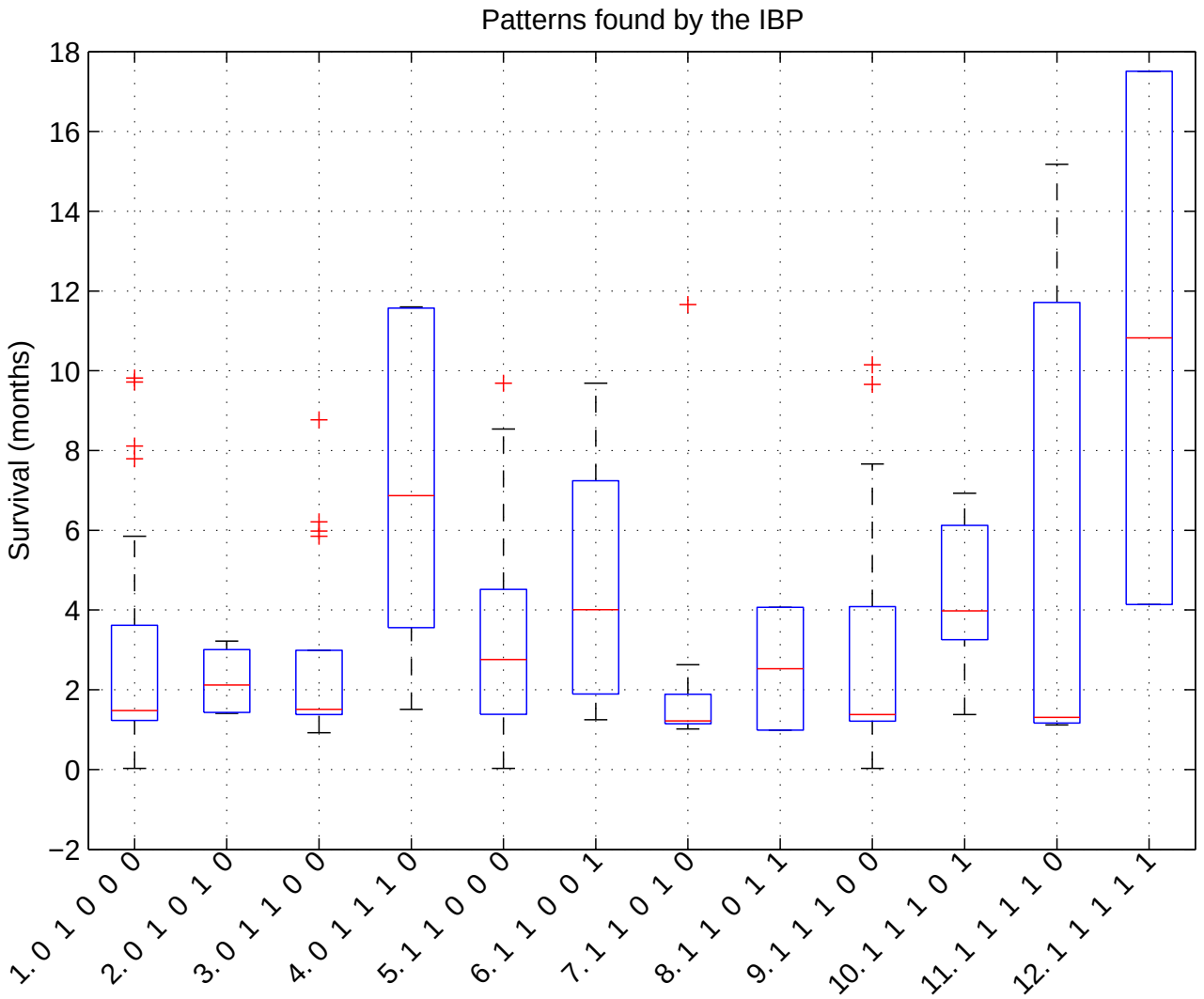
Partition according to Feature Patterns

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

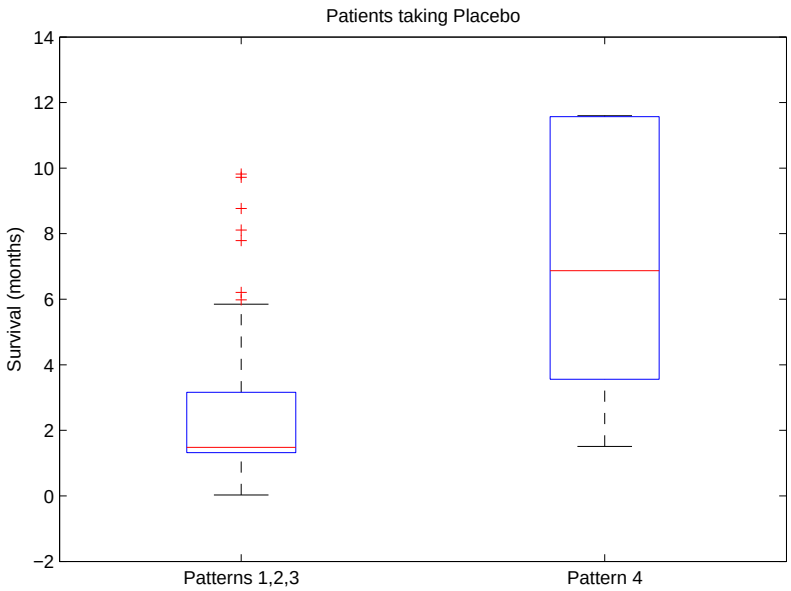
Partition according to Feature Patterns

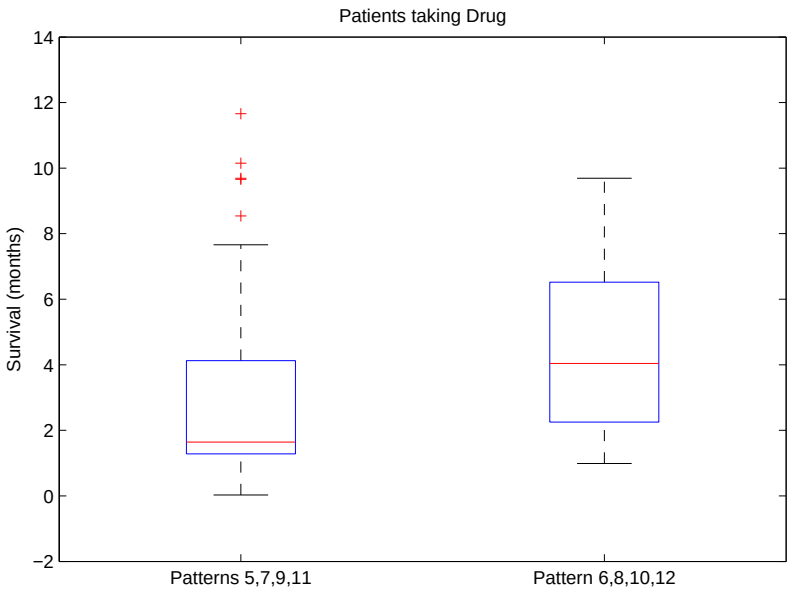
Nr.	Patterns					Occur. (number patients)	Mean TFPD (months)	Median TFPD (months)
	F1	F2	F3	F4	F5			
1.	0	1	0	0	0	33.37	3.06	1.65
2.	0	1	0	1	0	4.07	2.29	2.24
3.	0	1	1	0	0	17.84	2.72	1.81
4.	0	1	1	1	0	4.72	7.05	7.18
5.	1	1	0	0	0	51.52	3.22	2.55
6.	1	1	0	0	1	16.77	4.17	3.65
7.	1	1	0	1	0	8.38	1.74	1.33
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



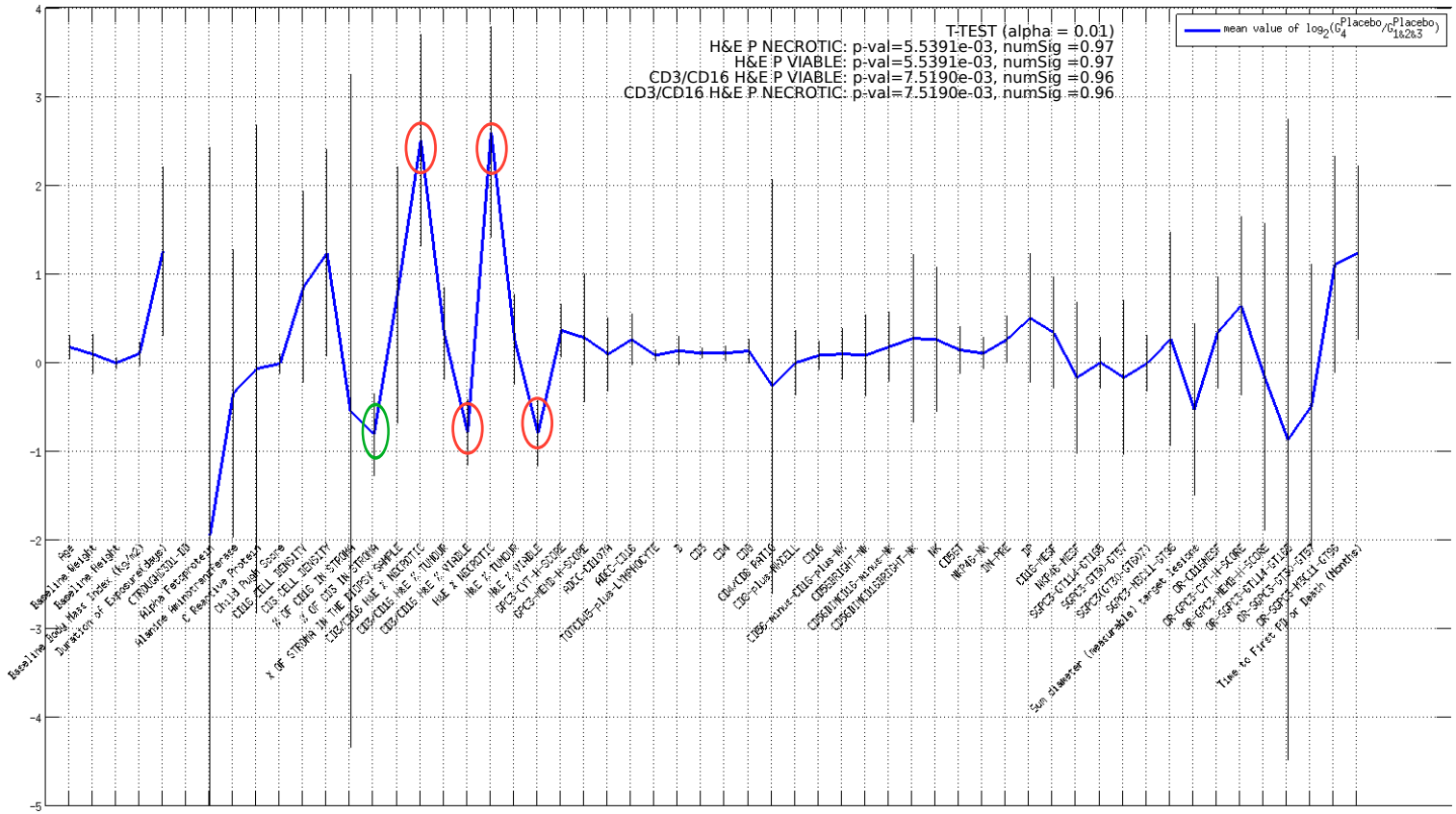
Different Survival in Subpopulations





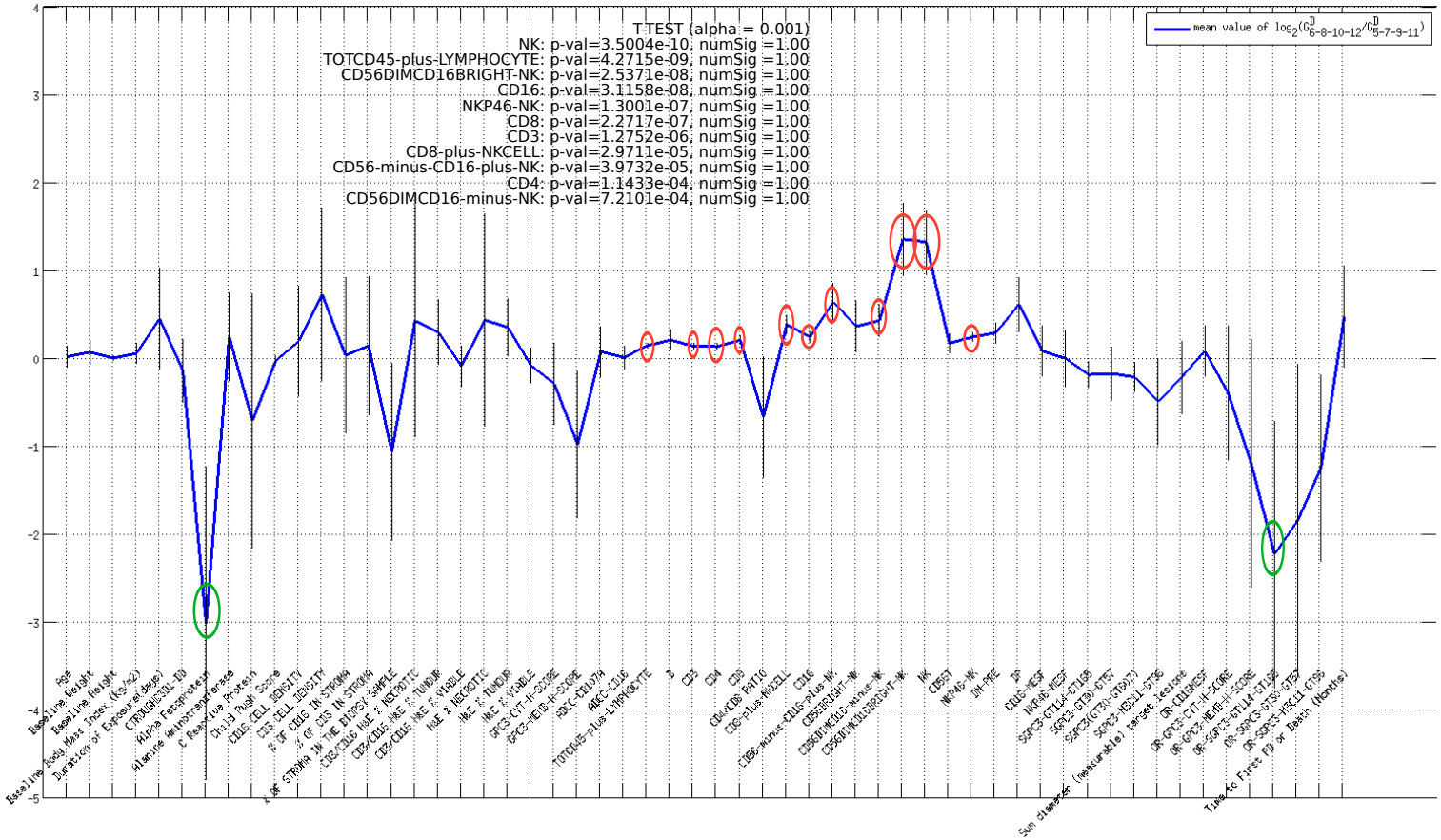
Strong Placebo Vs Normal Placebo

1. Which observations have an impact on survival?



Strong Drug Vs Normal Drug

2. Which observations make the drug work?



Outline

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

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...

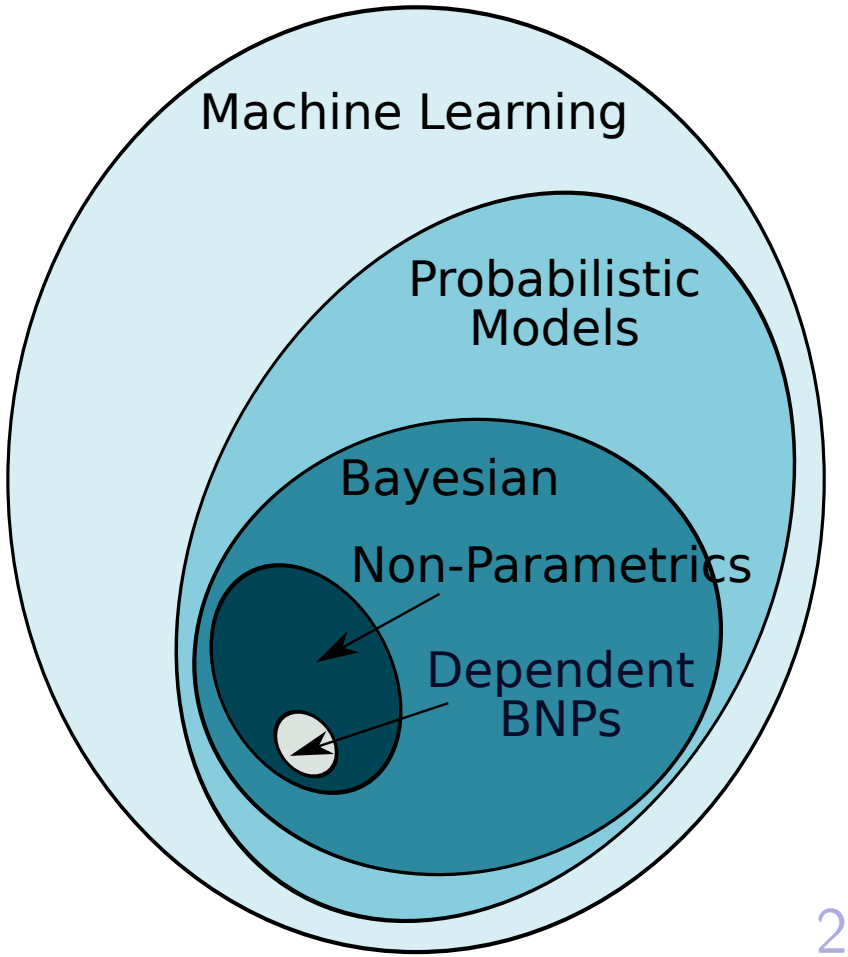
1. Identification of subpopulations
2. Potential prognostic and predictive variables

My contribution to the MLPM network

- Probabilistic models for data exploration

Challenges I address

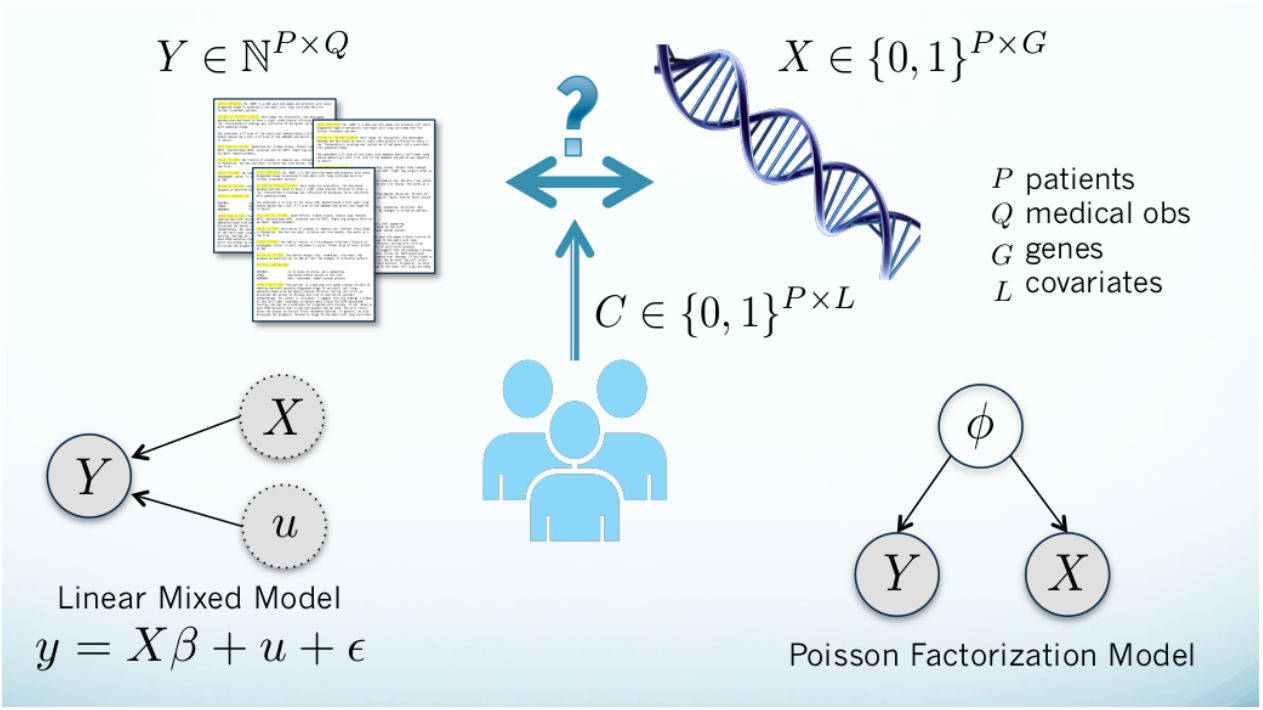
- Noisy/missings
- Uncertainty
- Complexity
- Heterogeneity
- $N << D$
- Confounders



Probabilistic Modeling for Genetic Associations in Cancer

Current and future research results at: www.melaniefpradier.work

- M. F. Pradier, S. Stark, S. Hyland, J. E. Vogt, G. Rätsch and F. Perez-Cruz. **Large-Scale Sentence Clustering from Electronic Health Records for Genetic Associations in Cancer**, Paper at Machine Learning for Computational Biology Workshop (NIPS 2015).
- M. F. Pradier, F. Perez-Cruz and G. Rätsch. **Sparse Poisson Factorization Model for Genetic Associations with Clinical Features in Cancer.** Working paper. 2016.

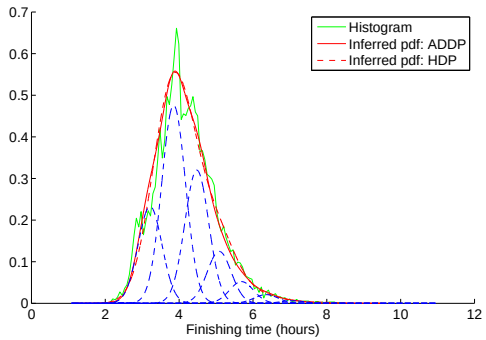


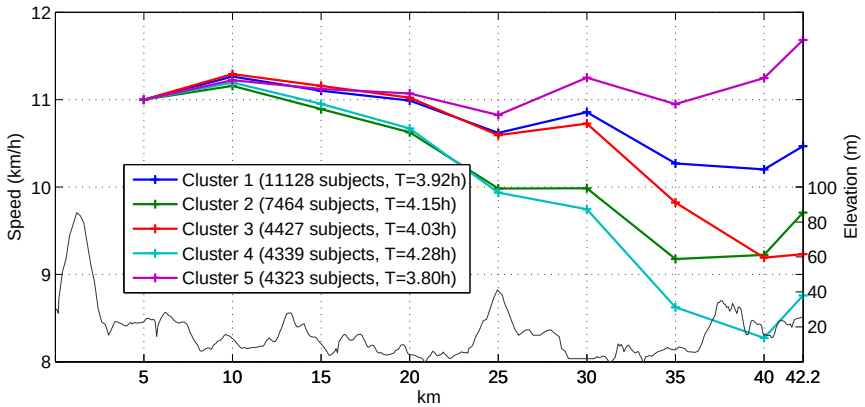
Past Projects

Current and future research results at: www.melaniefpradier.work

M. F. Pradier and F. Perez-Cruz. **Infinite Mixture of Global Gaussian Processes.** Paper at Bayesian Non-parametric: the Next Generation Workshop (NIPS 2015).

M. F. Pradier, F. J. R. Ruiz and F. Perez-Cruz. **Prior Design for Dependent Dirichlet Processes: An Application to Marathon Modeling.** Published at PlosONE. January 2016.





M. F. Pradier, P. G. Moreno, F. J.R. Ruiz, I. Valera, H. Mollina-Bulla and F. Perez-Cruz, **Map/Reduce Uncollapsed Gibbs Sampling for Bayesian Non Parametric Models.** Workshop in Software Engineering for Machine Learning (Software Workshop at NIPS). 2014.

Acknowledgments

Current and future research results at: www.melaniefpradier.work

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

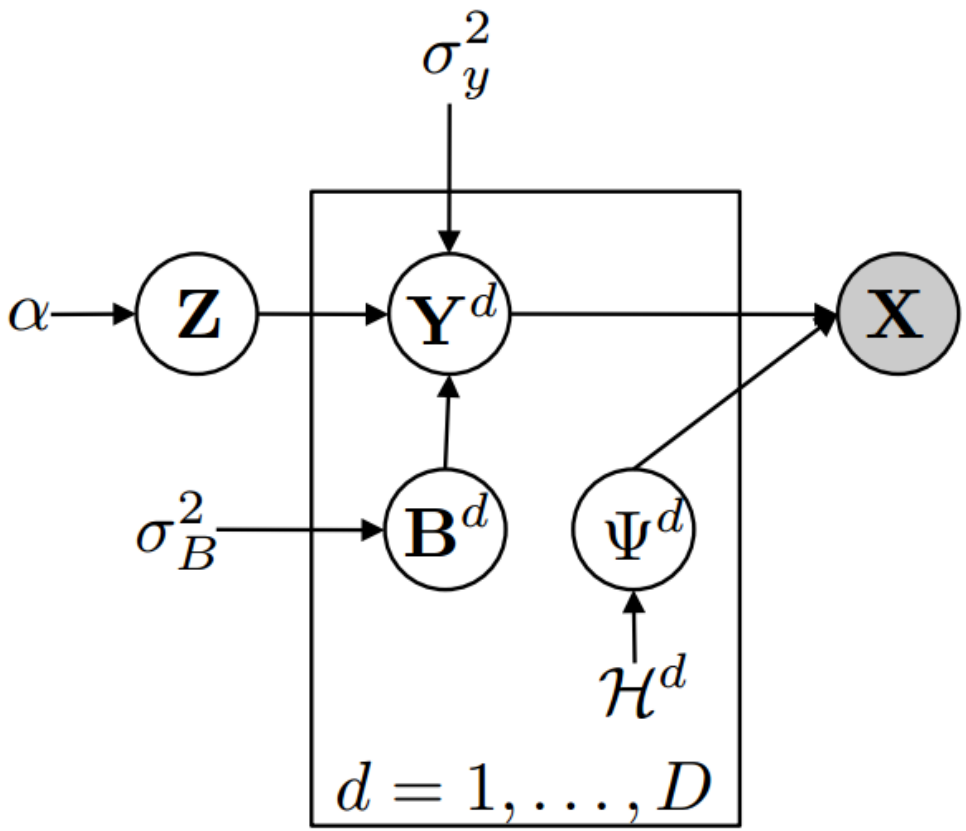
Thank you!

Appendix

Appendix

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:

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

- For categorical variable r:

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

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.

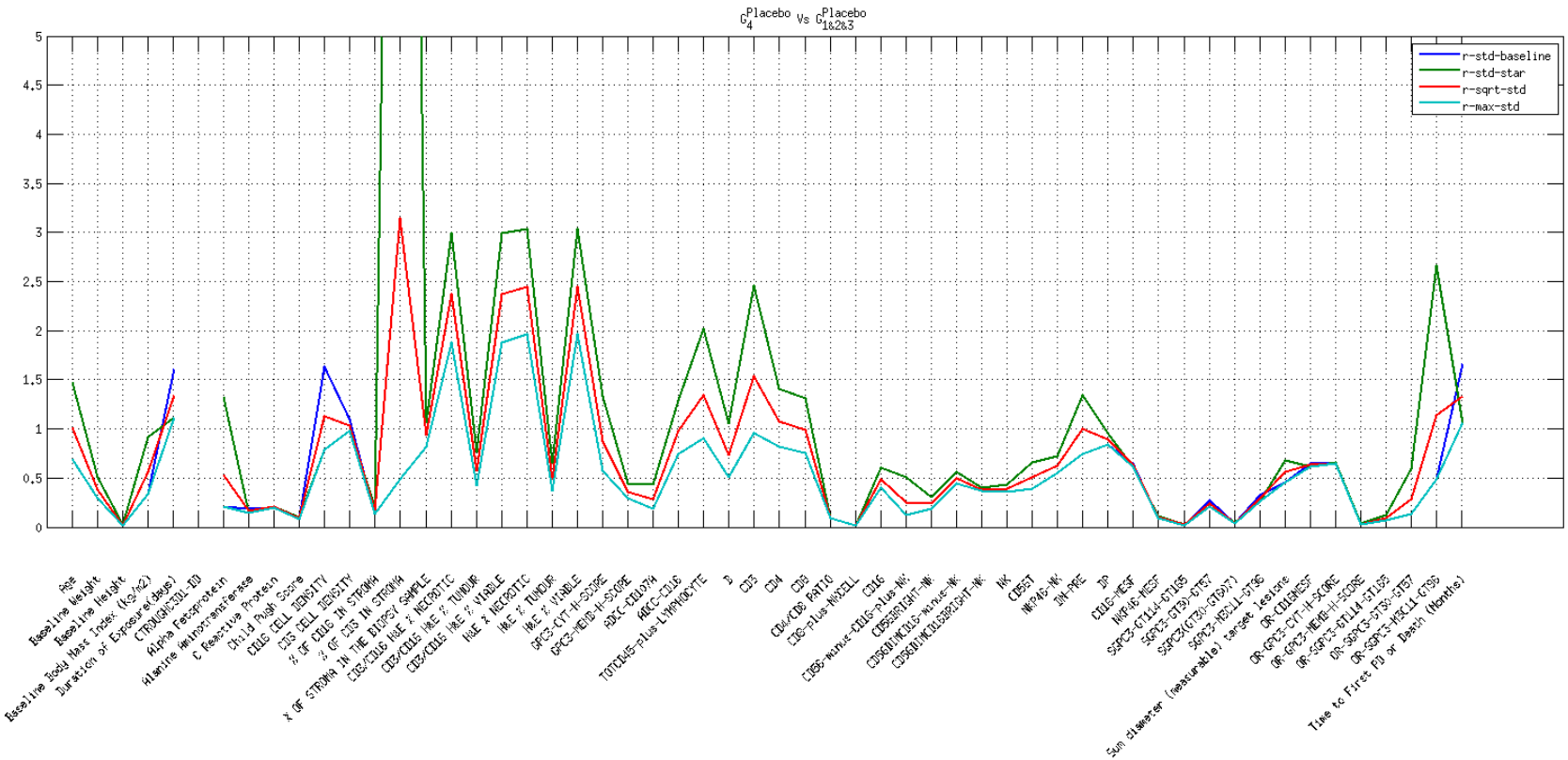
Measure of Significance
Categorical Variables

For categorical variables, compute:

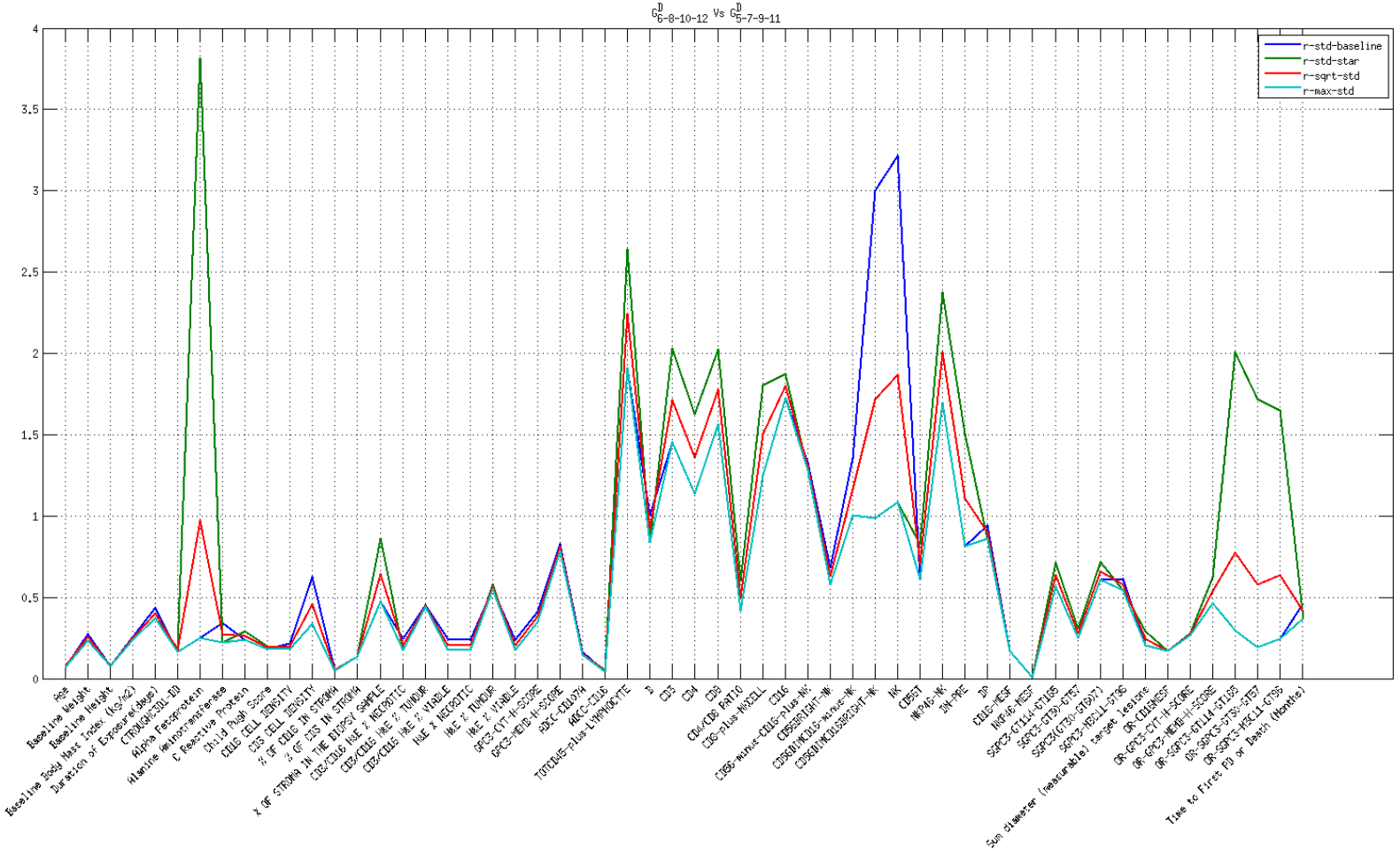
- Distance to Binomial Mean
  - Fit a Binomial distribution to G^B
  - A variable r is considered significant if $\mu_r(G^*)$ is outside confidence interval

- Fisher Exact Test: Standard statistical test for contingency tables.

Significance using mean deviation metrics: Placebo Patients



Significance using mean deviation metrics: Drug Patients



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. 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.
6. 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.