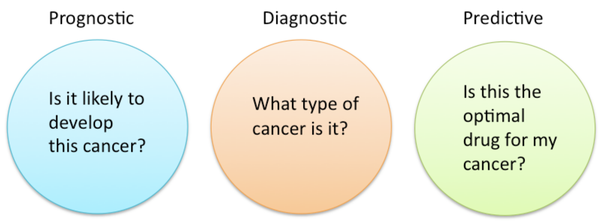
CASE-CONTROL INDIAN BUFFET PROCESS FOR BIOMARKER DISCOVERY IN CLINICAL TRIALS

Melanie F. Pradier[1], Fernando Perez-Cruz[1]
[1]Department of Signal Theory and Communications
Universidad Carlos III de Madrid,
{melanie,fernando}@tsc.uc3m.es

MOTIVATION

Biomarkers are used everywhere!!

- Prostate-specific antigen (PSA) to diagnose prostate cancer.
- Estrogen / progesterone to predict sensitivity to endocrine therapy in breast cancer.
- KRAS mutation to predict resistance to EGFr antibody treatment.



Cancer Drugs are ineffective for 75% of patient population (B. Spear et. al. *Clinical Trends in Molecular Medicine*, 2001).

OBJECTIVE

In a clinical trial scenario, we want to discover:

1. Indicators of disease progression: **prognostic** biomarkers
2. Indicators of (positive) drug response: **predictive** biomarkers
3. Actionable biomarkers as potential new targets for drugs

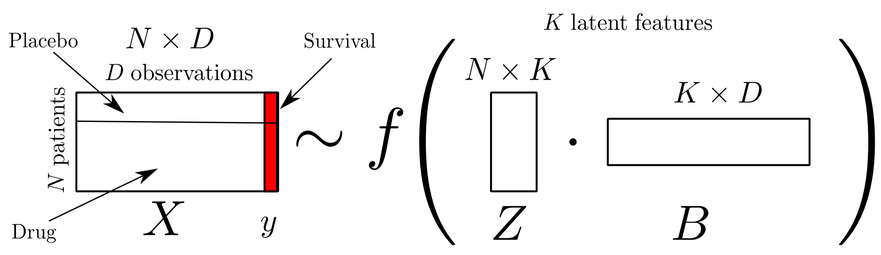
Challenges

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

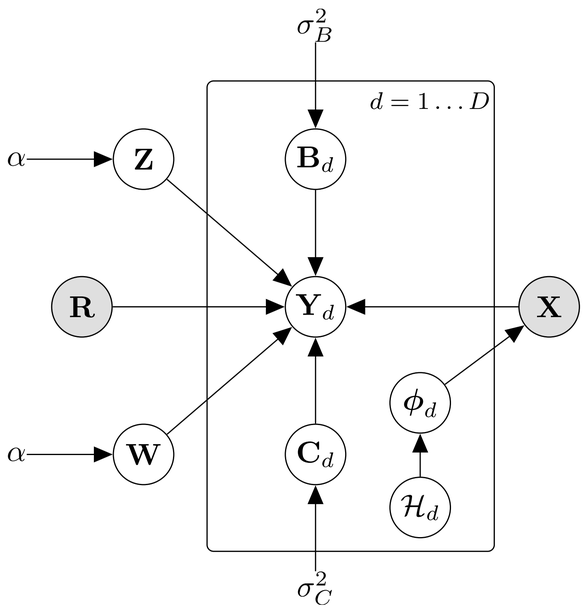
Solutions

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

CASE-CONTROL INDIAN BUFFET PROCESS (C-IBP)



- Model based on Generalized IBP for matrix completion (I. Valera, 2014)
- Let **Z** be a binary assignment matrix for **global** features
- Let **W** be a binary assignment matrix for **drug-specific** features



$$x_{nd}|y_{nd}, \phi_{\boldsymbol{d}} = T_d(y_{nd}, \phi_{\boldsymbol{d}})$$

$$y_{nd}|others \sim \mathcal{N}(z_{n\cdot}B_{\cdot d} + \mathbb{1}[\mathbf{R}_n = 1]\ w_{n\cdot}C_{\cdot d}, \sigma_y^2)$$

$$B_{kd} \sim \mathcal{N}(0, \sigma_B^2)$$

$$C_{kd} \sim \mathcal{N}(0, \sigma_C^2)$$

$$\mathbf{Z} \sim \text{IBP}(\alpha) \text{ and } \mathbf{W} \sim \text{IBP}(\alpha)$$

Nodes represent random variables, grey ones are observed. **X** is the matrix of observations and **R** is the drug indicator vector to distinguish placebo and drug patients.

INFERENCE

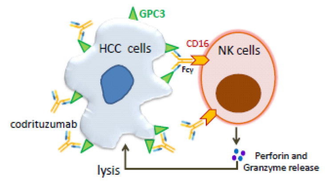
Here, let $\mathbf{Z^0}$, $\mathbf{Z^1}$ be the global feature assignment matrices for placebo and drug patients. We also distinguish the auxiliary variable matris $\mathbf{Y^0}$ and $\mathbf{Y^1}$ based on the drug indicator. Inference based on accelerated Gibbs sampling for the IBP (F. Doshi-Velez et.al. 2009).

1: **Initialize:** **Z**, **W**, and **Y**
2: **for** each iteration **do**
3: Sample $\mathbf{Z^0}$, $\mathbf{Y^0}$, and **B** given $\mathbf{X^0}$, using accelerated Gibbs sampling.
4: **for** $d = 1, \dots, D$ **do**
5: Sample $\mathbf{Z^1}$ given $\mathbf{Y^1}$, and **B** according to $p(z_{nk}^1 = 1|\mathbf{Z}_{-nk}, \mathbf{B}) \propto \frac{m_k}{N} \prod_{d=1}^{D} \mathcal{N}(y_{nd}|\sum_k z_{nk}^1 B_{kd}, \sigma_y^2)$
6: **end for**
7: Sample **W** given $\mathbf{Z^1}$ and $\mathbf{Y^1}$ using accelerated Gibbs sampling.
8: **for** $d = 1, \dots, D$ **do**
9: Sample $\mathbf{C}_d$ given **Z**, **W**, and $\mathbf{Y}_d$.
10: Sample $\mathbf{Y^1}_d$ given $\mathbf{X^1}$, **Z**, **W**, $\mathbf{B}_d$ and $\mathbf{C}_d$.
11: Sample $\phi_{\boldsymbol{d}}$ if needed. $\mathcal{H}_d$ are the hyperparameters over $\phi_{\boldsymbol{d}}$.
12: **end for**

DATABASE

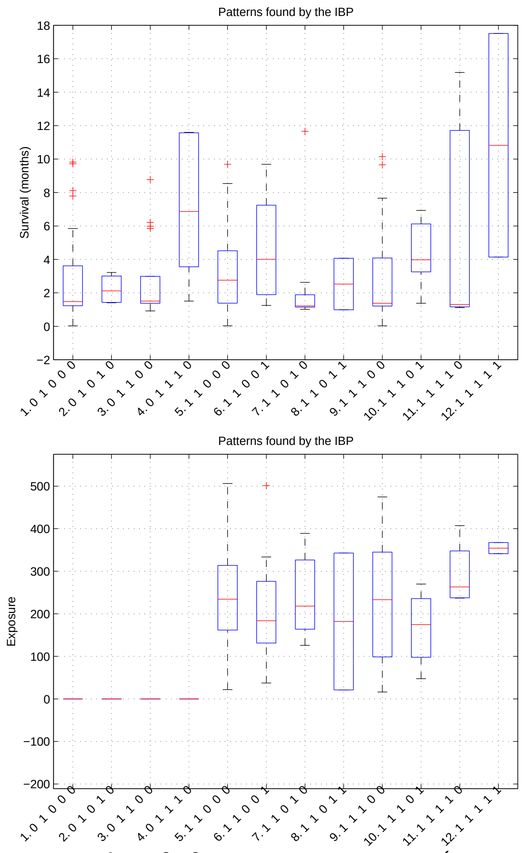
We demonstrate the utility of our approach in a **randomized phase II clinical trial** for the assessment of a cutting-edge immunotherapy treatment called Codrituzumab against **liver cancer** (J. Hepatology. 2016 Apr 13, Abou-Alfa et.al.)

- Clinical trial with 180 patients: 60 placebo, 120 drug.
- 80 obs. (demographics, clinical data, and survival).
- 48000 variables of RNA-seq data.



RESULTS

Sub-population	Drug Indicator	F1	F2	F3	Size (number of patients)	Mean TFPD (months)	Median TFPD (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



Robustness: we combine bootstrapping techniques with soft-feature assignments (posterior averages). We assess statistical significance using T-test, Fisher test, and Wald test. We also correct for multiple hypothesis testing using the Benjamini-Hochberg procedure.

CONCLUSIONS

We propose a **general method for biomarker discovery in clinical trials**. The C-IBP can identify both prognostic and predictive variables both global or specific to subpopulations.

1. **Complex correlations**: Captured by latent features.
2. **Population Heterogeneity**: Patients have different feature assignments.
3. **Natural Vs Drug Response**: Distinction via drug-specific features.

FUTURE WORK

- Release open-source package.
- Dependency based on drug exposure.
- Discriminative model towards survival.

FUNDING

This project has received funding from the European Union's Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861.

SIGNAL PROCESSING GROUP
DEPARTMENT OF SIGNAL THEORY AND COMMUNICATIONS
UNIVERSIDAD CARLOS III DE MADRID