Bayesian Poisson Factorization for Genetic Associations with Clinical Features in Cancer

M. F. Pradier[1], T. Karaletsos[3], S. Stark[3], J.E. Vogt[3], F. Perez-Cruz[1,2], and G. Rätsch[3]

1. University Carlos III in Madrid, Spain, 2. Bell Labs, Alcatel-Lucent, New Jersey, USA and 3. Memorial Sloan-Kettering Cancer Center, New York, USA

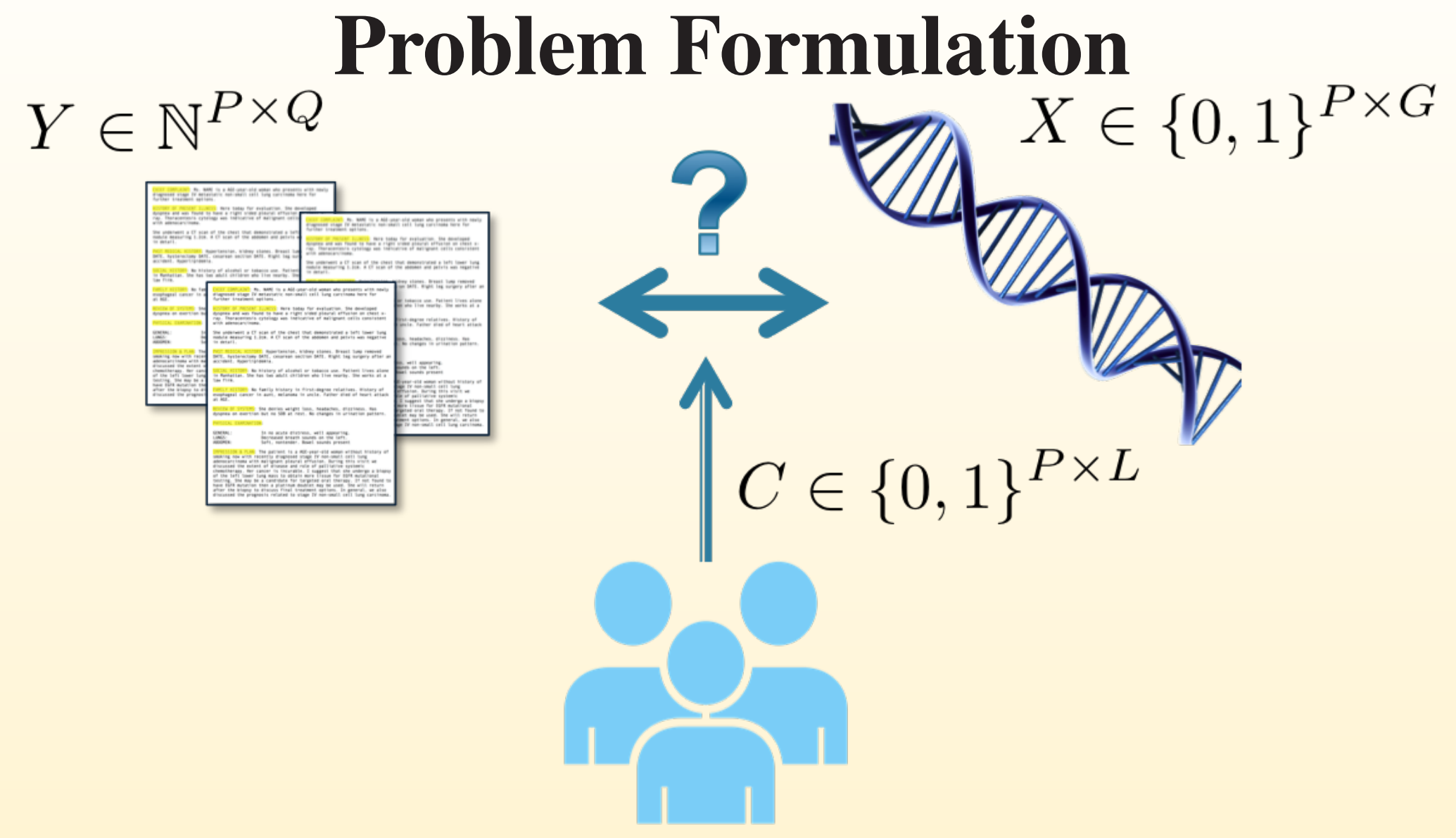
{melanie,fernando}@tsc.uc3m.es {karaletsos,starks,vogt,raetsch}@cbio.mskcc.org

Introduction

Motivation

- Cancer ≡ set of complex genetic diseases not very well understood yet.
- Our aim: Exploratory Analysis. Get meaninful genotype-phenotype associations by looking at Electronic Health Records (EHR) and genomic data.
- We would like our model to account for:
  - Confounders
  - Pleitropy
  - Epistasis

Problem Formulation


$Y \in \mathbb{N}^{P\times Q}$
$X \in \{0,1\}^{P\times G}$
?
$C \in \{0,1\}^{P\times L}$

Confounder-corrected Poisson Factorization Model

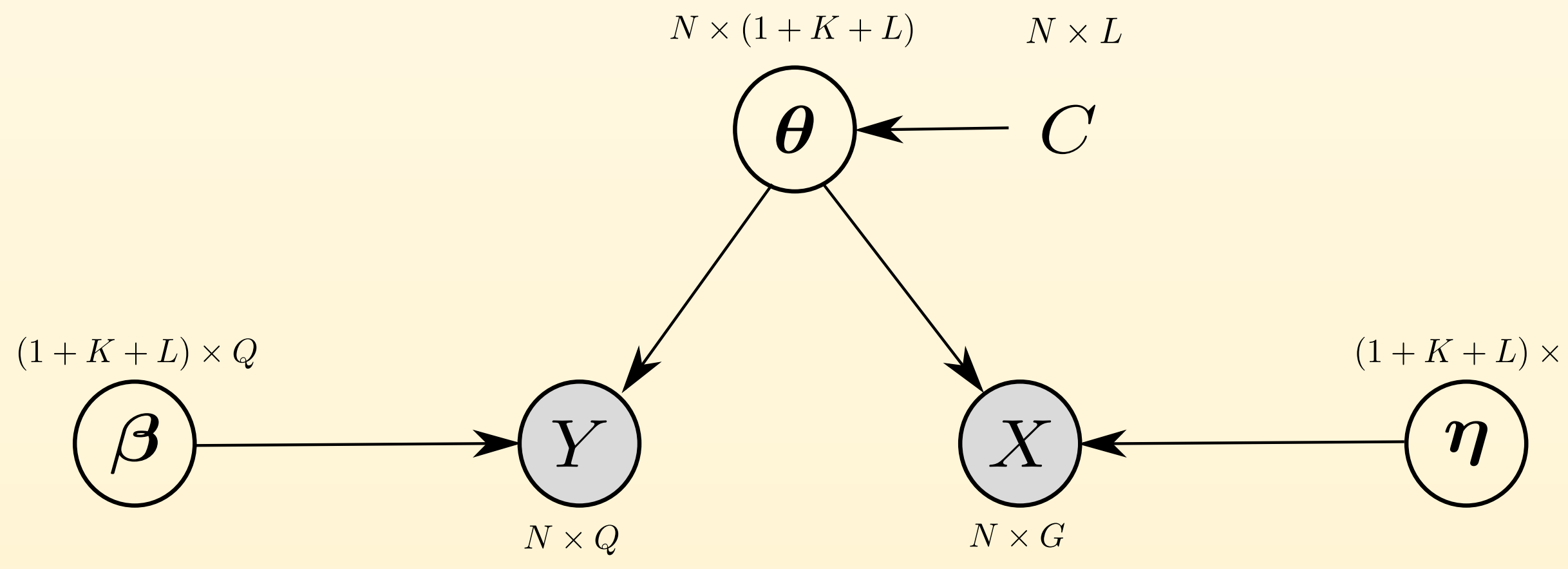
- We propose a model that directly finds associations between somatic mutations and clinical features through hidden factors. We call it Bayesian Confounder-corrected Poisson Factorization model (CC-PFM).
- Based on a PFM model for recommendation systems [1].
- Notation:
  - N patients, G mutations, Q words, L cancer types and K latent topics.
  - $\boldsymbol{\beta}$ and $\boldsymbol{\eta}$ are the clinical and genetic factors respectively.
  - $\boldsymbol{\theta}$ is the weight matrix: it captures the presence/activation of each factor per patient.

$$y_{nq}|\boldsymbol{\theta_{n.}},\boldsymbol{\beta_{.q}} \sim \text{Poisson}(\beta_{0q} + \boldsymbol{\theta}'_{\boldsymbol{n.}}\boldsymbol{\beta}'_{\boldsymbol{.q}} + \boldsymbol{\theta}''_{\boldsymbol{n.}}\boldsymbol{\beta}''_{\boldsymbol{.q}}) \tag{1}$$

$$x_{ng}|\boldsymbol{\theta_{n.}},\boldsymbol{\eta_{.g}} \sim \text{Poisson}(\eta_{0g} + \boldsymbol{\theta}'_{\boldsymbol{n.}}\boldsymbol{\eta}'_{\boldsymbol{.g}} + \boldsymbol{\theta}''_{\boldsymbol{n.}}\boldsymbol{\eta}''_{\boldsymbol{.g}}) \tag{2}$$

$$\theta_{nr} \sim \text{Gamma}(a,b), \quad \beta_{rq} \sim \text{Gamma}(c,d), \quad \eta_{rg} \sim \text{Gamma}(e,f), \tag{3}$$

where $\boldsymbol{\theta} = [\mathbf{1}_N,\, \boldsymbol{\theta}'_{\boldsymbol{N\times K}},\, \boldsymbol{\theta}''_{\boldsymbol{N\times L}} \odot C_{N\times L}]$ is an $N \times (1+K+L)$ matrix, $\mathbf{1}_N$ is a column vector of ones for the bias term, $\odot$ is the Hadamard product, $\boldsymbol{\beta} = [\boldsymbol{\beta}_0;\, \boldsymbol{\beta}'_{\boldsymbol{K\times Q}};\, \boldsymbol{\beta}''_{\boldsymbol{L\times Q}}]$ is a $(1+K+L) \times Q$ matrix, and $\boldsymbol{\eta} = [\boldsymbol{\eta}_0;\, \boldsymbol{\eta}'_{\boldsymbol{K\times G}};\, \boldsymbol{\eta}''_{\boldsymbol{L\times G}}]$ is a $(1+K+L) \times Q$ matrix of genetic factors.



We can also write the likelihood as:

$$y_{nq}|\boldsymbol{\theta_{n.}},\boldsymbol{\beta_{.q}} \sim \text{Poisson}(\boldsymbol{\theta_{n.}}\boldsymbol{\beta_{.q}}) \tag{4}$$

$$x_{ng}|\boldsymbol{\theta_{n.}},\boldsymbol{\eta_{.g}} \sim \text{Poisson}(\boldsymbol{\theta_{n.}}\boldsymbol{\eta_{.g}}) \tag{5}$$

- We use mean-field variational inference for learning.

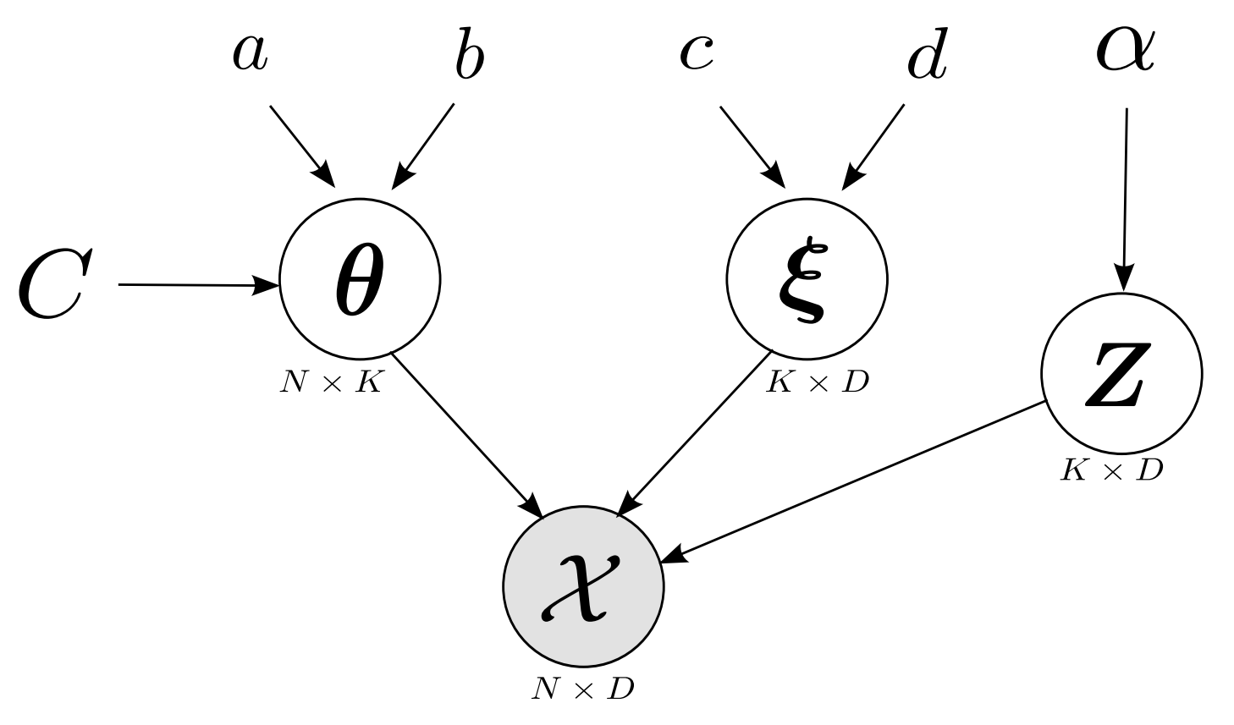
Non-parametric Sparse CC-PFM

- We introduce 2 binary matrices $\boldsymbol{Z_\beta}$ and $\boldsymbol{Z_\eta}$ that work as a mask on the factor matrices. In other words, we replace the Gamma priors by Spike-and-slab priors. The likelihoods are now:

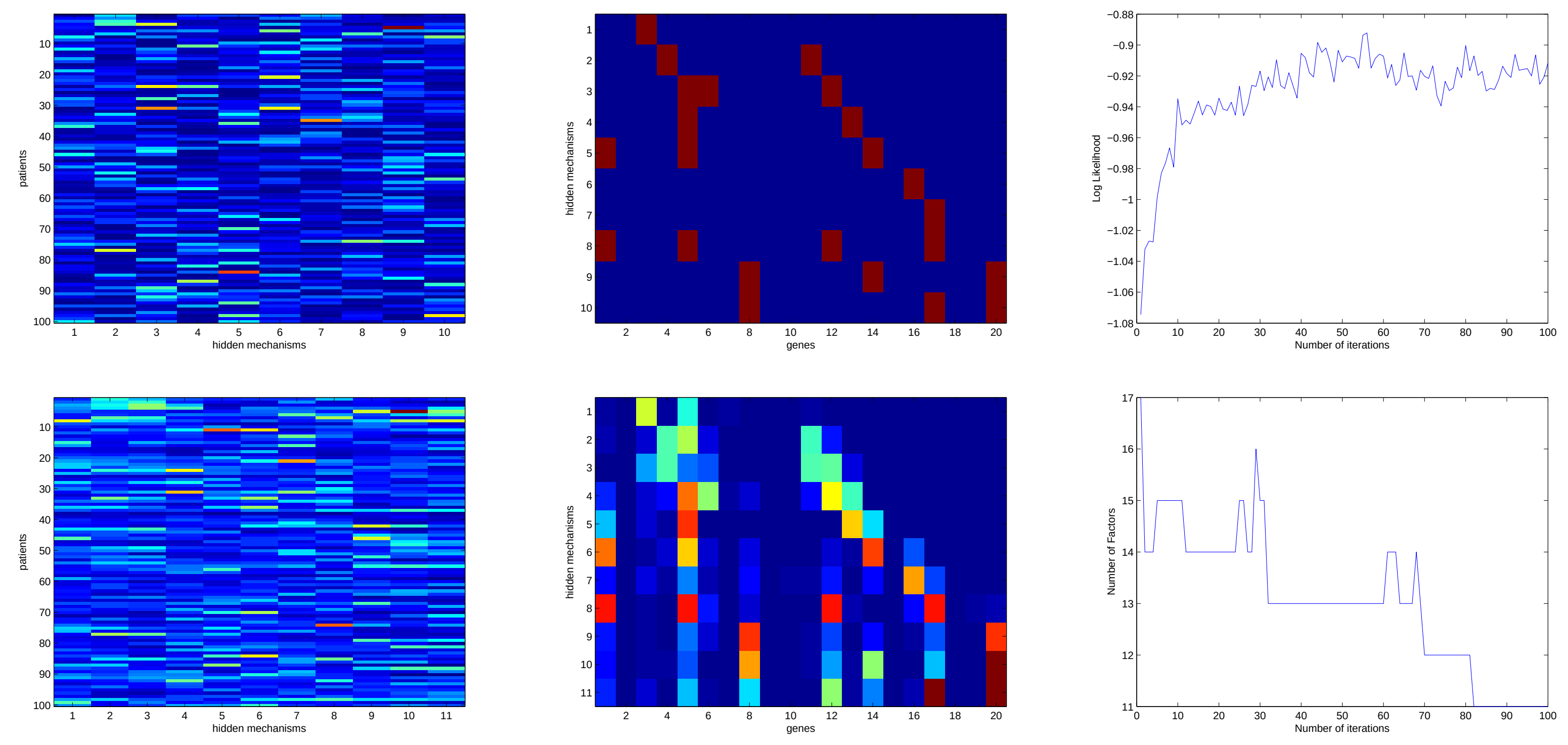
$$y_{nq}|\boldsymbol{\theta_{n.}},\boldsymbol{\beta_{.q}} \sim \text{Poisson}\big(\boldsymbol{\theta_{n.}}(\boldsymbol{Z_\beta} \odot \boldsymbol{\beta_{.q}})\big) \tag{6}$$

$$x_{ng}|\boldsymbol{\theta_{n.}},\boldsymbol{\eta_{.g}} \sim \text{Poisson}\big(\boldsymbol{\theta_{n.}}(\boldsymbol{Z_\eta} \odot \boldsymbol{\eta_{.g}})\big) \tag{7}$$

- We use an Indian Buffet Process prior on the matrices $\boldsymbol{Z_\beta}$ and $\boldsymbol{Z_\eta}$ to make $K \to \infty$. That is, $\boldsymbol{Z} \sim \text{IBP}(\alpha)$ where $\boldsymbol{Z} = [\boldsymbol{Z_\eta},\ \boldsymbol{Z_\beta}]^{\mathrm{T}}$. This model extends the model in [2].
- Simpler notation: $\boldsymbol{\xi} = [\boldsymbol{\beta},\, \boldsymbol{\eta}]$ and $\mathcal{X} = [\boldsymbol{Y},\, \boldsymbol{X}]$.
- Inference: Gibbs sampling + slice sampling for matrix $\boldsymbol{Z}$ [3].



The Idea: We show $\boldsymbol{\theta}$ and $\boldsymbol{Z}$, ground truth (top) and inferred (bottom).



Results

		Textual topics $\boldsymbol{\beta}'_{\boldsymbol{k.}}$	Genetic topics $\boldsymbol{\eta}'_{\boldsymbol{k.}}$
Free Associations	Factor 0 (Bias)	demonstrated, oncologist, suv, died, involvement	TP53
	Factor 1	adenocarcinoma, pleural, woman, smoker	PIK3CA, RB1
	Factor 2	pelvic, female, woman, endometrial, vaginal	NRAS
	Factor 3	cisplatin, squamous, icterus, kg, exertion	SPEN
	Factor 4	m, icterus, colon, fluid, ascites, cavity, hepatomegaly	KRAS
	Factor 5	folfox, colorectal, anc, colon, oxaliplatin	APC, KRAS, CIC
	Factor 6	brain, hemangiopericytoma, female, parietal	FUBP1, AXL
	Factor 7	breast, woman, adjuvant, female, mastectomy	PIK3CA, CDH1
		Textual topics $\boldsymbol{\beta}''_{\boldsymbol{l.}}$	Genetic topics $\boldsymbol{\eta}''_{\boldsymbol{l.}}$
Cancer Specific	Appendiceal	mucinous, debulking, intraperitoneal, appendectomy	KRAS, GNAS
	Bladder	bladder, urothelial, gemcitabine, invasive, cisplatin	TERT, KDM6A
	Breast Carcinoma	breast, mastectomy, invasive, husband, female	PIK3CA, GATA3
	Melanoma	melanoma, ipilimumab, database, toe, temozolomide	MYCN, RAD51
	Small Cell Lung	etoposide, smoker, cisplatin, reassessment, irinotecan	RB1
	Soft Tissue	sarcoma, gentleman, adjuvant, ifosfamide, c, adriamycin	MYOD1

Future work

- *Sparsity*: IBP ties the total number and per-row number of hidden features. This property is undesired a priori, and should be relaxed.
- *Validation*: What is the statistical significance of the inferred factors? Direct testing using statistical tests for stratified categorical data [4].
- *Scalability*: Inference using Stochastic Variational Inference.
- *Flexibility*: Introduce better conditional dependence on the confounders.

Funding

This project has received funding from the European Union Seventh Framework Programme for research, technological development and demonstration under grant agreement no. 316861.
This work is based on unpublished MSKCC data.

European Commission

Machine Learning for Personalized Medicine

Basic References

[1] P. Gopalan and D. Blei, ”Content-based recommendations with Poisson factorization”, presented at the Advances in Neural Information Processing Systems 27, 2014.

[2] S. K. Gupta, D. Phung, and S. Venkatesh, “A nonparametric Bayesian Poisson gamma model for count data,” in 2012 21st International Conference on Pattern Recognition (ICPR), 2012, pp. 1815–1818.

[3] Y. W. Teh, “Stick-breaking construction for the Indian buffet process,” in In Proceedings of the International Conference on Artificial Intelligence and Statistics, p. 2007.

[4] K. Fukumizu, A. Gretton, X. Sun, and B. Schölkopf, “Kernel measures of conditional dependence,” in In Adv. NIPS, 2008.