Probabilistic Analysis of Genetic Associations with Clinical Features in Cancer

Melanie F. Pradier[1,2], Julia E. Vogt[2], Stefan Stark[2],
Theofanis Karaletsos[2], Fernando Perez-Cruz[1] and Gunnar Rätsch[2]
1. University Carlos III in Madrid, Spain and 2. Memorial Sloan-Kettering Cancer Center, New York, USA
Email to: melanie@tsc.uc3m.es

Motivation

Objective: Finding meaningful relationships between genetic information and clinical features.

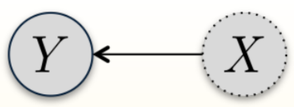
Why does it matter?

- Improved Diagnosis
- Risk Identification
- Biological Insight

Our models should be able to deal with...

- Epistasis
- Pleiotropy
- **Confounders**

Linear Mixed Model (State-of-the-art)



$$y = X\beta + u + \epsilon \quad (1)$$

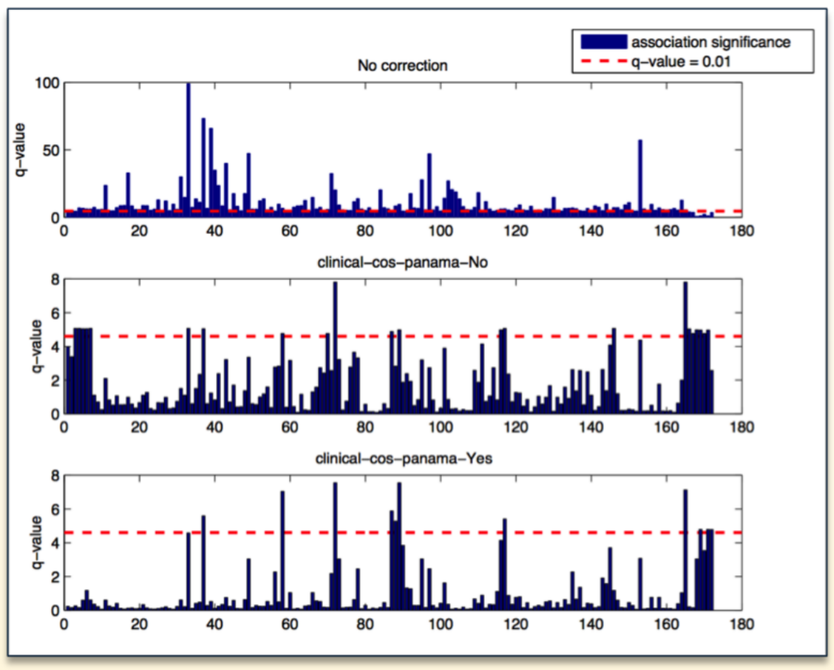
where

$$u \sim \mathcal{N}(0, K) \quad (2)$$
$$\epsilon \sim \mathcal{N}(0, \sigma_e^2 I) \quad (3)$$

- medical observation $y \in \mathbb{R}^{P\times 1}$
- genetic variation matrix $X \in \{0,1\}^{P\times G}$
- fixed effects $\beta \in \mathbb{R}^{G\times 1}$, random effects $u \in \mathbb{R}^{P\times 1}$
- K: covariance matrix (e.g. $K = CC^{\mathrm{T}}$)

- Classical model in statistical genetics [1]
- Rank Transformation to make Y observations Gaussian
- Inference of hidden confounders W using PANAMA [2], resulting in $y = X\beta + u + \epsilon + W\gamma$

Results LMM - Most Conservative



10 significant hits with clinical covariates, 165 without correction. Examples:

Gene	MAF	q-value	β	Sentence prototype
APC	112	0.0037	0.33	He underwent a colonoscopy which revealed a pedunculated polyp in the ascending colon.
ALK	40	0.0063	0.57	The patient showed a mild decrease in her blood counts.
HNF1A	13	0.0028	0.70	The patient was tearful presented with depressed affect and mood.
TRAF7	11	0.0008	0.59	He has a history of adenoid cystic carcinoma of the salivary gland.

Conclusion and Future Work

- LMM is easy to interpret and gives False Discovery Rate estimate
- PFM can deal with epistasis naturally and finds better clinical feature representation
- Sparsity of topics
- Conditional model of θ given C
- Non-parametric extension

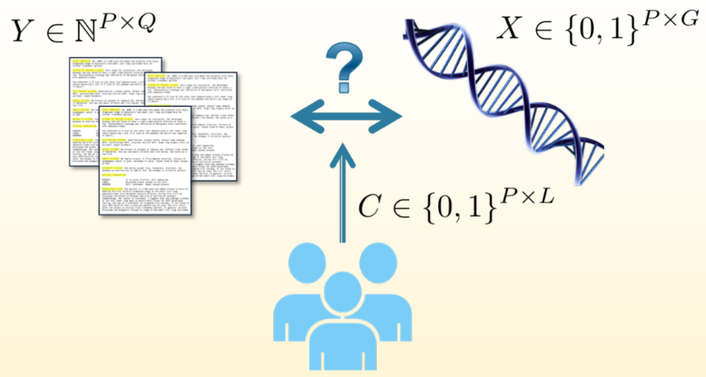
Funding

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

Problem Setting

Notation

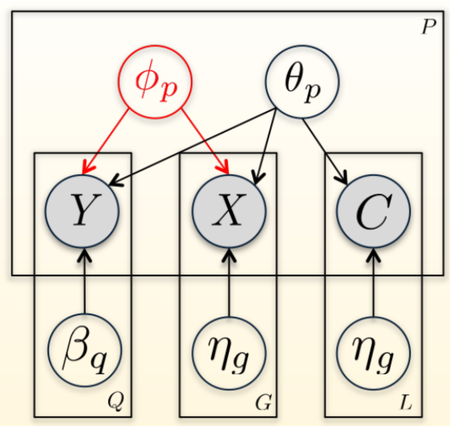
- P patients
- G genes
- Q features
- L covariates



Features are medical observations extracted from the Medical Health Records. This work considers sentence clusters and word topics.

Poisson Factorization Model

The idea is to build a joint generative model of the clinical text and genetic information.

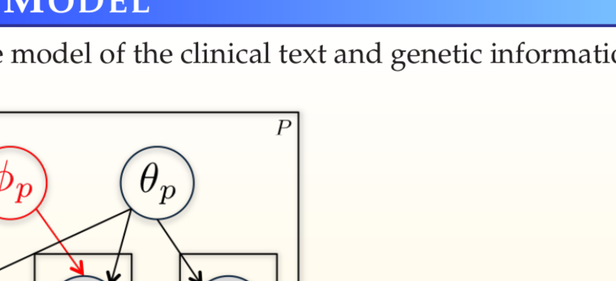


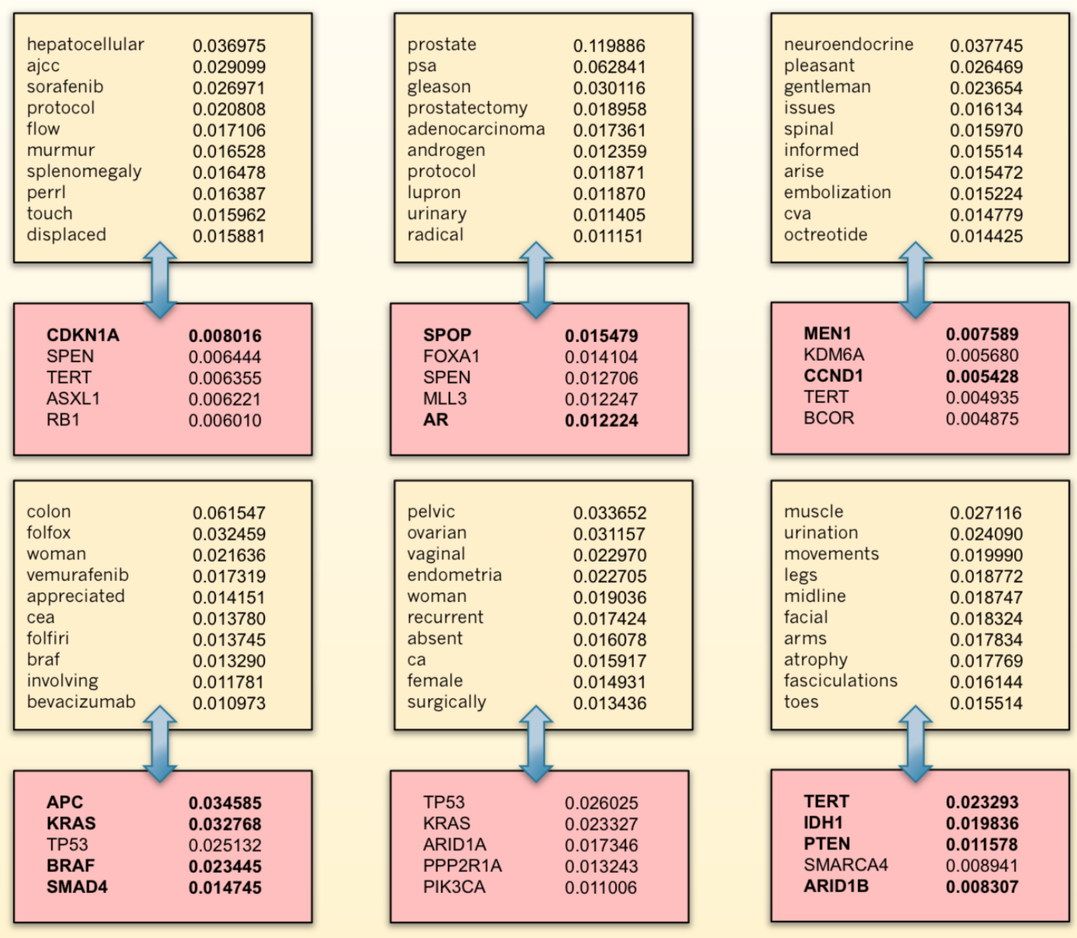
- Medical factors β_q
- Genetic factors η_g
- New interesting associations captured by ϕ, confounders by θ
- Based on [3], variational inference

$$x_{pg} \sim \text{Poisson}(\phi_p \eta_g' + \theta_p \eta_g) \quad (4)$$
$$y_{pq} \sim \text{Poisson}(\phi_p \beta_q' + \theta_p \beta_q) \quad (5)$$
$$c_{pl} \sim \text{Poisson}(\theta_p \zeta_q) \quad (6)$$
$$\text{rest} \sim \text{Gamma}(a, b) \quad (7)$$

Results PFM - Preliminary

- $P = 1265$ patients, $\sim 15k$ documents, G =343 genes, K =20 factors, $Q = 1k$ words out of $30k$ selected with highest it-idf index.
- Associations in grey generated by θ, in color by ϕ.
- Associations with genes in bold have a q-value $< 5.10^{-4}$.





- A False Discovery Rate estimate is obtained by plugging the inferred word topics to the LMM as new input matrix Y.

Basic References

1 C. Lippert, F. P. Casale, B. Rakitsch, and O. Stegle, "LIMIX: genetic analysis of multiple traits", bioRxiv, 2014.

2 N. Fusi, O. Stegle, and N. D. Lawrence, "Joint Modelling of Confounding Factors and Prominent Genetic Regulators Provides Increased Accuracy in Genetical Genomics Studies", PLoS Comput Biol, vol. 8, no. 1, p. e1002330, Jan. 2012.

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