Bayesian Poisson Factorization for Genetic Associations with Clinical Features in Cancer

M. F. Pradier, F. Perez-Cruz[*]
Department of Signal Theory
and Communications
Univ. Carlos III in Madrid, Spain
melanie@tsc.uc3m.es

T. Karaletsos, S. Stark, J.E. Vogt, G. Rätsch
Department of Computational Biology
Memorial Sloan-Kettering
Cancer Center, New York, USA
karaletsos@cbio.mskcc.org

Cancer encompasses not one, but a vast group of genetic diseases that are not very well understood yet. In the last decade, high-throughput genotyping technologies have led to the discovery of cancer-correlated gene mutations, most of which were not previously suspected to be related to carcinogenesis [1]. However, only the gene mutations with very strong effects have been discovered and many other genes with weaker effects still remain to be found [2]. Genetic-association studies have been widely used in the search for such genes, but success has been limited so far [3].

A first difficulty in cancer association studies is the immense phenotypic heterogeneity, which reduces statistical power in the discovery method and causes some associations to remain hidden [4, 5, 6]. Second, cohort sizes especially in rare cancers tend to be small, which makes the discovery of small effect size associations difficult [2]. Third, cancers are driven by the accumulation of mutations that may act epistatically or pleiotropically during the course of the disease [7, 8, 9]. Epistasis refers to complex interactions between genetic variants that have an effect on the same phenotype, while pleiotropy means that multiple phenotypes are influenced by the same single mutation. New approaches need to be found in order to overcome these difficulties.

In recent years, the adoption of Electronic Health Records (EHR) in hospitals has increased dramatically, and has become an interesting resource for phenotyping [10, 11]. Although structured data in EHR is very valuable, most of the clinical data, e.g., around 98% of the EHRs, comes as unstructured notes [12]. These include a broad spectrum of clinically-relevant phenotypic information, and might be useful to identify new phenotypes, still unclassified in ontologies [11, 12]. In this work, we use a a bag-of-words representation, e.g., we count how many times each word has ever been written in the medical history of each patient. Stop words and very frequent ones are removed in a pre-processing step, in which we also compute the tf-idf score for each word and only keep the top 1000 words with highest score. From these word counts, we build a generative model that extracts sparse topics associated to groups of genetic mutations.

This work presents a Bayesian Poisson Factorization model to find associations between somatic mutations and clinical features in cancer that deals with phenotype heterogeneity, small cohort size, epistasis and pleiotropy in a straightforward way. Our method is a generative model that infers latent topics $\boldsymbol{\beta_{k\cdot}}$ from the text, associated to genetic factors $\boldsymbol{\eta_{k\cdot}}$ that directly capture complex genetic interactions. Given N patients, G mutations, Q words, L cancer types and K latent topics, the basic generative model can be written as:

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

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

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

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. With such

[*] Also Machine Learning Scientist at Bell Labs.

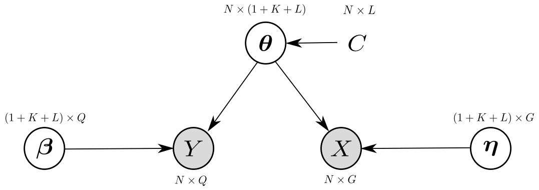

Figure 1: **Basic Poisson Factorization Model for genetic associations with clinical features.** X is the somatic mutation matrix, Y is the bag-of-words matrix, C is the cancer type indicator matrix. Each patient belongs to a single cancer type, so C is a binary matrix with a single one per row.

notation, the model likelihood can also be written as:

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

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

Our model is directly inspired on the Collaborative-Topic Poisson Factorization model for recommendation systems in [13], with three important modifications. First, we introduce confounding effects as conditional variables, as shown in Figure 1. In particular, we add cancer type specific hidden factors to analyse multiple cancers jointly. Indeed, most cancers are known to share a common pathogenesis despite specificities of the cell type and tissue origin [14]. By doing so, we avoid getting already well-known per-cancer mutations, and are able to obtain less well-known associations with gene mutations of smaller effect size. Second, we force sparsity on the textual and genetic topics by putting an hyperprior on the shape parameter of the Gamma distribution priors. Third, we present a non-parametric extension to the work in [13] by replacing $\boldsymbol{\theta}^{'}$ with an Indian Buffet Process (which also helps in terms of sparsity), and compare our extension to the approach followed in [15].

Bayesian modeling has already been proven useful for epistasis [16], pleiotropy [16, 17] or subphenotyping [18, 19] aplications. Our model combines ideas of these previous contributions and is the first one to deal with clinical text data and genetic information, capturing phenotypic heterogeneity, epistasis and pleiotropy in a straightforward way while correcting for possible confounders. Table 1 shows a partial list of the associations found by our model. These can be validated using statistical tests for stratified categorical data [20, 21, 22]. We hope that such associations might be useful for further research in oncology. Studies like these can inform us about actionable pathways when considering cancer therapy, where interventions through drug administration can be designed, and ultimately help with accurate diagnosis.

		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

Table 1: **Associations found using a Poisson Factorization Model with $K = 25$.** We only report topics associated to at least one mutation. 9 out of 25 *free* topics did not have any mutation associated. A genetic mutation is reported if its probability is above 10% in the topic. We show the ordered top most frequent words for each textual topic.

References

[1] D. F. Easton and R. A. Eeles, *Human Molecular Genetics* **17**, R109 (October 2008).

[2] U. Andersson, R. McKean-Cowdin, U. Hjalmars and B. Malmer, *Acta Oncologica (Stockholm, Sweden)* **48**, 948 (2009).

[3] P. D. P. Pharoah, A. M. Dunning, B. A. J. Ponder and D. F. Easton, *Nature Reviews Cancer* **4**, 850 (November 2004).

[4] E. Quintana, M. Shackleton, H. R. Foster, D. R. Fullen, M. S. Sabel, T. M. Johnson and S. J. Morrison, *Cancer Cell* **18**, 510 (November 2010).

[5] E. Lalonde, A. S. Ishkanian, J. Sykes, M. Fraser, H. Ross-Adams, N. Erho, M. J. Dunning, S. Halim, A. D. Lamb, N. C. Moon, G. Zafarana, A. Y. Warren, X. Meng, J. Thoms, M. R. Grzadkowski, A. Berlin, C. L. Have, V. R. Ramnarine, C. Q. Yao, C. A. Malloff, L. L. Lam, H. Xie, N. J. Harding, D. Y. F. Mak, K. C. Chu, L. C. Chong, D. H. Sendorek, C. P'ng, C. C. Collins, J. A. Squire, I. Jurisica, C. Cooper, R. Eeles, M. Pintilie, A. Dal Pra, E. Davicioni, W. L. Lam, M. Milosevic, D. E. Neal, T. van der Kwast, P. C. Boutros and R. G. Bristow, *The Lancet Oncology* **15**, 1521 (December 2014).

[6] M. D. Ritchie, L. W. Hahn, N. Roodi, L. R. Bailey, W. D. Dupont, F. F. Parl and J. H. Moore, *American Journal of Human Genetics* **69**, 138 (July 2001).

[7] X. Wang, A. Q. Fu, M. E. McNerney and K. P. White, *Nature Communications* **5**, p. 4828 (November 2014).

[8] A. D. N. J. de Grey, *Mechanisms of Ageing and Development* **128**, 456 (August 2007).

[9] L. C. Sakoda, E. Jorgenson and J. S. Witte, *Nature Genetics* **45**, 345 (April 2013).

[10] T. Adamusiak and M. Shimoyama, *AMIA Summits on Translational Science Proceedings* **2014**, 9 (April 2014).

[11] S. M. Meystre, G. K. Savova, K. C. Kipper-Schuler, J. F. Hurdle and others, *Yearb Med Inform* **35**, 128 (2008).

[12] P. B. Jensen, L. J. Jensen and S. Brunak, *Nature Reviews. Genetics* **13**, 395 (June 2012).

[13] P. K. Gopalan, L. Charlin and D. Blei, Content-based recommendations with Poisson factorization, in *Advances in Neural Information Processing Systems 27*, eds. Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence and K. Q. Weinberger (Curran Associates, Inc., 2014) pp. 3176–3184.

[14] M. R. Stratton, P. J. Campbell and P. A. Futreal, *Nature* **458**, 719 (April 2009).

[15] P. Gopalan, F. J. Ruiz, R. Ranganath and D. M. Blei, *Artificial Intelligence and Statistics (AISTATS)* **33**, 275 (2014).

[16] Y. Zhang and J. S. Liu, *Nature Genetics* **39**, 1167 (September 2007).

[17] W. Zhang, J. Zhu, E. E. Schadt and J. S. Liu, *PLoS Computational Biology* **6** (January 2010).

[18] L. Parts, O. Stegle, J. Winn and R. Durbin, *PLoS Genetics* **7**, p. e1001276 (January 2011).

[19] D. Knowles and Z. Ghahramani, *The Annals of Applied Statistics* **5**, 1534 (June 2011).

[20] J. Wittes and S. Wallenstein, *Journal of the American Statistical Association* **82**, 1104 (December 1987).

[21] K. Fukumizu, A. Gretton, X. Sun and B. Schlkopf, Kernel measures of conditional dependence, in *In Adv. NIPS*, 2008.

[22] G. Doran, K. Muandet, K. Zhang and B. Schlkopf, A Permutation-Based Kernel Conditional Independence Test, in *Proceedings of the 30th Conference on Uncertainty in Artificial Intelligence (UAI2014)*, 2014.