Power Law Models in Somatic Mutation Profiles

Rahul Mehta, Hui Lu
Department of Bioengineering
University of Illinois at Chicago
mehta5@uic.edu, huilu@uic.edu

Abstract

Like many natural phenomena, genetic mutations that cause cancer have a very heavy tail - the frequency of a few driver mutations is very high as compared to the majority. As a result, infrequent driver mutations make it difficult to identify the distinct mutated genetic pathways that cause cancer progression. Despite the success of computational methods to understand the underlying structure of somatic mutations based on stochastic processes, frequency counts, or factor analysis, they have limitations capturing datasets with power law behavior. Generalizations such as the Pittman-Yor process or placing sparse priors on factor analysis have been used to model power law behavior, but they are limited by scale and computational tractability. In this paper we extended the use of completely random measures based on the gamma process and stable processes to a variational autoencoder (VAE) framework that can incorporate massive datasets, while remaining computational tractable. We outline the generative mixture model that is used as a prior for the VAE to illustrate the utility for modeling power law behavior in somatic mutations and create more interpretable mutation profiles.

1 Introduction

Somatic mutation profiling often leverages expert knowledge in terms of annotations of functional consequences (e.g. AnnoVar and ENSEMBL VEP annotations) and biological knowledge of pathways (e.g. KEGG and GO annotations) to create a list of driver mutations that induce carcinogenesis. Distinguishing between driver mutations and passenger mutations, mutated genes that are not directly related with tumor genesis, is crucial for therapeutic decisions as well as the basic understanding of carcinogenesis. Moreover, driver mutations reflect changes in distinct pathways, such that the same pathway can be affected by various driver mutations [Sjöblom *et al.*, 2006]. As a result, research has been divided into finding individual mutated genes or pathways, our model focuses on the latter.

There are a variety of models to discover mutated pathways, such as linear programming, sparse factor analysis, or mixture models based on stochastic processes, like the beta-Bernoulli process. The primary goal of these models, and our model, is to create a diverse set of partitions that accurately render the relationship between mutations. Despite the simple generative construct of mixture models and factor analysis, the postulated distributions are often inappropriate for modeling real data, in that they fall into the scheme of the "rich get richer". Our goal here is to complement these approaches to scale with power law behavior to identify the affected latent pathways.

To account for sparsity in real data, power laws for latent models have yielded algorithms that use the Pittman-Yor process with its corresponding inference scheme. Unfortunately, with the large size of genetic datasets, current power law models do not scale computationally. We demonstrate a generative framework using normalized random measure mixture models (NRMM) with a variational inference scheme that incorporates the scalability of variational autoencoders [Kingma and Welling, 2013]. The model is applied to a mutation dataset based on lung adenocarcinoma, with preliminary results suggesting scaling between mutations and latent pathways follows a power law model.

2 Methods

We first give a general description of our model and how it incorporates a completely random measure. We then show the inference model using a variational autoencoder.

2.1 Definitions

Let $(\Omega, \Sigma, \mathbf{P})$ be a probability space and u on (Ω, Σ) is a random measure values element such that $u(A)$ is a random variable for any measure set $A \in \Sigma$. A completely random measure (CRM) is a random measure with additional requirement such that for any finite disjoint measurable sets $A_1 ... A_n \in \Sigma$ the random variables $u(A_1) ... u(A_n)$ are independent [Kingman, 1967]. We can represent a CRM using the Lévy-Khintchine formula of its Laplace functional transform as follows

$$\mathbb{E}[e^{\int g(y)u(dy)}] = e^{-\phi(s)} = exp(-\int (1 - e^{-sg(y)})p(ds)H_0(dy)) \quad (1)$$

where $p(ds)$ is the Lévy measure, g is a positive measurable function on $\mathcal{Y}$ and H_0 is the base distribution. A nor-

malized random measure can then be represented as $P(\cdot) = u(\cdot)/u(\mathfrak{Y})$, where $u(\mathfrak{Y})$ is the total mass that must be finite and almost positive surely. This is satisfied when $p(ds) = \infty$ and the Laplace transform is finite for any positive s. Two popular Lévy measures that incorporate power law are the generalized gamma process (GGP) and the stable beta processes (SBP), however, they rely on variants of MCMC, which have difficulty scaling to large datasets. Recent research has shown the little known Bertoin-Fujita-Roynette-Yor, $BFRY(\alpha)$, distribution has a simple density, $f(s) = \frac{Gamma(1-\alpha,1)}{Beta(\alpha,1)}$ exhibiting power law behavior [Bertoin *et al.*, 2006] which can scale to the GGP and SBP by incorporating exponential tilting, thus allowing for simpler simulation [Lee *et al.*, 2016].

In our model we use the properties of the BFRY distribution to create a normalized random measure (NRM) and then use the techniques described in [James *et al.*, 2009] to create the joint distribution of a NRMM (2).

$$p(x, \theta, u) = \frac{u^{N-1}}{\Gamma(N)} e^{-\phi(u)} du \prod_{k=1}^{K} \kappa_{|\theta_k|} H_0(d\theta) \qquad (2)$$

where κ_m denotes the mth moment of the Lévy measure, θ is the latent variable, and an auxiliary variable $U \sim Gamma(N, \theta)$ is introduced for tractability. For inference we use a VAE and must optimize the ELBO with respect to the parameters of the variational distribution $E_q[\log p(x, \theta, u)] - E_q[\log q(\theta, u)]$. Since the parameters of the variational distribution are not differentiable as required for a VAE, we use the reparameterization trick using an acceptance-rejection algorithm as in [Naesseth *et al.*, 2017], where $\theta = h(\epsilon, \alpha)$ and h is an estimation of the ratio of the gamma and beta distributions parameterized by α.

3 Experiments

We demonstrate the ability of BFRY mixture models as a prior for VAE to learn meaningful latent somatic pathways by using the lung dataset available on TCGA [Network and others, 2014]. We truncated the number of latent pathways to 700, set the alpha parameter to 0.1, and removed any outlier mutations resulting in 100 distinct mutations within 300 patients. In Figure 1 we show preliminary results that indicate the model produces power law behavior in the somatic mutation profile in comparison to the mixture models based on Gaussian or Dirichlet processes.

Acknowledgments

We gratefully acknowledge the support of NVIDIA Corporation with the donation of the Titan X Pascal GPU used for this research.

References

[Bertoin *et al.*, 2006] J Bertoin, T Fujita, Bernard Roynette, and Marc Yor. On a particular class of self-decomposable random variables: the durations of bessel excursions straddling independent exponential times. 2006.

[James *et al.*, 2009] Lancelot F James, Antonio Lijoi, and Igor Prünster. Posterior analysis for normalized random measures with independent increments. *Scandinavian Journal of Statistics*, 36(1):76–97, 2009.

[Kingma and Welling, 2013] Diederik P Kingma and Max Welling. Auto-encoding variational bayes. *arXiv preprint arXiv:1312.6114*, 2013.

[Kingman, 1967] John Kingman. Completely random measures. *Pacific Journal of Mathematics*, 21(1):59–78, 1967.

[Lee *et al.*, 2016] Juho Lee, Lancelot F James, and Seungjin Choi. Finite-dimensional bfry priors and variational bayesian inference for power law models. In *Advances in Neural Information Processing Systems*, pages 3162–3170, 2016.

[Naesseth *et al.*, 2017] Christian Naesseth, Francisco Ruiz, Scott Linderman, and David Blei. Reparameterization gradients through acceptance-rejection sampling algorithms. In *Artificial Intelligence and Statistics*, pages 489–498, 2017.

[Network and others, 2014] Cancer Genome Atlas Research Network et al. Comprehensive molecular profiling of lung adenocarcinoma. *Nature*, 511(7511):543, 2014.

[Sjöblom *et al.*, 2006] Tobias Sjöblom, Siân Jones, Laura D Wood, D Williams Parsons, Jimmy Lin, Thomas D Barber, Diana Mandelker, Rebecca J Leary, Janine Ptak, Natalie Silliman, et al. The consensus coding sequences of human breast and colorectal cancers. *science*, 314(5797):268–274, 2006.

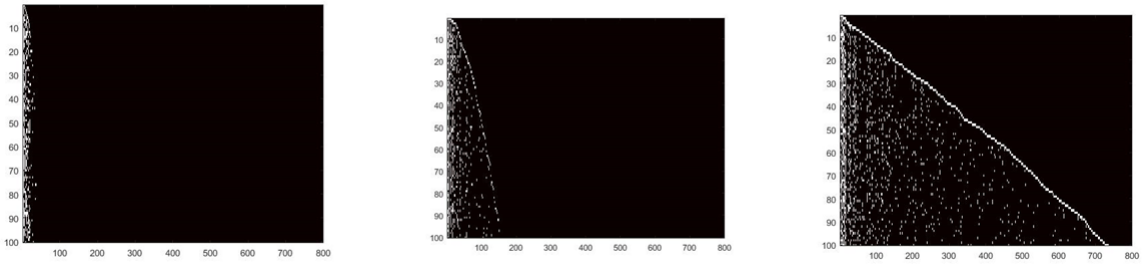

Figure 1: Gaussian Mixture Model, Dirichlet Process Mixture Model, BFRY-Mixture Model prior in VAE respectively from left to right. We can see that the BFRY mixture model shows power law behavior between the pathways and mutations.