# A Bayesian Feature Allocation Model for Tumor Heterogeneity

**Peter Müller**
*UT Austin, pmueller@math.utexas.edu*

Juhee Lee
*UCSC, juheelee@soe.ucsc.edu*

Yuan Ji
*Northshore and U. Chicago, jiyuan@uchicago.edu*

Yanxun Xu
*Johns Hopkins U., yxu.stat@gmail.com*

**Keywords**: feature allocation, latent structure, posterior simulation

**Abstract**: We develop a feature allocation model for inference on genetic tumor variation. We analyze data on mutations (single nucleotide variants). In [3] we characterize tumor variability by assuming hypothetical homogeneous cell subpopulations that are defined by the presence of some subset of the recorded mutations. Assuming that each sample is composed of sample-specific proportions of these subclones we can then fit the observed proportions of mutations (variant allele fractions) for each sample. Taking a Bayesian perspective, we proceed with a prior probability model for all relevant unknown quantities, including in particular a prior probability model on the binary indicators that characterize the latent subpopulations by selecting (or not) the recorded mutations. Such prior models are known as feature allocation models [2]. We define a simplified version of the Indian buffet process, one of the most traditional feature allocation models.
We discuss some limitations and variations of the model, including an extension that adds inference on structural variations (copy number variation) [4] and an implementation of efficicient posterior simualtion using small variance asymptotics [5].
Another important limitation of the approach is the a priori independent nature of the model, independent across the latent features. We argue that such independence complicates interpretation in any application where latent structure, such as the features in this model, should be interpreted as meaningful biologic structure. We briefly discuss an alternative class of models based on the determinental point process (DPP) (e.g., [1]). The DPP implements a repulsive prior for the latent structure, thereby facilitating the interpretation of imputed features etc. as meaningfully distinct biologic structure. We show an application of this model to the problem of inference for tumor heterogeneity.

## References

[1] Affandi, R. H., Fox, E. B., and Taskar, B. (2013). Approximate inference in continuous determinantal point processes. *arXiv preprint arXiv:1311.2971*.

[2] Broderick, T., Pitman, J., and Jordan, M. I. (2013). Feature allocations, probability functions, and paintboxes. *Bayesian Analysis*, page to appear.

[3] Lee, J., Müller, P., Gulukota, K., and Ji, Y. A Bayesian feature allocation model for tumor heterogeneity. *Annals of Applied Statistics*, in press, 2015.

[4] Lee, J., Müller, P., Sengupta, S., Gulukota, K., and Ji, Y. Bayesian inference for tumor subclones accounting for sequencing and structural variants. *arXiv:1409.7158v1* , 2015.

[5] Xu, Y., Müller, P., Yuan, Y., Gulukota, K., and Ji, Y. MAD Bayes for Tumor Heterogeneity Feature Allocation with Non-Normal Sampling. *Journal of the American Statistical Association*, in press, 2015.