Scaled process priors: Improved predictions and uncertainties for new-feature counts via random scaling in Bayesian nonparametrics

Lorenzo Masoero, Federico Camerlenghi, Stefano Favaro and Tamara Broderick

MIT, CSAIL [lom@mit.edu] University of Milan University of Turin MIT, CSAIL

- Modern genomics for personalized medicine: *study **rare functional variants.***
- To unlock full potential of genomics based approach, need effective catalogue of rare variants
  - Quantitative framework/prediction problem: how much can we ***expect to learn*** from new data?
  - *Provide* fast, **improved prediction and uncertainty** for # new genomic variants to be seen in future samples

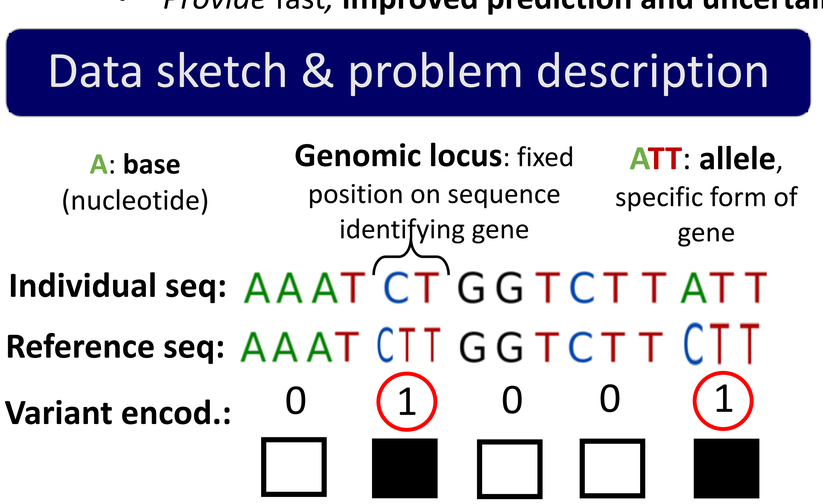
Online during session C25: Bayesian Computation and Modeling for Complex Data
Thu, July 1st 2021, C25: 3:45 pm - 5:00 pm [EDT]

Data sketch & problem description

A: **base** (nucleotide)

Genomic locus: fixed position on sequence identifying gene

ATT: **allele**, specific form of gene



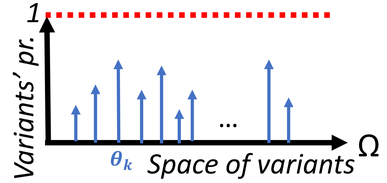
Variant ``called'' whenever **allele differs from reference genome**

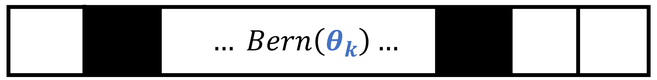
- **Goal**: form exhaustive catalogue to "map" rare variants
- **Problem:** present in few individuals, need a lot of data;
  - Need predictive methods to guide future experiments
  - BNP methods based on posterior predictive distribution of **completely random measures** [CRMs] are *simple, fast*, but:
    - Weak sample dependence: # future new variants depends explicitly on data only via # datapoints N
    - Misspecification**:** # future new variants ***must*** follow ***Poisson*** distribution; real data often over dispersed
- **Our contribution:** Revisit alternative class of "scaled" random measures [James, Orbanz, Teh 2015];
  - ***Scaled stable process:*** *simple, closed form* predictive distr.
  - Overcomes CRM framework limitations, better predictions
    - additional sample dependence (# distinct variants K_N)
    - flexible predictive distribution (negative binomial)

BNP models for genetic variants

Old: Completely Random Measures

- Variants probabilities mutually *independent*
- # new variants depends explicitly on observations only via sample size *N*

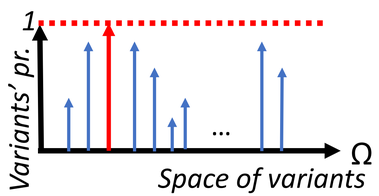


$X_n =$ 

Proposition [informal]: For any CRM prior paired with a Bernoulli likelihood process, # new features in *M* new samples given *N* datapoints follows $\text{Poisson}(f(v, N, M))$

New: Scaled Random Measures

- Scale CRM by largest jump Δ_1
- Resulting measure **not** a CRM: dependence across rates
- Posterior predictive can have richer sample dependence, but not necessarily nice form



Special case: we focus on Scaled Stable process (SSP):

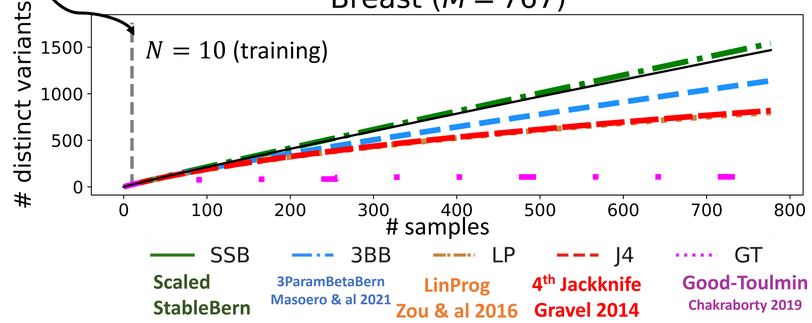
- Simple, closed form posterior predictive
- Overcomes CRM limitations

Proposition [informal]: For $\mu \sim$SSP paired with Bernoulli likelihood process, # new features in *M* new samples given *N* datapoints follows $\mathbf{NegBin}(f(v, N, \boldsymbol{K_N}, M))$

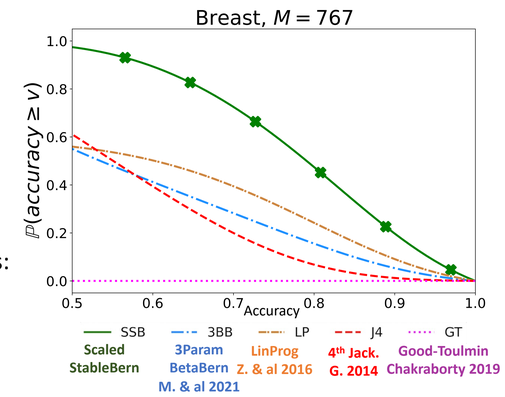
Experimental results

Prediction: # new variants will be observed in cancer samples:

- Retain N observations for training
- Extrapolate up to M such that $N \ll M$



- Future directions:
  - Are there priors which exploit the full frequency spectrum ***and*** have tractable posterior predictive distributions?
  - Coverage and accurate calibration of BNP models: can we obtain fully calibrated posterior predictives?



*Masoero L., Camerlenghi F., Favaro S. and Broderick T. "***Improved uncertainties for predicting new-feature counts via random scaling in Bayesian nonparametrics***" [In preparation]*
*Masoero L., Camerlenghi F., Favaro S. and Broderick T. "*More for less: predicting and maximizing genomic variants discovery via Bayesian nonparametrics*" [Biometrika, forthcoming]*
*James L. F., Orbanz P., Teh Y. W. "*Scaled subordinators and generalizations of the Indian buffet process*" [arXiv:1510.07039]*