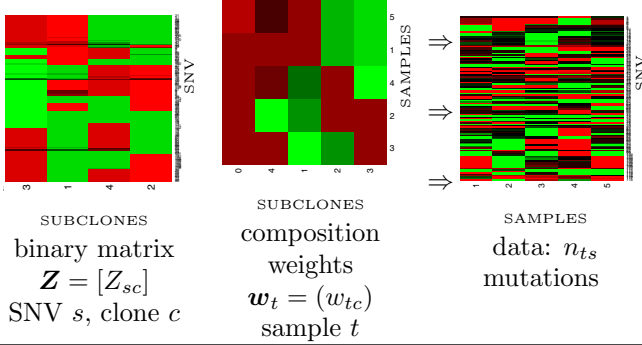


- (iii) Prior $p(\mathbf{Z})$ on $(S \times C)$ binary matrix \mathbf{Z} ,
prior $p(\mathbf{w})$ on mixture weights w_{tc} for composition
(i).

Modeling Tumor Heterogeneity

PETER MÜLLER, UT Austin

JUHEE LEE, UCSC, YUAN JI, U Chicago & NorthShore, K.
GULUKOTA, NorthShore Health System



Tumor Heterogeneity

- Mutations acquired over a tumor's life history
- Every new mutation gives rise to a new subpopulation of cells ("subclone")
- \rightarrow heterogeneous population of cells, composed of subpopulations with varying numbers of mutations.
- Tumor history imprinted in each sample as the mosaicism of mutations.

Data

SNV: point mutations, $s = 1, \dots, S$

Data: $N_{st} = \#$ reads mapped to locus of SNV s in sample t .
 $n_{st} = \#$ of these *with* SNV.

Sampling model: $n_{st} \sim \text{Bin}(N_{st}, p_{st})$

Prior: in words,

- p_{st} arises as a composition of sample t as a mixture of C latent cell subclones.
- Mutation s in subclone c is either present ($Z_{sc} = 1$) or not ($Z_{sc} = 0$).
 $\mathbf{Z}_c = (Z_{sc}, s = 1, \dots, S)$ defines subclone, c .

Inference

Goal: Reconstruct cell subpopulations = estimate \mathbf{Z} and C .

Problem: Deconvolution of p_{st} as a mixture of binary indicators Z_{sc}

$$p_{st} = \sum_c w_{tc} Z_{sc} + w_{t0} p_{s0}$$

plus "background noise"

Real problem: \mathbf{Z} is latent, need to infer \mathbf{Z} from the data.

Identifiability: In principle even feasible with one sample.
Weights are identified across mutations s .

Prior

Latent cell types: $p(\mathbf{Z})$ on $(S \times C)$ binary matrix, w. random C .

Feature allocation: Think of SNV s selecting cell types c
Features (dishes) = c ; experimental units (customers) = s

Random feature allocation: define $p(\mathbf{Z})$ as

- $p(Z_{sc} = 1 | \pi_c) = \pi_c, c = 1, \dots, C$
- $\pi_c \sim \text{Be}(\frac{\alpha}{C}, 1)$
- Drop unselected features

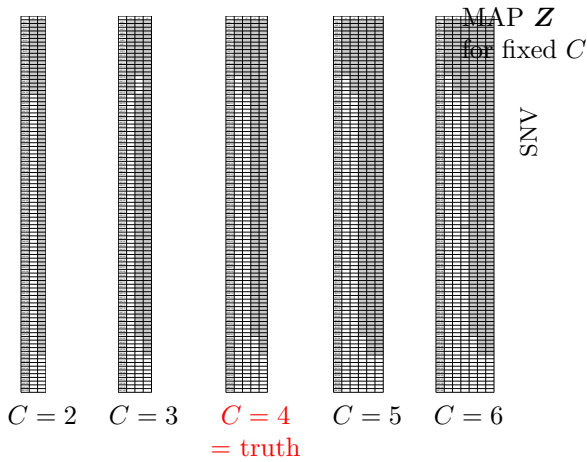
IBP as $C \rightarrow \infty$.

Composition of sample t as mix of cell types:
 $(w_{tc}, c = 1, \dots, C) \sim \text{Dir}(\cdot)$.

This is for normal sampling, asymptotically for small variance and shrinking total mass.

Slide 6

Simulation



IBP: Broderick et al. (2013) extend a similar argument to the IBP, with normal sampling and small variance and shrinking rate of new features,

IBP with binomial sampling: same argument can be made :-)
using increasing scaling of Bin with β and shrinking IBP par γ , using $\gamma = \exp(-\beta\lambda^2)$

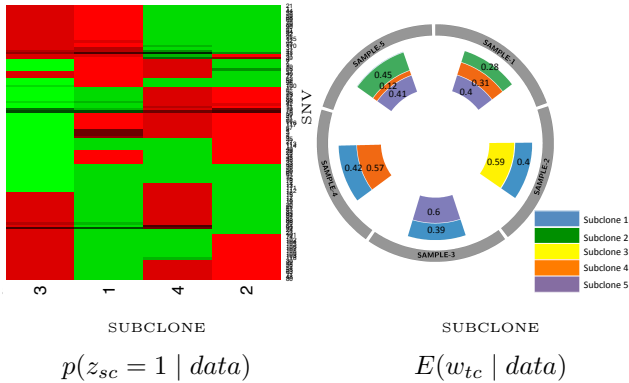
Approx posterior: use k-means with different starting values to characterize posterior.

Slide 10

Slide 7

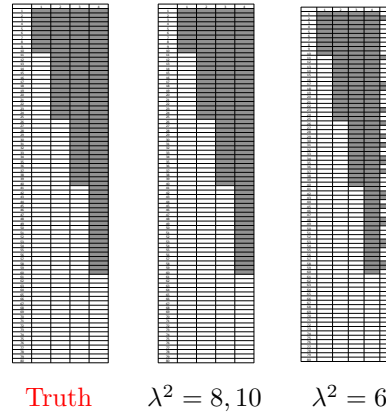
Results – Pancreatic Cancer

$n = 5$ samples of pancreatic cancer (PDAC, pancreatic ductal adenocarcinoma).



Simulation

True and estimated Z



Slide 11

Slide 8

Computation

... is a pain.

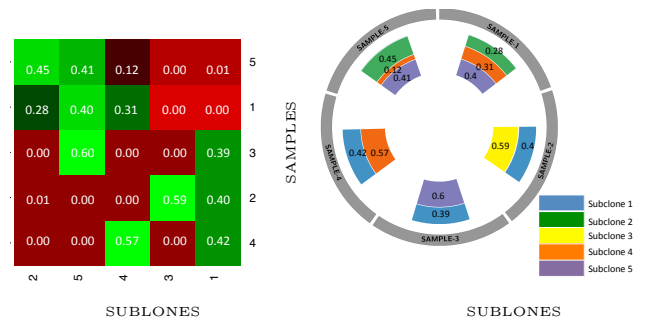
Slide 9

MAD Bayes for TH

with YANXUN XU, UT Austin; YUAN YUAN, Baylor C.of Med.;
YUAN JI and KAMALAKAR GULUKOTA, NorthShore Hospital.

DP mixture: Kulis & Jordan (2012) recognize log posterior \approx criterion function in k-means – voila!

Results – Pancreatic Cancer
 $n = 5$ samples of pancreatic cancer (PDAC, pancreatic ductal adenocarcinoma). Estimated w_{tc} :



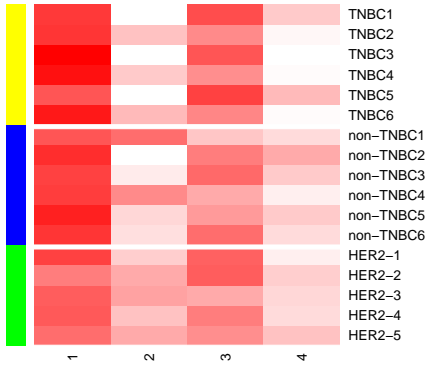
$S = 118$ SNV's in KEGG pathway

$S = 7000$ SNV's

Slide 12

Results – Breast Cancer

Horvath et al. (2013): $n = 17$ BC patients, $S = 329$ SNV's.



Estimated w_{tc} with $C = 4$.

Slide 13

Summary

TH: Model-based estimation of cell subpopulations is possible – and seems to work.

Big data: MCMC is not feasible anymore – alternative approaches remain feasible.

Limitations: and extensions

Tumor phylogenetics: Without condition on phylogenetic tree of subclones

A priori independent cell types: independent $\mathbf{z}_c = (Z_{1\dots S,c})$, with $p(\mathbf{z}_c = \mathbf{z}_{c'}) > 0$, a priori (i know – arrgh!)

Alternative dependent prior using DPP or others.

CNV: we conditioned on N_{st} .

Could use N_{st} to learn about CNV.

SOFTWARE: R code at <http://www.ma.utexas.edu/users/yxu/>