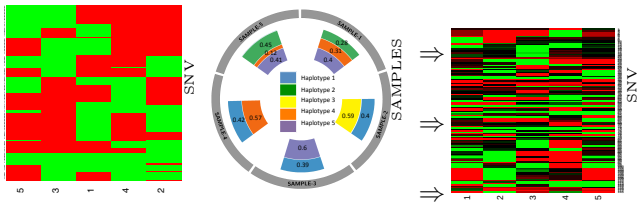
Slide 1

Feature allocation models for tumor heterogeneity

Peter Müller and Yanxun Xu, UT Austin

Juhee Lee, UCSC, Yuan Ji, U Chicago & NorthShore, Subhajit SENGUPTA and K. Gulukota, NorthShore Health System



HAPLOTYPES
binary matrix
$\boldsymbol{Z} = [Z_{sc}]$
SNV s,
haplotype c

HAPLOTYPES
composition
weights
$\boldsymbol{w}_t = (w_{tc})$
sample t

SAMPLES
data: n_{ts}
mutations

1 Tumor Heterogeneity

Slide 2

Tumor Heterogeneity

- Mutations acquired over a tumor's life history
- Every new mutation gives rise to a new subpopulation of cells ("subclone" = pair of haplotypes)
- $\longrightarrow$ heterogeneous population of cells, composed of subpopulations with varying numbers of mutations (e.g., Gerlinger et al. (2012, NEJM).
- Tumor history imprinted in each sample as the mosaicism of mutations.

Slide 3

Data

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

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,

(i) p_{st} arises as a composition of sample t as a mixture of C latent haplotypes.
(pairs of haplotypes define subclones).

(ii) Mutation s in haplotype c is either present ($Z_{sc} = 1$) or not ($Z_{sc} = 0$).
$\boldsymbol{Z}_c = (Z_{sc},\ s = 1, \ldots, S)$ defines haplotype c.

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

Slide 4

Inference

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

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

$$p_{st} = \sum_c w_{tc} Z_{sc} \quad + w_{t0} p_0 \qquad (1)$$

plus "background noise"

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

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

Alternatives:
- cluster variant allele fractions (VAF), $f_{st} = \frac{n_{st}}{N_{st}}$;
$\rightarrow$ subclones (e.g., PyClone: Roth et al, 2014 Nature Meth)
- mixture of Beta's for observed VAF (SciClone: Miller et al., 2014 PLOS Comp Bio); variational Bayes
- CNV data, fit mle for the prob of read alignments (THeta: Oesper et al., 2013 Genome Bio); mle mixture decomposition.
- instead (1) explicitely models decomposition of samples into (hypothetical) subclones.

2 Feature allocation

Slide 5

Feature allocation

Feature allocation: binary matrix $\boldsymbol{Z}$
rows = mutations $s = 1, \ldots, S$;
cols = haplotypes ($\subseteq \{1, \ldots, S\}$)

- Each mutation is in *multiple* subsets (haplotypes)

- Binary matrix $\boldsymbol{Z} = [Z_{sc}]$ records membership of mutations in haplotypes.

vs.

Clustering: each mutation is in *exactly one* subset (partition), row sum $\sum_c Z_{sc} = 1$

Slide 6

Indian buffet process

(Griffiths & Ghahramani, 2005 NIPS)

Equivalent (original) definition of $p(\boldsymbol{Z})$. Let Z_{sc} = customer s selects dish c.

First customer: selects $C_1 \sim \text{Poi}(\alpha)$ dishes, $Z_{1c} = 1$, $c = 1, \ldots, C_1$.
Let $C = C_1$

s-th customer: Let $m_{sc} = \sum_{r<s} Z_{rc}$;

- select dish $c = 1, \ldots, C$ with prob $p(Z_{sc} = 1) = \frac{m_{sc}}{s}$;
- $C_s \sim \text{Poi}(\alpha/s)$ new dishes, $Z_{sc} = 1$,
- $c = C + 1, \ldots, C + C_s$
- Set $C \equiv C + C_s$

Prior for feature allocaiton $\boldsymbol{Z}$.

3 Model for TH

Slide 7

Prior

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

Feature allocation prior: Indian buffet process $p(\boldsymbol{Z})$ (IBP),
with customers (experimental units) $s = 1, \ldots, S$ selecting dishes (features) $c = 1, \ldots, C$
Think of SNV s selecting haplotype (feature) c

IBP: define $p(\boldsymbol{Z})$, first for fixed C,

- $\pi_c \sim \text{Be}\left(\frac{\alpha}{C}, 1\right)$ for each feature $c = 1, \ldots, C$
- $p(Z_{sc} = 1 \mid \pi_c) = \pi_c$, $s = 1, \ldots, S$
- Drop unselected features

$C \to \infty$ defines the IBC (Indian buffet process) $p(\boldsymbol{Z})$

Composition of sample t as mix of haplotypes: $(w_{tc}, \quad c = 1, \ldots, C) \sim \text{Dir}(\cdot)$, for each sample, $t = 1, \ldots, T$.

Slide 8

Results – Pancreatic Cancer

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

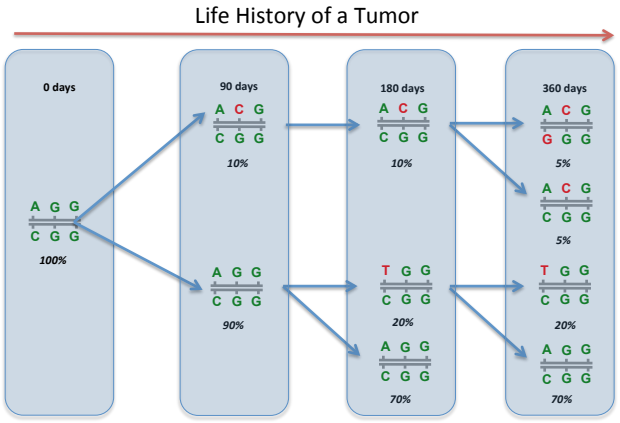
see figures on slide 1.

4 Adding copy number variation

Slide 9

Haplotypes vs. Subclones

So far our discussion is on haplotypes.
But cell subpopulations (subclones) are defined by pairs of haplotypes.



for a diploid organism, pairs of haplotypes define a unique genome.

- Next: will change to subclones (pairs of haplotypes) as experimental units.
- Subclone can have $Z_{sc} \in \{0, 1, 2\}$ copies of each mutation.

Slide 10

Alternative models and generalizations

To represent subclones and more we relax assumptions, using

(i) cIBP for **subclones** (= pairs of haplotypes)

(ii) CNV's: use data on **copy number variation** N_{st}

(iii) repulsive priors (DPP): IBP includes independence across columns!

Slide 11

Extension (i): cIBP

Subclones: pair of haplotypes, $Z_{sc} \in \{0, 1, 2\}$

Categorical IBP: Categorical generalization of the IBP, $p(\boldsymbol{Z})$ for a random trinary matrix with $z_{sc} \in \{0, 1, 2\}$

	1	2	3	4	5
1	0.5	1	0	1	0
2	1	0.5	1	1	1
3	0.5	0	0	0	0.5
4	0.5	0	0.5	0	0.5
5	1	1	0.5	0.5	0.5
6	1	0	0.5	0	0
7	1	0	0	0	0
8	1	0.5	0	0.5	1
9	1	0.5	1	1	1
10	0.5	0	0	0	1

Slide 12

Extension (ii): CNV & SNV

CNV & SNVs: modeling of copy number variation *and* SNV given CNV.
Earlier: $p(n_{st} \mid N_{st}, \boldsymbol{Z}, \ldots)$ conditional on N_{st} and $\boldsymbol{Z}$ =SNV's.
Now: $p(n_{st} \mid N_{st}, \boldsymbol{Z}, \ldots)$ and $p(N_{st} \mid \boldsymbol{L}, \ldots)$,
with $\boldsymbol{L} = [\ell_{sc}]$ = CNV.

CNV: Latent matrix $\boldsymbol{L} = [\ell_{sc}]$ reports copy gain ($\ell_{sc} > 2$), loss ($\ell_{sc} < 2$) or neutral ($\ell_{sc} = 2$) for each locus s and *subclone* c under consideration.

$$\boldsymbol{L} \sim \text{cIBP}$$

Sampling model:

$$N_{st} \mid \phi_t, M_{st} \sim \text{Poi}(\phi_t M_{st}/2). \text{ with } M_{st} = \sum_{c=1}^{C} w_{tc}\ell_{sc},$$

Slide 13

SNV: conditional on ℓ_{sc}:

$$z_{sc} \mid \ell_{sc} \sim \text{Unif}\{0, \ldots, \ell_{sc}\}$$

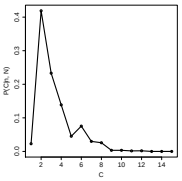
and sampling model

$$\begin{aligned} n_{st} \mid N_{st}, \boldsymbol{Z}, \boldsymbol{L}, \ldots &\sim \text{Bin}(N_{st}, p_{st}) \\ p_{st} &= \frac{p_0 z_{s0} w_{t0} + \sum_{c=1}^{C} w_{tc} z_{sc}}{M_{st}} \end{aligned}$$

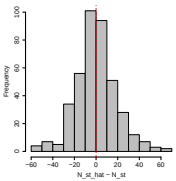
That's all!

Slide 14

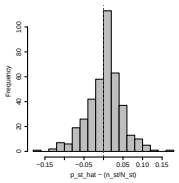
Lung Cancer – 4 Samples



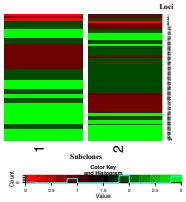
(a) $p(C \mid \boldsymbol{n}, \boldsymbol{N})$



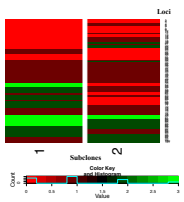
(b) $\hat{N}_{st} - N_{st}$



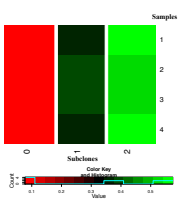
(c) $\hat{p}_{st} - (n_{st}/N_{st})$



(a) $\boldsymbol{L}^\star$



(b) $\boldsymbol{Z}^\star$



(c) $\boldsymbol{w}^\star$

5 Repulsive Prios

Slide 15

Extension (iii): Repulsive priors

Independence: IBP implies independence across features! This should be *proibido*, verboten, vietato...

Latent biologic structure: common theme – inference for latent biologic structure

- feature allocation: used here, in TH
- mixture model: interpreting components as biolog. meaningful
- clustering: e.g., of patient population

Desiderata: distinct features should be diverse to allow interpretation, to serve as meaningful clinical targets, etc.;

Independent priors: over features, mixture components etc., is usually inappropriate, but often used.
E.g., IBP implies independent prior over columns $\boldsymbol{z}_c$.

Repulsive priors: replace independent prior by repulsive priors, e.g.,
Determinantal point process (DPP; Macchi, 1975; Lavancier et al. 2015, JRSSB; Affandi et al., 2013)

Slide 16

DPP on images "Jaguar"

$X = \{x_1, \ldots, x_K\}$, $x_k \in$ "Jaguar images"

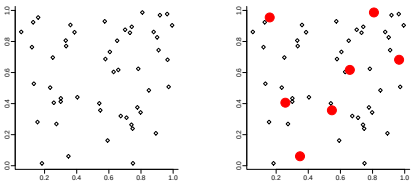
Want more diversity, fewer cats – more football teams etc. :-)

Slide 17

DPP

DPP: point process $X = \{x_1, \ldots, x_K\}$ on $x_i \in S$ for some space S, e.g. $S \subseteq \Re^2$ or $S =$ images.

Idea: instead of many similar points x_j, generate only few distinct ones.



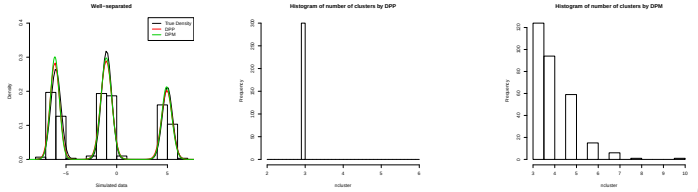
indep (unif) DPP

Repulsive point process: avoid duplication of similar values x_j;
for example, google-ing "Jaguar", you want $X = \{$ cat, car, football team, $\ldots\}$

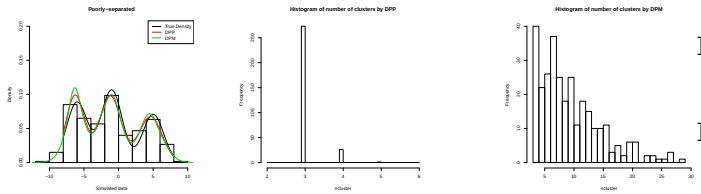
Slide 18

DPP – Examples

Simulation truth = mix of normals (left). Clustering model with DPP (center) vs. DP mixture (right) prior.



data & sim truth DPP: n clusters DP mixture: n clusters



Density estimation for a mix of normals.

Slide 19

Simulation truth Z^o (left). Feature allocation with DPP prior vs. IBP on features

Truth Z^o DPP: Estimated $\hat{Z}$ IBP: Estimated $\hat{Z}$

Slide 20

DPP

DPP: point process $X = \{x_1, \ldots, x_K\}$ on $x_i \in S$ for some space S;
penalizes "similar" x_j.

DPP on finite discrete space: $p(X) \propto \det C_X$
where $C_{X,ij} = C(x_i, x_j)$ for a p.d. kernel $C(x, x')$.

DPP on (bounded) continuous space: density w.r.t. unit rate Poisson process

$$f(X) = \det C_X / \prod (1 + \lambda_h)$$

with λ_h = eigenvalues of operator $T : h \to \int_S C(x, y) h(y) dy$.

MCMC: easy for finite S (e.g., Kulesza, A. and Taskar, B. (2012 Machine Learn.)
for continuous S: reversible jump MCMC with $f(X)$ (Xu et al., 2015 arXiv)

Slide 21

Summary

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: indpendent $\boldsymbol{z}_c = (Z_{1\ldots S,c})$, with $p(\boldsymbol{z}_c = \boldsymbol{z}_{c'}) > 0$, a priori (i know — arrgh!)
Alternative dependent prior using DPP or others.

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

6 Extra Slides – MAD Bayes

Slide 22

Posterior inference for IBP using MAD Bayes

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

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

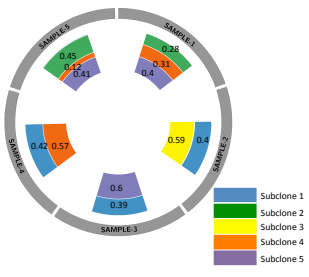
IBP with binomial sampling: same argument can be made :-)
using increasing scaling of Bin with β and shrinking IBP par γ, using $\gamma = \exp(-\beta\lambda^2)$

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

Slide 23

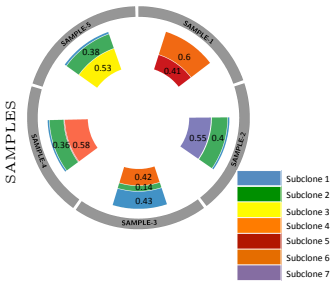
Results – Pancreatic Cancer

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



HAPLOTYPES

$S = 118$ SNV's in KEGG pathway



HAPLOTYPES

$S = 7000$ SNV's

References

Slide 24

Affandi, R. H., Fox, E., and Taskar, B. (2013). Approximate inference in continuous determinantal processes. *Adv in Neural Inf Processing Systems*, 1430–1438.

Lavancier, F., Møller, J., and Rubak, E. (2015). Determinantal point process models and statistical inference. *JRSSB*, 77, 853–877.

Lee, J., Müller, P., Ji, Y. and Gulukota, K. (2015) "A Bayesian Feature Allocation Model for TH," *Ann. of Applied Stat*, 9, 621-639.

Lee, J., Müller, P., Sengupta, S., Gulukota, K. and Ji, Y. (2014). "Bayesian Inference for Intra-Tumor Heterogeneity in Mutations and Copy Number Variation," *JRSSC*, final(?) revision.

Roth, A. et al. (2014). "PyClone ...", *Nature Methods*, 11, 396-

Sengupta, S., Guluokta, K., Lee, J., Müller, P., and Ji, Y. (2015) "Bayclone: Bayesian Nonparametric Inference of Tumor Subclones Using NGS Data." In *Proceedings of The Pacific Symposium on Biocomputing (PSB) 2015*, 467-78.

Xu Y, Müller P, Yuan Y, Gulukota K and Ji Y, (2015). "MAD Bayes for Tumor Heterogeneity," *JASA*, 110, 503-514.