

Bayesian Feature Allocation Models for Natural Killer Cell Repertoire Studies Using Mass Cytometry Data

Arthur Lui

Advisor: Juhee Lee

Department of Applied Mathematics and Statistics
UC Santa Cruz

8 June 2018

► Cytometry at time-of-flight (CyTOF)

- commercialized in 2009
- makes use of time-of-flight mass spectrometry to accelerate, separate, and identify ions by mass
- enables detection of many parameters (biological, phenotypic, or functional markers) in less time and at a higher resolution [Cheung and Utz, 2011]
- led to greater understanding of natural killer (NK) cells



- **Natural Killer cells** play a critical role in cancer immune surveillance and are the first line of defense against viruses and transformed tumor cells.

Introduction

- ▶ NK cell diversity refers to number of NK cell sub-populations within NK cells, and also affects antiviral response
- ▶ Drs. Thall and Rezvani, collaborators at UT MD Anderson Cancer Center, have conducted clinical trials to study the potential clinical efficacy of umbilical cord blood (UCB) transplantation as a therapy for leukemia.
- ▶ UCB NK cell therapy has the advantage of low risk of viral transmission from donor to recipient [Sarvaria et al., 2017].
- ▶ In the trials, leukemia patients received UCB cell transplants, and NK cell surface protein markers are measured using mass cytometry.

	2B4	2DL1	2DL3	2DS4	3DL1	CCR7	...
1	47.60	0.00	30.90	1.35	82.49	0.00	...
2	81.84	0.44	0.88	0.51	176.99	2.38	...
3	13.33	0.00	0.00	0.00	0.00	8.81	...
4	23.64	3.37	43.39	0.27	0.73	0.00	...
5	156.19	0.00	9.04	0.00	0.00	11.43	...
6	273.86	0.00	9.71	2.41	0.84	0.52	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Cutoff	7.62	6.07	13.60	3.79	15.50	9.52	...

Table 1: Cord-blood sample marker expression levels for 6 of 32 NK-cell markers (columns), and 6 of 41474 cells (rows). Last row contains cutoff values returned by CyTOF instrument.

- ▶ Data missing not at random
 - ▶ Some markers contain up to 85% missing values
- ▶ Cutoff values are computed after measurement

	2B4	2DL1	2DL3	2DS4	3DL1	CCR7	...
1	1	0	1	0	1	0	...
2	1	0	0	0	1	0	...
3	1	0	0	0	0	0	...
4	1	0	1	0	0	0	...
5	1	0	0	0	0	1	...
6	1	0	0	0	0	0	...
⋮	⋮	⋮	⋮	⋮	⋮	⋮	
Cutoff	7.62	6.07	13.60	3.79	15.50	9.52	...

Table 2: Cell phenotypes (rows)

- Obtaining cell phenotypes using overly-simplistic methods may yield unreasonably high number of sub-populations.

Existing Methods

- ▶ Most existing methods use traditional clustering methods (K-means, hierarchical clustering, density-based clustering, nearest-neighbor clustering, etc.)
- ▶ For high-dimensional cytometry data, Weber and Robinson [2016] compared existing clustering methods including **FlowSOM** [Van Gassen et al., 2015], **PhenoGraph** [Levine et al., 2015], **Rclusterpp** [Linderman et al., 2013], and **flowClust** [Lo et al., 2009]
- ▶ Existing methods do not directly model latent phenotypes or quantify model uncertainty

Proposed Projects

- ▶ **Project I:** Bayesian Feature Allocation Model for Heterogeneous Cell Populations
- ▶ **Project II:** Repulsive Feature Allocation Model
- ▶ **Project III:** Feature Allocation Model with Regression for Abundances of Features in Longitudinal Data

Project I

Feature Allocation Model – Indian buffet process

Griffiths et al. Griffiths and Ghahramani [2011] introduced the IBP as follows:

Let \mathbf{Z} be a $J \times K$ matrix such that

$$\begin{array}{l|l} z_{jk} & \pi_k \sim \text{Bernoulli}(\pi_k) \\ \pi_k & \alpha \sim \text{Beta}(\frac{\alpha}{K}, 1) \end{array}$$

for $j = 1, \dots, J$ and $k = 1, \dots, K$, and where α is positive.

Feature Allocation Model – Indian buffet process

Griffiths et al. Griffiths and Ghahramani [2011] introduced the IBP as follows:

Let \mathbf{Z} be a $J \times K$ matrix such that

$$\begin{array}{l|l} z_{jk} & \pi_k \sim \text{Bernoulli}(\pi_k) \\ \pi_k & \alpha \sim \text{Beta}(\frac{\alpha}{K}, 1) \end{array}$$

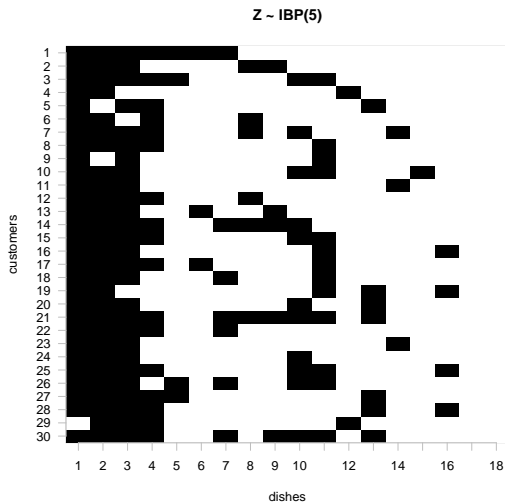
for $j = 1, \dots, J$ and $k = 1, \dots, K$, and where α is positive.

Then as $K \rightarrow \infty$, $\mathbf{Z} \sim \text{IBP}(\alpha)$. It can be shown that

$$P(\mathbf{Z}) = \frac{\alpha^{K_+}}{\prod_{h=1}^{2^J-1} K_h!} \exp\{-\alpha H_J\} \prod_{k=1}^{K_+} \frac{(J - m_k)!(m_k - 1)!}{J!},$$

where $H_J = \sum_{j=1}^J j^{-1}$ is the harmonic number, K_+ is the number of non-zero columns in \mathbf{Z} , m_k is the k^{th} column sum of \mathbf{Z} , and K_h the number of columns having history h (some binary number).

One draw from the IBP



Project I: Bayesian Feature Allocation Model for Heterogeneous Cell Populations

Notation

- ▶ I : Number of samples
- ▶ J : Number of markers
- ▶ N_i : Number of observations in sample i
- ▶ \tilde{y}_{inj} : Raw expression levels for observation n in sample i for marker j . ($\tilde{y}_{inj} \geq 0$)
- ▶ c_{ij} : Cutoff for marker j , sample i
- ▶ y_{inj} : Transformed expression levels for observation n , sample i , marker j

$$y_{inj} = \log \left(\frac{\tilde{y}_{inj}}{c_{ij}} \right) \in \mathbb{R}.$$

- ▶ ($y_{inj} \gg 0$) likely corresponds to expression
- ▶ ($y_{inj} \ll 0$) likely corresponds to non-expression

Project I: Bayesian Feature Allocation Model for Heterogeneous Cell Populations

- ▶ K is a sufficiently large constant
- ▶ \mathbf{Z} : $(J \times K)$ binary matrix defining the latent phenotypes.
 - ▶ if $Z_{jk} = 1$, then marker j is expressed in phenotype k
 - ▶ if $Z_{jk} = 0$, then marker j is not expressed in phenotype k
- ▶ $\lambda_{in} \in \{1, \dots, K\}$: The latent phenotype of observation n , sample i

Project I: Sampling Distribution

$$y_{inj} \mid \eta_{ij}, \mu^*, \sigma_i^{2*} \stackrel{ind}{\sim} \begin{cases} F_0, & \text{if } z_{j,\lambda_{in}} = 0, \\ F_1, & \text{if } z_{j,\lambda_{in}} = 1. \end{cases}$$

- ▶ $F_0 = \sum_{\ell=1}^{L^0} \eta_{ij\ell}^0 \text{Normal}(\mu_{0\ell}^*, \sigma_{0i\ell}^{2*})$, where $\mu_{0\ell}^* < 0$
- ▶ $F_1 = \sum_{\ell=1}^{L^1} \eta_{ij\ell}^1 \text{Normal}(\mu_{1\ell}^*, \sigma_{1i\ell}^{2*})$, where $\mu_{1\ell}^* > 0$

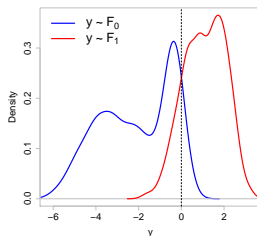


Figure 1: Kernel density estimate of samples from F_0 (blue) and F_1 (red)

Project I: Missing Mechanism

$$m_{inj} \mid p_{inj} \overset{ind}{\sim} \text{Bernoulli}(p_{inj})$$

Project I: Missing Mechanism

$$m_{inj} \mid p_{inj} \stackrel{ind}{\sim} \text{Bernoulli}(p_{inj})$$

$$\text{logit}(p_{inj}) = \begin{cases} \beta_{0i} - \beta_{1i}(y_{inj} - c_0)^2, & \text{if } y_{inj} < c_0 \\ \beta_{0i} - \beta_{1i}c_1 (y_{inj} - c_0)^{1/2}, & \text{otherwise,} \end{cases}$$

where $m_{inj} = 1$ if y_{inj} is missing, and 0 otherwise.

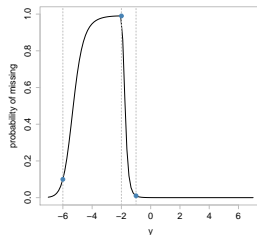


Figure 2: Example missing mechanism

Latent Phenotypes

$$\begin{aligned}\mathbf{Z} \mid \alpha &\sim \text{IBP}_K(\alpha) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha)\end{aligned}$$

Latent Phenotypes

$$\begin{aligned} z_{jk} \mid v_k &\stackrel{ind}{\sim} \text{Bernoulli}(v_k) \\ v_k \mid \alpha &\stackrel{iid}{\sim} \text{Beta}(\alpha/K, 1) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \end{aligned}$$

Latent Phenotypes

$$\begin{aligned}z_{jk} \mid h_{jk}, v_k &= \mathbb{I}\{\Phi(h_{jk} \mid 0, 1) < v_k\} \\v_k \mid \alpha &\stackrel{iid}{\sim} \text{Beta}(\alpha/K, 1) \\ \mathbf{h}_k &\stackrel{iid}{\sim} \text{Normal}_J(\mathbf{0}, \mathbf{I}) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha)\end{aligned}$$

Latent Phenotypes

$$\begin{aligned} z_{jk} \mid h_{jk}, v_k &= \mathbb{I}\{\Phi(h_{jk} \mid \mathbf{0}, \Gamma_{jj}) < v_k\} \\ v_k \mid \alpha &\stackrel{iid}{\sim} \text{Beta}(\alpha/K, 1) \\ \mathbf{h}_k &\stackrel{iid}{\sim} \text{Normal}_J(\mathbf{0}, \Gamma) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha) \end{aligned}$$

- ▶ Dependent IBP (dIBP) [Williamson et al., 2010] construction to model correlations between markers
- ▶ dIBP reduces to IBP [Griffiths and Ghahramani, 2011] when Γ is the identity matrix

Latent Phenotypes

$$\begin{aligned}\mathbf{Z} \mid \alpha, \Gamma &\sim \text{dIBP}_K(\alpha, \Gamma) \\ \alpha &\sim \text{Gamma}(a_\alpha, b_\alpha)\end{aligned}$$

- ▶ Dependent IBP (dIBP) [Williamson et al., 2010] construction to model correlations between markers
- ▶ dIBP reduces to IBP [Griffiths and Ghahramani, 2011] when Γ is the identity matrix

Phenotype Abundance

Let w_{ik} denote an abundance level of phenotype k in sample i . Let $\mathbf{w}_i = (w_{i1}, \dots, w_{iK})$. Then, $\mathbf{w}_i \mid K \stackrel{iid}{\sim} \text{Dirichlet}_K(d/K)$.

Latent Cell Phenotype Indicators

$$p(\lambda_{in} = k \mid \mathbf{w}_i) = w_{ik}$$

Project I: Priors

$$\mu_{0\ell}^* \mid \psi_0, \tau_0^2 \stackrel{iid}{\sim} \text{Normal}_-(\psi_0, \tau_0^2), \quad \ell \in \{1, \dots, L^0\}$$

$$\mu_{1\ell}^* \mid \psi_1, \tau_1^2 \stackrel{iid}{\sim} \text{Normal}_+(\psi_1, \tau_1^2), \quad \ell \in \{1, \dots, L^1\}$$

$$\sigma_{0i\ell}^2 \mid s_i \stackrel{ind}{\sim} \text{Inverse-Gamma}(a_\sigma, s_i), \quad \ell \in \{1, \dots, L^0\}$$

$$\sigma_{1i\ell}^2 \mid s_i \stackrel{ind}{\sim} \text{Inverse-Gamma}(a_\sigma, s_i), \quad \ell \in \{1, \dots, L^1\}$$

$$s_i \stackrel{iid}{\sim} \text{Gamma}(a_s, b_s), \quad i \in \{1, \dots, I\}$$

$$\eta_{ij}^0 \stackrel{iid}{\sim} \text{Dirichlet}_{L^0}(a_{\eta^0}/L^0), \quad i \in \{1, \dots, I\}, j \in \{1, \dots, J\}$$

$$\eta_{ij}^1 \stackrel{iid}{\sim} \text{Dirichlet}_{L^1}(a_{\eta^1}/L^1), \quad i \in \{1, \dots, I\}, j \in \{1, \dots, J\}$$

$$\beta_{0i} \stackrel{iid}{\sim} \text{Normal}(m_{\beta_0}, s_{\beta_0}^2), \quad i \in \{1, \dots, I\}$$

$$\beta_{1i} \stackrel{iid}{\sim} \text{Normal}_+(m_{\beta_1}, s_{\beta_1}^2), \quad i \in \{1, \dots, I\}$$

Project I: Posterior Estimate for \mathbf{Z}

- ▶ \mathbf{Z} susceptible to label switching, especially when K is random
- ▶ We summarize the posterior distribution of \mathbf{Z} using sequentially-allocated latent structure optimization (SALSO) [Dahl and Müller, 2017]

Project I: Posterior Estimate for \mathbf{Z}

For each posterior sample $b \in \{1, \dots, B\}$ and patient sample i ,

1. compute a $(J \times J)$ Adjacency Matrix $A_i^{(b)}$, where

$$A_{i,j,j'}^{(b)} = \sum_{k=1}^K w_{ik}^{(b)} \mathbb{1} \left\{ z_{jk}^{(b)} = z_{j'k}^{(b)} = 1 \right\}$$

2. compute the mean adjacency matrix $\bar{A}_i = \sum_{b=1}^B A_i^{(b)} / B$.
3. $\hat{\mathbf{Z}}_i = \operatorname{argmin}_{\mathbf{Z}} \sum_{j,j'} (A_{i,j,j'}^{(b)} - \bar{A}_{i,j,j'})^2$

Project I: Simulation Study

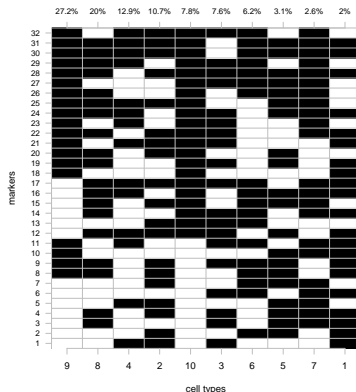


Figure 3: Simulation truth for Z in Sample 1, with markers in rows and latent phenotypes in columns. Black and white represents $z_{jk} = 1$ and 0, respectively. The phenotypes and w_1 are shown at the bottom and on top, respectively. The markers are sorted by w_{ik} .

Project I: Simulation Study

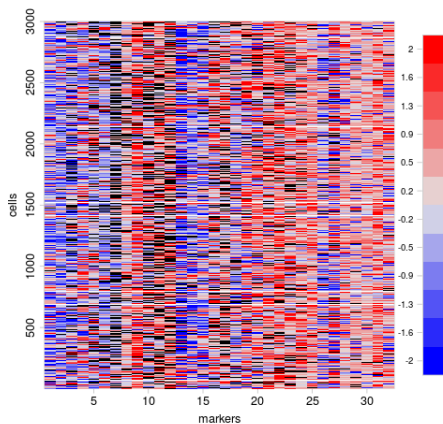
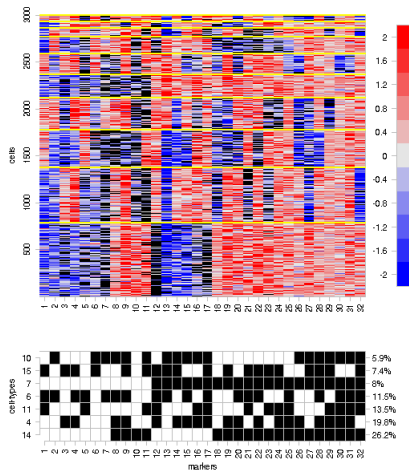
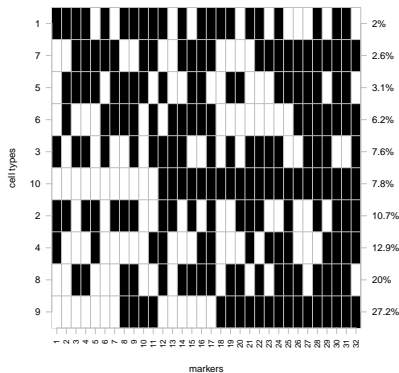


Figure 4: Simulated data for one sample

Project I: Simulation Results – FAM



(a) Sample 1



(b) Z true

Figure 5: FAM Simulation Study

Missing Mechanism Posterior

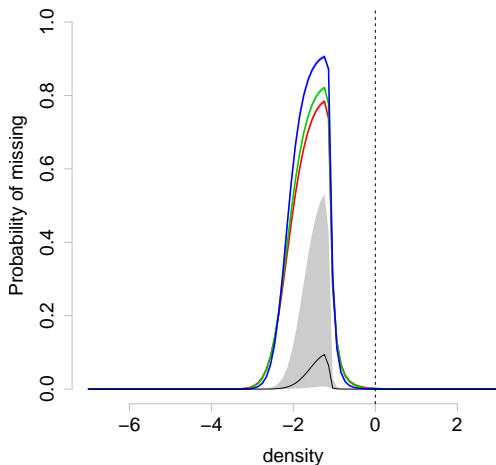
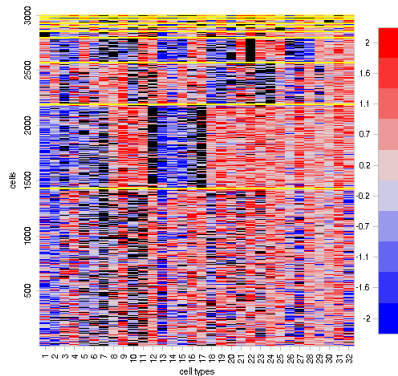
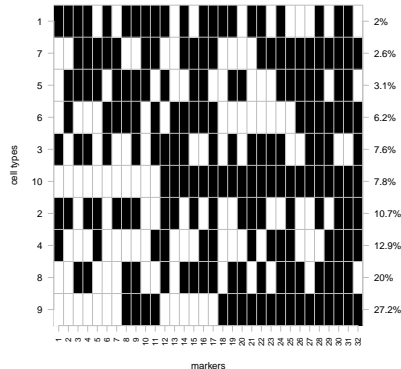


Figure 6: Posterior missing mechanism for simulation study in Project I for sample 1 (red), 2 (green), and 3 (blue). Prior missing mechanism in grey.

Project I: Simulation Results – FlowSOM



(a) Sample 1



(b) Z true

Figure 7: FlowSOM Simulation Study

Project I: Simulation Study – Comparing FAM and FlowSOM

We use the F-score to summarize the accuracy of the computed cluster labels. The F-score is defined as the harmonic mean of precision and recall.

$$F_1 = \frac{2}{\text{precision}^{-1} + \text{recall}^{-1}},$$

- ▶ precision = (true positives) / (true positives + false positives)
- ▶ recall = (true positives) / (true positives + false negatives)
- ▶ $F_1 \in [0, 1]$ with $F_1 = 1$ being a perfect score

	F-score	Elapsed time (seconds)
FlowSOM	0.490	6
FAM	0.999	17472

Table 3: F-score and elapsed time for simulated data for FAM and FlowSOM

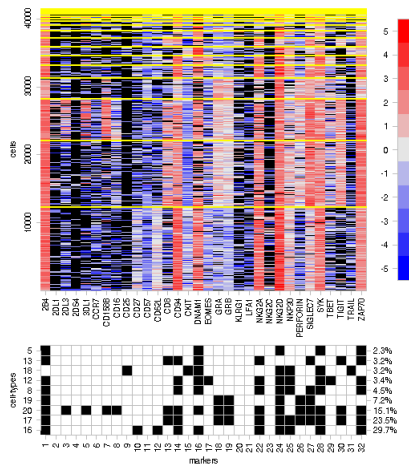
Project I: Conclusions for Comparing FAM and FlowSOM

- ▶ FAM produces posterior distribution and estimates of latent phenotypes and their weights
- ▶ Choose K sufficiently large
- ▶ Graph first two principal components to visually estimate K
- ▶ FlowSOM quickly produces point estimates of clusters
- ▶ In FlowSOM, additional ad-hoc criteria are needed to produce estimates of latent phenotypes and their abundance
- ▶ In FlowSOM, missing values need to be pre-imputed

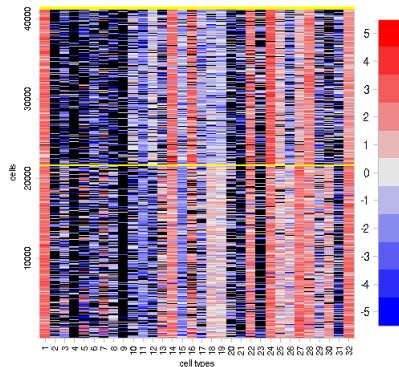
Project I: Cord Blood Samples Data

- ▶ Three cord blood (CB) samples from MD Anderson Cancer Center
- ▶ Number of cells in each sample $N = (41474, 10454, 5177)$
- ▶ Number of markers $J = 32$
- ▶ $K = 20$
- ▶ MCMC: 2000 iterations, 1000 burn-in

Project I: CB Study



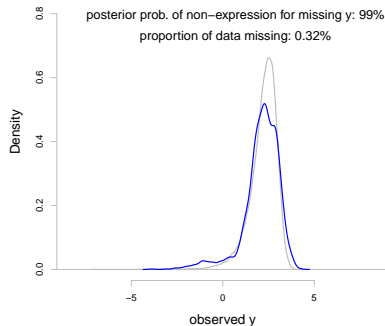
Sample 1 (FAM)



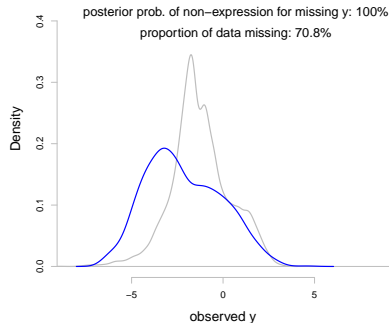
Sample 1 (FlowSOM)

Figure 8: CB data Sample 1, analyzed using FAM (left) and FlowSOM (right)

Project I: Posterior Predictive for Observed Data



(a) Sample 1, marker 1



(b) Sample 1, marker 2

Figure 9: Observed data density for y_{ij} in grey. Posterior predictive density for observed data in blue.

Project I: Probability of Non-expression for Imputed Values

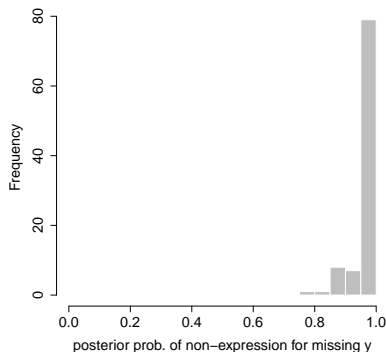


Figure 10: Histogram of posterior probabilities \hat{q}_{ij} of non-expression for missing for all (i, j) . The peak at 1 suggests a marker is not likely expressed if its expression level is missing.

R code for installing cytof3

```
library(devtools)
repo = 'luiarthur/ucsc_litreview'
subdir = 'cytof/src/model3/cytof3'
install_github(repo, subdir=subdir)
```

Project II

Project II: Repulsive Feature Allocation Model

- ▶ Similar or duplicated features may occur in feature allocation matrix under IBP prior
- ▶ Repulsion penalizes similar features in the prior, resulting in a parsimonious model
- ▶ Different approaches for repulsive models have been developed mostly in the context of mixture models [Petrulia et al., 2012, Quinlan et al., 2017b, Xie and Xu, 2017, Quinlan et al., 2017a].

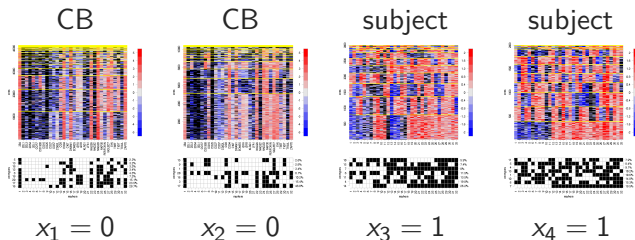
Project II: Repulsive Feature Allocation Model

Objective

- ▶ Develop a repulsive feature allocation model for parsimonious feature matrix
- ▶ Compare phenotypes present in healthy subject samples and cord blood samples

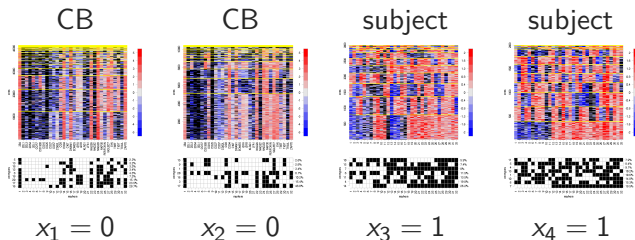
Project II: Feature Selection for Different Samples

- Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).



Project II: Feature Selection for Different Samples

- Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).



- Learn binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k

Project II: rep-FAM Prior

$$P(\mathbf{Z} \mid \nu) \propto \prod_{k=1}^K \left\{ \prod_{j=1}^J v_k^{z_{jk}} (1 - v_k)^{1-z_{jk}} \right\}$$
$$v_K \mid \alpha \sim \text{Beta}(\alpha/K, 1)$$

$$P(\mathbf{Z} \mid \mathbf{v}, C_\phi) \propto \prod_{k=1}^K \left\{ \prod_{j=1}^J v_k^{z_{jk}} (1 - v_k)^{1-z_{jk}} \right\} \times \\ \prod_{k_1=1}^{K-1} \prod_{k_2=k_1+1}^K \{1 - C_\phi(\rho(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}))\} \\ \mathbf{v}_K \mid \alpha \sim \text{Beta}(\alpha/K, 1)$$

where

- ▶ $C_\phi(\cdot)$ is a continuous decreasing function in distance with $C_\phi(0) = 1$ and $\lim_{d \rightarrow \infty} C_\phi(d) = 0$.
- ▶ $\rho(\mathbf{z}_{k_1}, \mathbf{z}_{k_2})$ is a distance metric
- ▶ We use $C_\phi(d) = C(d) = \exp\{-d\}$, and $\rho(\mathbf{z}_{k_1}, \mathbf{z}_{k_2}) = \sum_{j=1}^J |z_{jk_1} - z_{jk_2}|$

Project II: Simulation Study for rep-FAM Prior

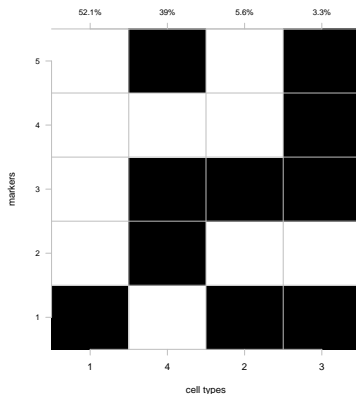
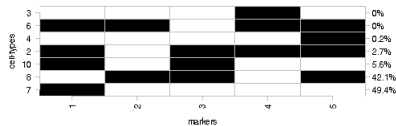
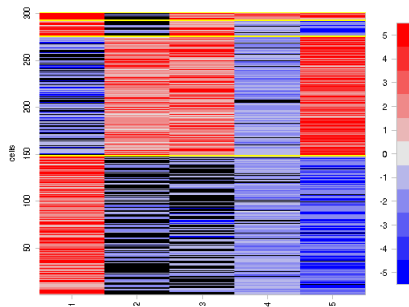
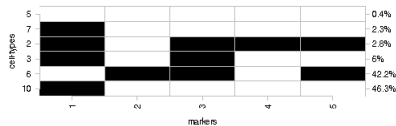
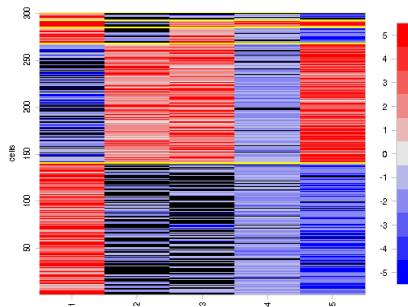


Figure 11: Simulation truth for Z in Sample 1, with markers in rows and latent phenotypes in columns. Black and white represents $z_{jk} = 1$ and 0, respectively. The phenotypes and w_1 are shown at the bottom and on top, respectively. The markers are sorted by w_{ik} .

Project II: Simulation Study for rep-FAM Prior



(a) rep-FAM



(b) IBP

Figure 12: Heatmaps of y for simulated data ordered by phenotypes (Sample 1)

Project III

Project III: Feature Allocation Model with Regression for Feature Abundances Over Time

Objectives:

1. Extend the previous models to analyze samples taken at multiple time points from a patient after NK cell infusion.
2. Model the evolution of NK cell populations over time

Project III: Feature Allocation Model with Regression for Feature Abundances Over Time

- ▶ I samples taken at time points t_1, \dots, t_I
- ▶ Phenotype abundances $\mathbf{w}_{t_i} = (w_{t_i,1}, \dots, w_{t_i,K})$ as a function of time (t_i) after treatment
- ▶ \mathbf{Z} includes possible cell types possessed across all samples

Timeline

Project	Academic Quarter
Project 1	Fall 17 - Fall 18
Project 2	Fall 18 - Spring 19
Project 3	Winter 19 - Fall 19
Thesis	Fall 19 - Winter 20

I

Appendix

References I

- Regina K Cheung and Paul J Utz. Screening: Cytof - the next generation of cell detection. *Nature Reviews Rheumatology*, 7(9): 502, 2011.
- David B. Dahl and Peter Müller. Summarizing distributions of latent structures. *Bayesian Nonparametric Inference: Dependence Structures & Applications Oaxaca, Mexico*, 2017.
- Thomas L Griffiths and Zoubin Ghahramani. The indian buffet process: An introduction and review. *Journal of Machine Learning Research*, 12(Apr):1185–1224, 2011.
- Jacob H Levine, Erin F Simonds, Sean C Bendall, Kara L Davis, D Amir El-ad, Michelle D Tadmor, Oren Litvin, Harris G Fienberg, Astraea Jager, Eli R Zunder, et al. Data-driven phenotypic dissection of aml reveals progenitor-like cells that correlate with prognosis. *Cell*, 162(1):184–197, 2015.

References II

- Michael Linderman, Robert Bruggner, and Maintainer Robert Bruggner. Package 'rclusterpp'. 2013.
- Kenneth Lo, Florian Hahne, Ryan R Brinkman, and Raphael Gottardo. flowclust: a bioconductor package for automated gating of flow cytometry data. *BMC bioinformatics*, 10(1):145, 2009.
- Francesca Petralia, Vinayak Rao, and David B Dunson. Repulsive mixtures. In *Advances in Neural Information Processing Systems*, pages 1889–1897, 2012.
- José Quinlan, Fernando A Quintana, and Garritt L Page. Density regression using repulsive distributions. 2017a.
- José J Quinlan, Fernando A Quintana, and Garritt L Page. Parsimonious hierarchical modeling using repulsive distributions. *arXiv preprint arXiv:1701.04457*, 2017b.

References III

- Anushruti Sarvaria, Dunia Jawdat, J Alejandro Madrigal, and Aurore Saudemont. Umbilical cord blood natural killer cells, their characteristics, and potential clinical applications. *Frontiers in Immunology*, 8, 2017.
- Sofie Van Gassen, Britt Callebaut, Mary J Van Helden, Bart N Lambrecht, Piet Demeester, Tom Dhaene, and Yvan Saeys. Flowsom: Using self-organizing maps for visualization and interpretation of cytometry data. *Cytometry Part A*, 87(7): 636–645, 2015.
- Lukas M Weber and Mark D Robinson. Comparison of clustering methods for high-dimensional single-cell flow and mass cytometry data. *Cytometry Part A*, 89(12):1084–1096, 2016.
- Sinead Williamson, Peter Orbanz, and Zoubin Ghahramani. Dependent indian buffet processes. In *International Conference on Artificial Intelligence and Statistics*, pages 924–931, 2010.

Fangzheng Xie and Yanxun Xu. Bayesian repulsive gaussian mixture model. *arXiv preprint arXiv:1703.09061*, 2017.

Missing Mechanism Posterior

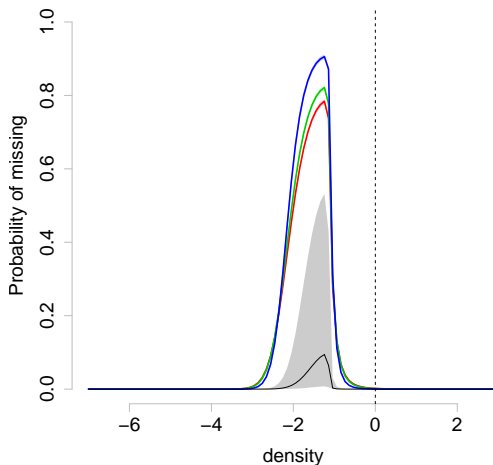
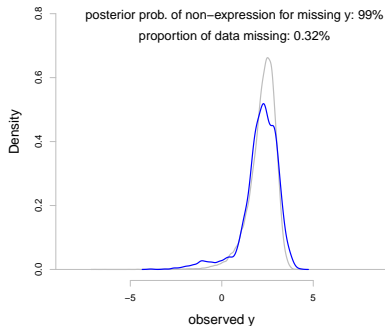
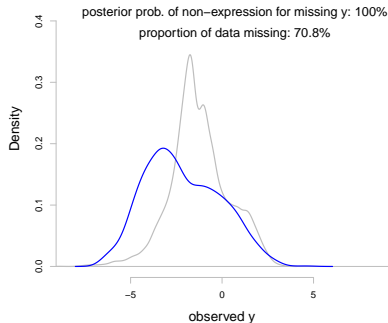


Figure 13: Posterior missing mechanism for simulation study in Project I for sample 1 (red), 2 (green), and 3 (blue). Prior missing mechanism in grey.

Posterior Predictive for Observed Data in Project I



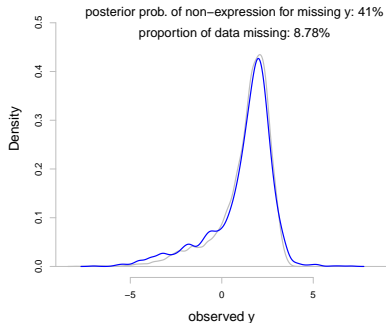
(a) Sample 1, marker 1



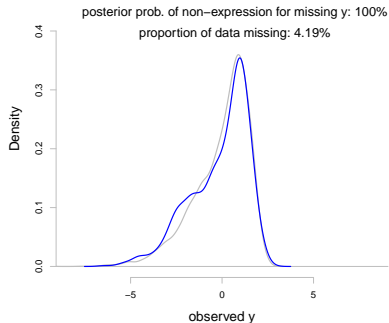
(b) Sample 1, marker 2

Figure 14: Observed data density for y_{ij} in grey. Posterior predictive density for observed data in blue.

Posterior Predictive for Observed Data in Project I



(c) Sample 2, marker 22



(d) Sample 3, marker 19

Figure 15: Observed data density for y_{ij} in grey. Posterior predictive density for observed data in blue.

Probability of Non-expression for Imputed Values in Project I CB Analysis

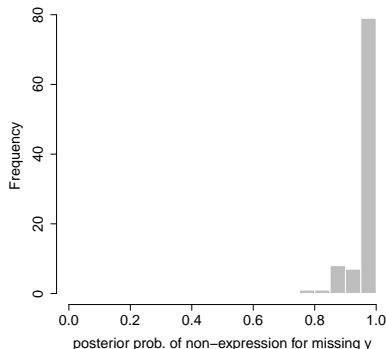
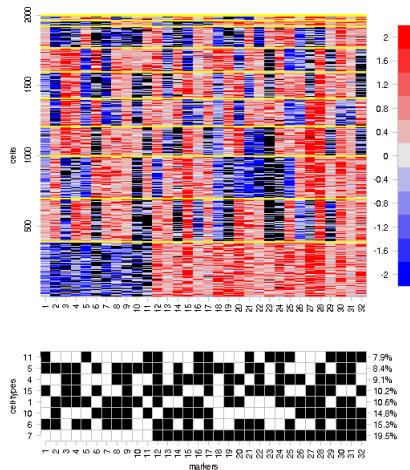
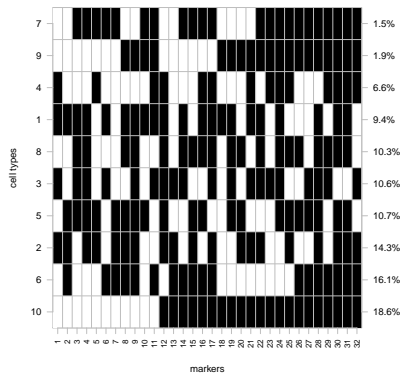


Figure 16: Histogram of posterior probabilities \hat{q}_{ij} of non-expression for missing for all (i, j) . The peak at the value of 1 suggests that most of the time, a marker is estimated as no expression if its expression level is missing.

Project I: Simulation Results – FAM



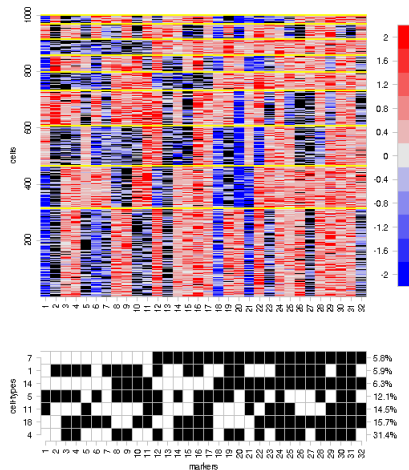
(a) Sample 2



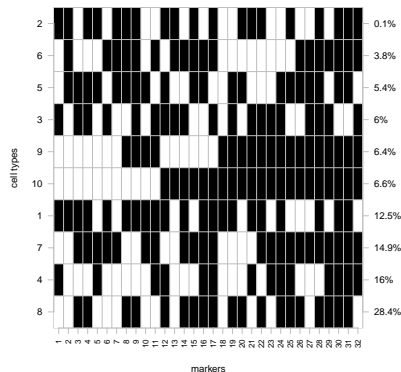
(b) Z true

Figure 17: FAM Simulation Study

Project I: Simulation Results – FAM



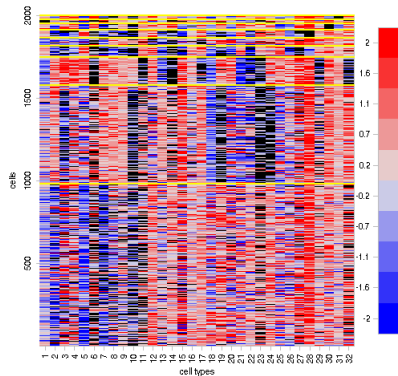
(a) Sample 3



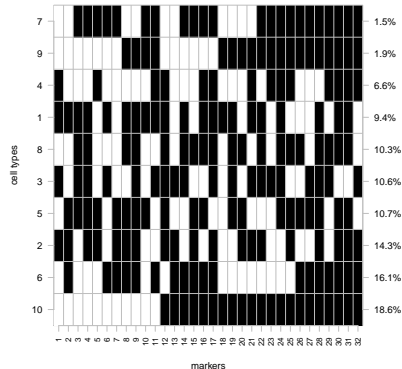
(b) Z true

Figure 18: FAM Simulation Study

Project I: Simulation Study – FlowSOM



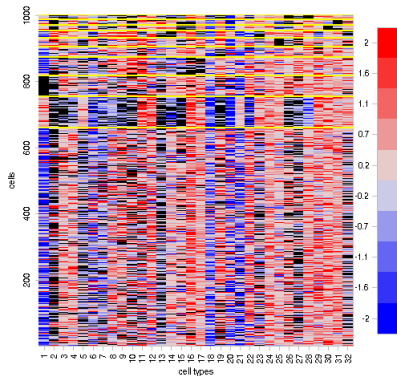
(a) Sample 2



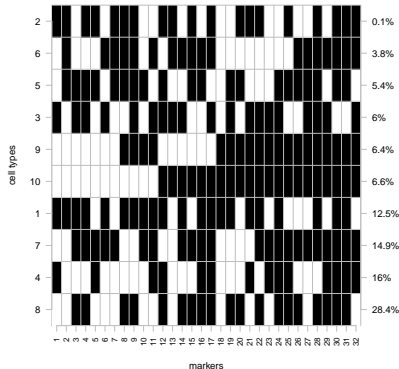
(b) Z true

Figure 19: FlowSOM Simulation Study

Project I: Simulation Study – FlowSOM



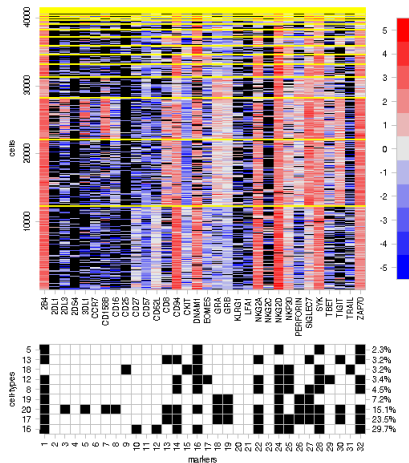
(a) Sample 3



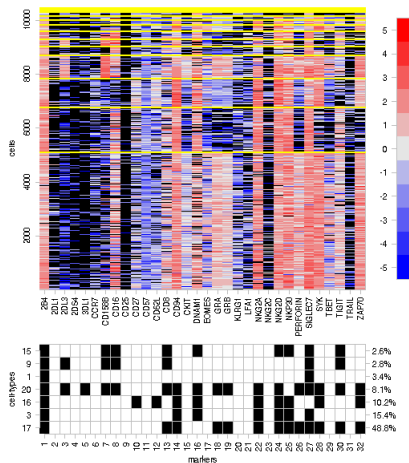
(b) Z true

Figure 20: FlowSOM Simulation Study

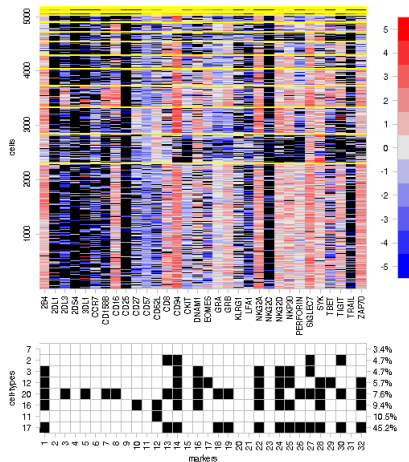
Project I: CB Study – FAM



(a) Sample 1



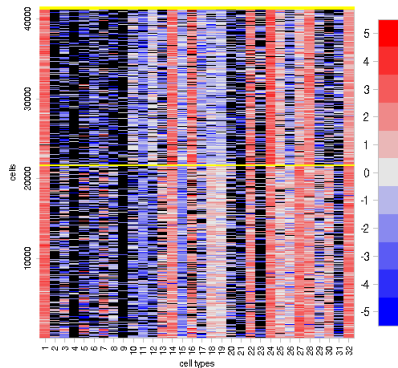
Project I: CB Study (FAM)



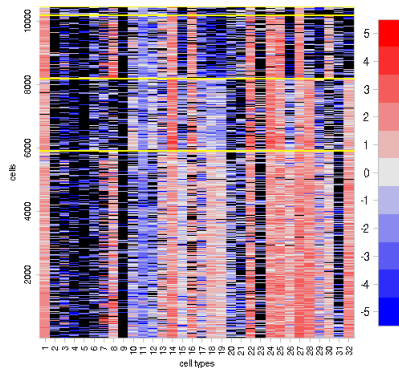
(c) Sample 3

Figure 22: CB data analyzed using FAM

Project I: CB Study – FlowSOM



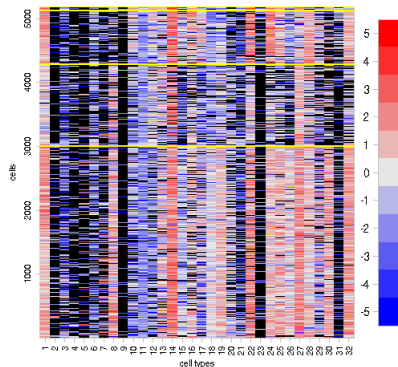
(a) Sample 1



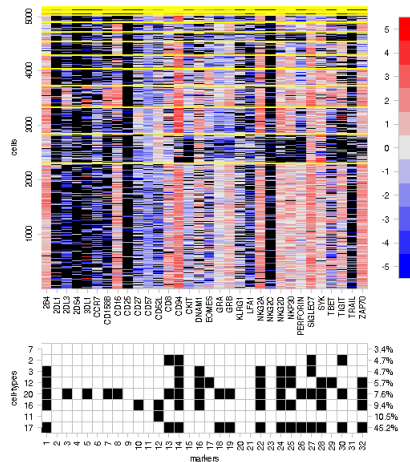
(b) Sample 2

Figure 23: CB data analyzed using FlowSOM

Project I: CB Study – FlowSOM



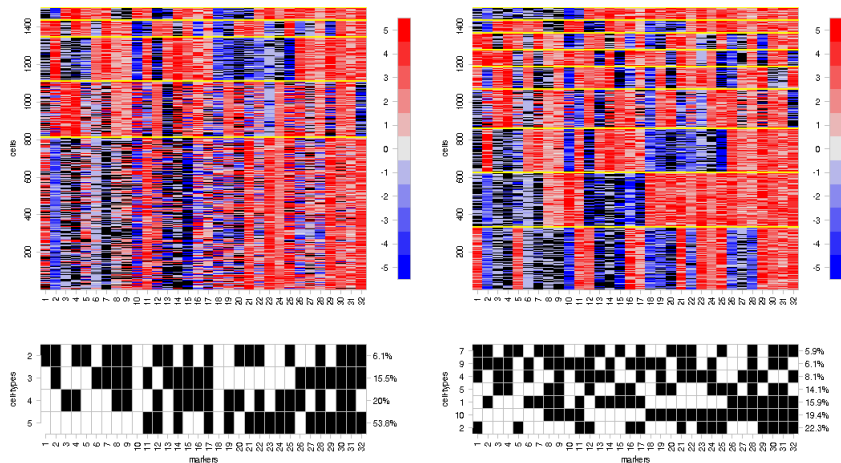
Sample 3 (FlowSOM)



Sample 3 (FAM)

Figure 24: CB data Sample 3, analyzed using FlowSOM (left) and FAM (right)

Project I: Simulation Study – Sensitivity to K

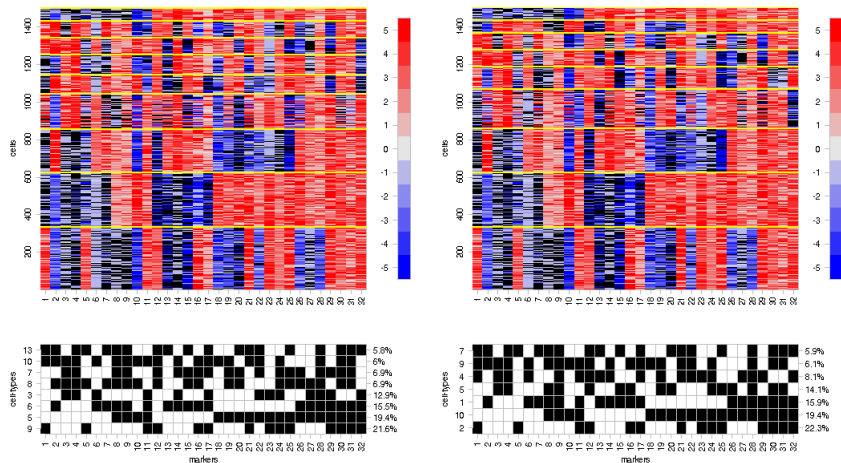


(a) Sample 1. $K = 5$

(b) Sample 1. $K = 10$

Figure 25: FAM sensitivity to K

Project I: Simulation Study – Sensitivity to K



(c) Sample 1. $K = 20$

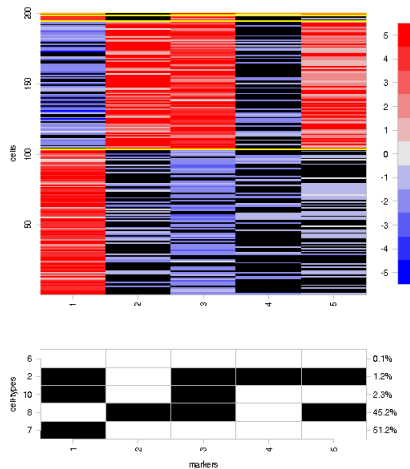
(b) Sample 1. $K = 10$

Figure 26: FAM sensitivity to K

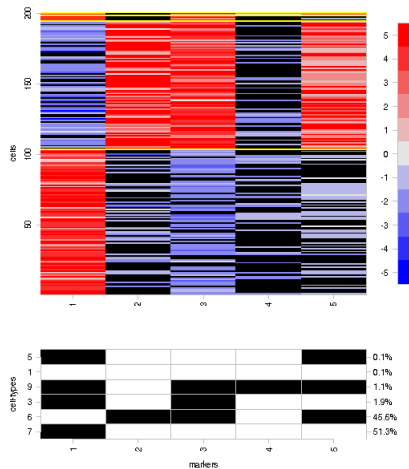
Project I: Conclusions for FAM's Sensitivity to K

- ▶ Choose K sufficiently large
- ▶ May graph first two principal components to visually estimate K

Project II: Simulation Study for rep-FAM Prior



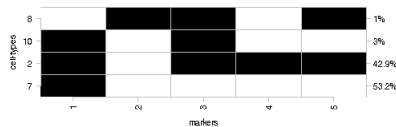
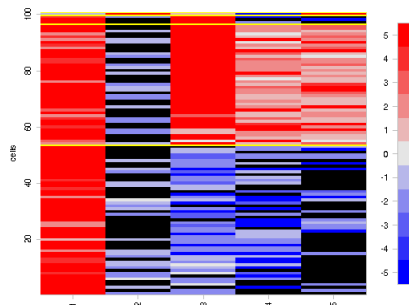
(a) rep-FAM



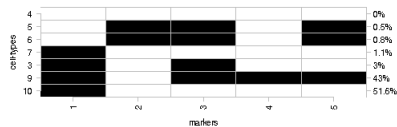
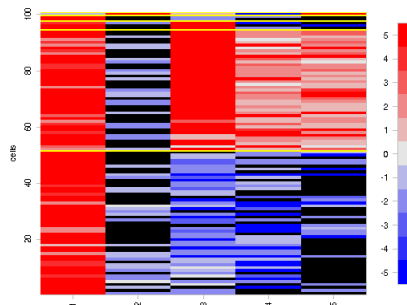
(b) IBP

Figure 27: Heatmaps of y for simulated data ordered by phenotypes (Sample 2)

Project II: Simulation Study for rep-FAM Prior



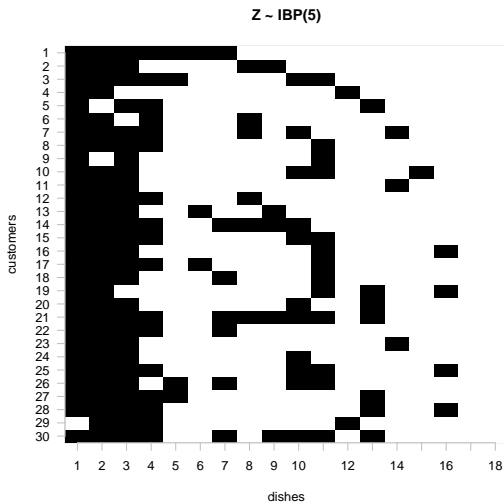
(a) rep-FAM



(b) IBP

Figure 28: Heatmaps of y for simulated data ordered by phenotypes (Sample 3)

Why is it called the Indian Buffet Process?



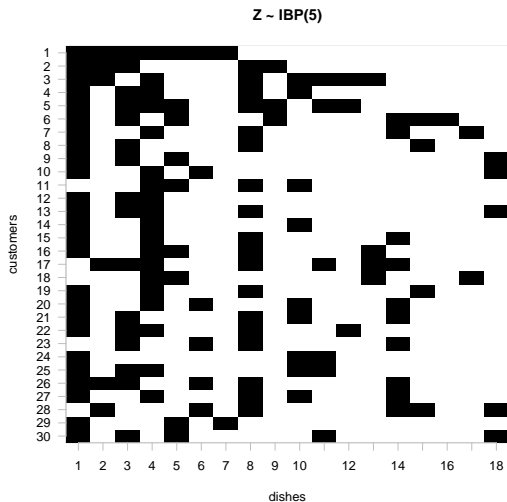
Why is it called the Indian Buffet Process?

In the IBP, a customer (j) taking a dish (k) is analogous to an observation possessing a feature. This is indicated by setting the value of z_{jk} to 1 if the customer takes the dish, and 0 otherwise. An $\text{IBP}(\alpha)$ for J observations can be simulated as follows:

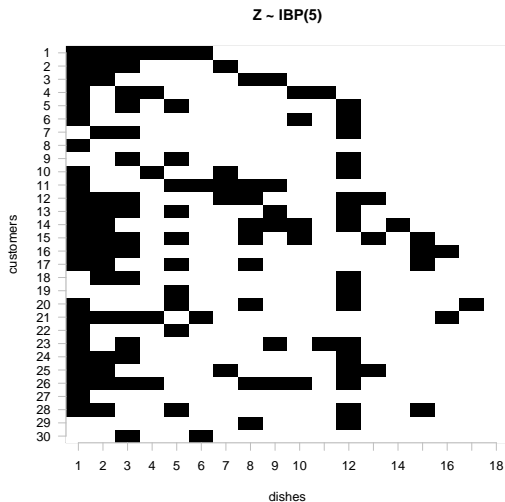
1. The 1st customer takes $\text{Poisson}(\alpha)$ number of dishes
2. For customers $j = 2$ to J ,
 - ▶ For each previously sampled dish, customer j takes dish k with probability m_k/j
 - ▶ after sampling the last sampled dish, customer j samples $\text{Poisson}(\alpha/j)$ new dishes

It can be shown that a matrix generated by this process has the same pmf up to a proportionality constant as the previous pmf.

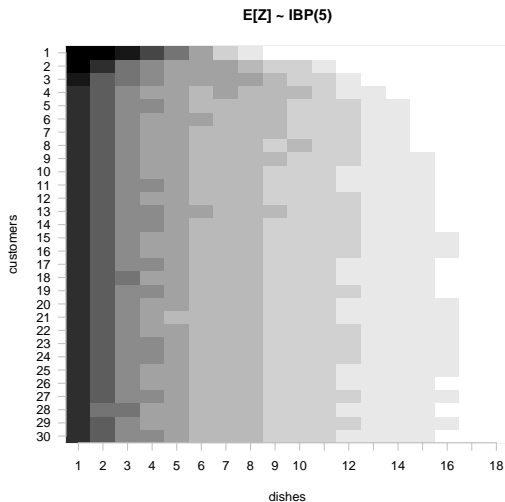
One draw from the IBP



One draw from the IBP



Expected value of the IBP



Project II: Simulation Study for rep-FAM Prior

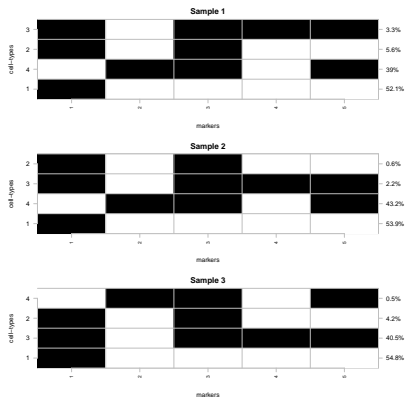


Figure 29: The transpose of \mathbf{Z}^{TR} with markers in columns and latent phenotypes in rows. Black and white represents $z_{jk}^{\text{TR}} = 1$ and 0 , respectively. The phenotypes and w_i^{TR} are shown on the left and right sides of each panel. All samples share the same \mathbf{Z}^{TR} and the phenotypes are arranged in order of w_{ik}^{TR} within each sample.

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).
- ▶ Binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).
- ▶ Binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k
- ▶ $\delta_{xk} \mid p_x \stackrel{ind}{\sim} \text{Bernoulli}(p_x)$ and $p_x \stackrel{iid}{\sim} \text{Beta}(a_p, b_p)$.

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).
- ▶ Binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k
- ▶ $\delta_{xk} \mid p_x \stackrel{ind}{\sim} \text{Bernoulli}(p_x)$ and $p_x \stackrel{iid}{\sim} \text{Beta}(a_p, b_p)$.
- ▶ Unnormalized cell phenotype abundances $\tilde{w}_{ik} \stackrel{ind}{\sim} \text{Gamma}(a_W/K, 1)$, $i = 1, \dots, I$, and $k = 1, \dots, K$.

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).
- ▶ Binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k
- ▶ $\delta_{xk} \mid p_x \stackrel{ind}{\sim} \text{Bernoulli}(p_x)$ and $p_x \stackrel{iid}{\sim} \text{Beta}(a_p, b_p)$.
- ▶ Unnormalized cell phenotype abundances $\tilde{w}_{ik} \stackrel{ind}{\sim} \text{Gamma}(a_W/K, 1)$, $i = 1, \dots, I$, and $k = 1, \dots, K$.
- ▶ Relative abundances in sample i from x_i as $w_{ik} = \tilde{w}_{ik} \delta_{x_i k} / \sum_{\ell=1}^K \tilde{w}_{i\ell} \delta_{x_i \ell}$.
- ▶ Relative abundance w_{ik} is exactly zero for $\delta_{x_i k} = 0$.

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).
- ▶ Binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k
- ▶ $\delta_{xk} \mid p_x \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p_x)$ and $p_x \stackrel{\text{iid}}{\sim} \text{Beta}(a_p, b_p)$.
- ▶ Unnormalized cell phenotype abundances $\tilde{w}_{ik} \stackrel{\text{ind}}{\sim} \text{Gamma}(a_W/K, 1)$, $i = 1, \dots, I$, and $k = 1, \dots, K$.
- ▶ Relative abundances in sample i from x_i as $w_{ik} = \tilde{w}_{ik} \delta_{x_i k} / \sum_{\ell=1}^K \tilde{w}_{i\ell} \delta_{x_i \ell}$.
- ▶ Relative abundance w_{ik} is exactly zero for $\delta_{x_i k} = 0$.
- ▶ Samples from x have the same subset of phenotypes but can have different relative abundances over the selected phenotypes.

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).
- ▶ Binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k
- ▶ $\delta_{xk} \mid p_x \stackrel{\text{ind}}{\sim} \text{Bernoulli}(p_x)$ and $p_x \stackrel{\text{iid}}{\sim} \text{Beta}(a_p, b_p)$.
- ▶ Unnormalized cell phenotype abundances $\tilde{w}_{ik} \stackrel{\text{ind}}{\sim} \text{Gamma}(a_W/K, 1)$, $i = 1, \dots, I$, and $k = 1, \dots, K$.
- ▶ Relative abundances in sample i from x_i as $w_{ik} = \tilde{w}_{ik} \delta_{x_i k} / \sum_{\ell=1}^K \tilde{w}_{i\ell} \delta_{x_i \ell}$.
- ▶ Relative abundance w_{ik} is exactly zero for $\delta_{x_i k} = 0$.
- ▶ Samples from x have the same subset of phenotypes but can have different relative abundances over the selected phenotypes.
- ▶ Phenotypes with $\delta_{0k} = \delta_{1k} = 1$ appear in all samples, while some are present in only one type of samples.

Project II: Feature Selection for Different Samples

- ▶ Sample covariate x_i . Cord blood samples ($x_i = 0$). Healthy subject samples ($x_i = 1$).
- ▶ Binary indicator $\delta_{xk} \in \{0, 1\}$, $x = 0, 1$, and $k = 1, \dots, K$, to indicate whether samples from x possess phenotype k
- ▶ $\delta_{xk} \mid p_x \stackrel{\text{iid}}{\sim} \text{Bernoulli}(p_x)$ and $p_x \stackrel{\text{iid}}{\sim} \text{Beta}(a_p, b_p)$.
- ▶ Unnormalized cell phenotype abundances $\tilde{w}_{ik} \stackrel{\text{iid}}{\sim} \text{Gamma}(a_W/K, 1)$, $i = 1, \dots, I$, and $k = 1, \dots, K$.
- ▶ Relative abundances in sample i from x_i as $w_{ik} = \tilde{w}_{ik} \delta_{x_i k} / \sum_{\ell=1}^K \tilde{w}_{i\ell} \delta_{x_i \ell}$.
- ▶ Relative abundance w_{ik} is exactly zero for $\delta_{x_i k} = 0$.
- ▶ Samples from x have the same subset of phenotypes but can have different relative abundances over the selected phenotypes.
- ▶ Phenotypes with $\delta_{0k} = \delta_{1k} = 1$ appear in all samples, while some are present in only one type of samples.
- ▶ The probability models for the other parameters remain unchanged as in Project 1.

Project III: Feature Allocation Model with Regression for Feature Abundances Over Time

- ▶ $w_{t_1,k} = \xi_{t_1,k} / \sum_{\ell=1}^K \xi_{t_1,\ell}$
- ▶ We fix $\xi_{t_1,1} = a$, an arbitrary positive number, to avoid potential identifiability issues, and let $\xi_{t_1,k} = \max(\xi'_{t_1,k}, 0)$, for $k \geq 2$, where $\xi'_{t_1,k} \stackrel{iid}{\sim} \mathcal{N}(0, s_1^2)$
- ▶ $\xi_{t_i,k} = \max(\xi'_{t_i,k}, 0)$, $i = 2, \dots, I$ and $k = 1, \dots, K$
- ▶ $\xi'_{t_i,k} = \xi'_{t_1,k} + f_k(t_i)$, where $f_k(t_i)$ is a phenotype-specific function of time
- ▶ One choice is $f_k(t) = \xi'_{t_1,k} + \beta_{k1}t + \beta_{k2}t^2$