Immune Profiling among Colorectal Cancer Subtypes using Dependent Mixture Models

Yunshan Duan[1], Shuai Guo[2], Wenyi Wang[2], and Peter Mueller[1, *]

[1]Department of Statistics and Data Science, University of Texas at Austin

[2]The University of Texas MD Anderson Cancer Center

[*]Corresponding author: Peter Mueller, pmueller@math.utexas.edu

Abstract

Comparison of transcriptomic data across different conditions is of interest in many biomedical studies. In this paper, we consider comparative immune cell profiling for early-onset (EO) versus late-onset (LO) colorectal cancer (CRC). EOCRC, diagnosed between ages 18-45, is a rising public health concern that needs to be urgently addressed. However, its etiology remains to be poorly understood. We work towards filling this gap by identifying homogeneous T cell subpopulations that show significantly distinct characteristics across the two tumor types, and to identify others that are shared between EOCRC and LOCRC. Such inference may reveal underlying determinants of clinically observed differences in the two disease subpopulations. We develop dependent finite mixture models where immune subtypes enriched under a specific condition are characterized by terms in the mixture model with common atoms but distinct weights across conditions, whereas common subtypes are characterized by sharing both atoms and relative weights. The proposed model defines a variation of mixtures of finite mixture models, facilitating the desired comparison by introducing highly structured multi-layer Dirichlet priors. The model allows us to explicitly compare features across conditions. We illustrate inference with simulation studies and data examples. Results identify EO-enriched and LO-enriched T cells subtypes whose biomarkers are found to be linked to mechanisms of tumor progression. The findings reveal distinct characteristics of the immune profiles in EOCRC and LOCRC, and potentially motivate insights into treatment and management of CRC.

1 Introduction

Early-onset colorectal cancer (EOCRC) refers to CRC diagnosis in individuals aged 18-45 (or ≤ 50) and is an increasing global health problem that needs to be urgently addressed [1]. An important gap in current research is that the differences in immune cell populations of

EOCRC and late-onset colorectal cancer (LOCRC) have not been well investigated [2]. We develop inference for immune profiling of EOCRC and LOCRC samples, aiming to identify, compare, and understand relations of immune cell subtypes across these two disease subpopulations. The problem is formalized as statistical inference in two dependent mixture models, where immune subtypes are represented as components in the mixture. The specific aim of identifying differences in immune subtype abundance between the two conditions is achieved by introducing a highly structured multi-level prior across the two linked models. Under each disease condition, the mixture model gives rise to a partition, i.e., a cluster arrangement of observed immune cells. Inference can thus alternatively be characterized as joint clustering of all immune cells under the two conditions, using synchronized random partitions.

We proceed under a model-based Bayesian paradigm. Bayesian inference for a mixture model $\sum_h w_h f(y \mid \theta_h)$ is often based on representing the model as a mixture of a kernel $f(y \mid \theta)$, e.g., a Gaussian kernel, with respect to a discrete mixing measure $G(\theta) = \sum_h \delta_{\theta_h}$ (here δ_x is point mass at x). The construction is motivated by Kingman's representation theorem [3] which shows that under certain symmetry conditions (exchangeability) any random partition can be thought of as arising this way. A prior probability model for the random mixing measure G completes the model construction. Priors for random distributions are known as nonparametric Bayes (BNP) models [4], and the by far most widely used BNP model is the Dirichlet process (DP) prior [4, Chapter 4]. We write $G \sim \mathrm{DP}(\alpha, G_0)$, where $\alpha > 0$ is a precision parameter and G_0 is the prior expectation of the random probability measure. One important property of the DP prior is the a.s. discrete nature of G, i.e., $G = \sum_h w_h \delta_{\theta_h}$. The DP provides a computationally attractive model for an unknown probability measure G. However, for inference on two related random mixtures, as needed in our motivating application, we need a prior probability model $p(G_1, G_2)$ for two dependent random probability measures G_1 and G_2, characterizing the immune subtypes under the two disease conditions. We build on an extensive literature on such models over the past 20 years. Some of the earliest constructions are additive decomposition models introduced in [5], which use a mixture of common components and idiosyncratic components, both defined by discrete random probability measures with DP priors. The strength of the dependence is controlled by the relative weight of common versus idiosyncratic parts. Kolossiatis et al. [6] introduced a restriction that ensures that the marginal prior on G_j remains a DP. Lijoi et al. [7] generalized such models using random probability measures constructed via completely random measures (CRM).

Other dependent models $p(G_j;\ j = 1, \ldots, J)$ that are built on DP priors include the hierarchical and the nested DP and variations thereof. Teh et al. [8] introduce the hierarchical DP (HDP) which constructs a probability model for dependent random measures G_j by linking submodels $G_j \sim \mathrm{DP}(M_j, G)$ with a common base measure G, which in turn is given a $\mathrm{DP}(M_0, G_0)$ hyperprior. In contrast to the HDP, the nested DP (NDP) proposed by Rodriguez

et al. [9] links the random probability measures G_j by assuming $G_j \sim Q$, using a common prior $Q = \sum_h \pi_h \delta_{G_h^\star}$. Here Q is a distribution of distributions. A sample G_j from Q selects one of the atoms $G_h^\star$ with probability π_h. Assuming a hyperprior $Q \sim \mathrm{DP}(\alpha, \mathrm{DP}(M, H))$ creates a discrete distribution Q with each atom $G_h^\star$ itself being generated by a DP, $G_h^\star \sim \mathrm{DP}(M, H)$. The latter implies marginally a DP prior $G_j \sim \mathrm{DP}(M, H)$. We shall later discuss and compare more properties of HDP and NDP, as they are relevant for our motivating application. A generalization of the HDP is studied in [10–12] by replacing the DP with more general normalized completely random measures (NRM). A completely random measure (CRM) is a random measure μ with the defining property that $X = \mu(A)$ and $Y = \mu(B)$ for two mutually exclusive events $A \cap B = \emptyset$ are independent random variables [13]. The hierarchical NRM includes the HDP and the hierarchical Pitman-Yor (PY) process, another variation with the same construction, as special cases. Camerlenghi et al. [14] extend the NDP by defining the latent nested process (LNP) as a nested process of NRM's, which is constructed by normalizing an additive structure including shared and idiosyncratic CRM components. More recently, Denti et al. [15] proposed a nested common atoms model (CAM), where different distributions are characterized by specific weights assigned to a common set of atoms. Yet another construction that combines elements of the HDP and NDP models is the hidden hierarchical Dirichlet process (HHDP) model of Lijoi et al. [16]. Like in the NDP they generate the G_j's from a prior Q. But the atoms $G_h^\star$ of Q are generated from a HDP, as discrete random probability measures with shared atoms. The HHDP blends features of the HDP and the NDP. Additionally, instead of assuming fully exchangeable data in the models above, models for partially exchangeable data have been investigated using nested clustering [17], and separate exchangeable structure [18].

Considering the immune profile comparison problem, differential immune cell behaviors are expected when comparing EO- with LOCRC, i.e., some immune subtypes are shared across conditions while some occupy a significantly larger proportion in one condition than the other. Suitable inference models should therefore include shared immune subtypes with common weights and common atoms across all conditions, and idiosyncratic but not exclusive components with shared atoms, but distinct weights. The previously summarized constructions do not introduce and favor such structure, making them less suitable for the application in comparative immune profiling. One notable limitation is that the idiosyncratic components that appear in some of the constructions are independent, lacking clusters with common atoms but distinct weights. The HDP and related priors, as well as the CAM and HHDP model do not allow for components with common weights and atoms. While the CAM and the HHDP model include common atoms across all conditions, there is no structure in the model to favor a distinction into shared clusters versus clusters that are enriched under different conditions. Finally, it is natural to assume a finite but unknown number of immune subtypes, advising against the use of infinite mixture models.

The discussed limitations prompt us to consider alternatives. We build on finite mixture models. Recent discussions of Bayesian inference for mixtures of finite mixtures (MFM) appear in Frühwirth-Schnatter et al. [19], Greve et al. [20] and Miller and Harrison [21]. These constructions create a single random probability model G as a finite mixture of random size K. MFM models appear to be more suitable than DP models in many examples and offer many of the same features. For example, when focusing on the implied clustering of experimental units, the probability function of the implied random partition is characterized by exchangeable partition probability functions (EPPF) with similar properties as under corresponding BNP mixtures. A natural extension of MFM to be suitable for the motivating application involves multiple, related mixtures.

In this paper, we introduce a structured dependent finite mixture model, including model features that anticipate the desired inference in the comparative immune profiling problem. We construct the model using an additive structure of common components and components that are specific to each condition. The model is useful for any inference problem concerned with clustering of features and comparison across different conditions.

The rest of this paper is organized as follows. Section 2 describes the motivating immune profiling problem and the data sources. We describe explanatory data analysis and preprocessing steps for data cleaning and dimension reduction. In Section 3.1, we introduce a dependent finite mixture model ("888 model") for a general inference problem. In Section 3.2 we introduce a specific implementation for the motivating application on immune profile comparison between EO and LO-CRC. We discuss the differences and commonalities with alternative models in Section 3.3. More properties and details of posterior inference are included in Section 4.1 and 4.2. In Section 5, we discuss simulation studies to validate finite sample performance of the model. Section 6 shows the results for the motivating comparative immune profiling problem. The findings in immune subtypes, especially the ones that are enriched under a certain condition reveal some underlying biological paradigm for the disease subtype.

2 Immune profiling of EOCRC vs. LOCRC

Recent studies have identified distinctive clinical features in EOCRC patients aged 18-45 (or $\leq$ 50) compared to LOCRC patients over 60 [1]. Also, it is known that the immune cells within the tumor microenvironment (TME) profoundly influence tumor development, progression, and response to treatment in CRC patients [22, 23]. This leads us to hypothesize that the immune microenvironment of EOCRC differs from that of LOCRC and promotes tumor progression. To further investigate the distinct immune microenvironment in patients with EOCRC, we develop a full probabilistic description of comparative immune profiling between EO and LO conditions. Such inference will help reveal some of the underlying determinants for the distinct

etiology of EOCRC and for the potentially different treatment response between the two disease conditions.

Our analysis is based on single-cell RNA sequencing data obtained from Zhang et al. (2018) [22], which was sequenced using the Smart-seq2 platform (see **Appendix A**). This study only sequenced T cells, allowing us to focus on investigating the subtypes of T cells. Because CRC samples with microsatellite instability (MSI) have distinct immune profiles compared to microsatellite stable (MSS) CRC samples and are known to be more responsive to checkpoint inhibitor (CPI) treatment, we further removed tumor samples with MSI [24]. We followed the standard analyzing pipeline of the *Seurat* package for data preprocessing. Only genes that were detected in a minimum of three cells were retained. Next, we excluded low-quality cells by setting thresholds for the number of expressed genes and detected Unique Molecular Identifier (UMI) counts: cells were retained if they expressed between 500 and 6,500 genes and had UMI counts ranging from 500 to 2,000,000. Taking into account the information provided by the original paper [22], our study focused on CD4+ and CD8+ cells in tumors and adjacent normal tissues, while other T cell types (such as double-negative and double-positive cells) and cells originating from the peripheral blood were filtered out. After these quality control steps, we were left with a total of 3,139 CD4+ and CD8+ cells from seven CRC patients, including two EO and five LO patients. For data visualization we used Uniform Manifold Approximation and Projection (UMAP) transformations implemented with the *Seurat* package, as shown in Figure 1 (a)-(b). In this figure, panel (a) organizes the data by disease condition (EO vs. LO), and panel (b) does so by the type of cell (CD4+ vs. CD8+).

As part of exploratory data analysis, we first carried out clustering as implemented in the *Seurat* package, setting the resolution at 0.75. We annotated T cells into immune subtypes, as illustrated in Figure 1 (c). The annotation is based on matching markers for the known T-cell sub populations. CD4+ cells were annotated into three subtypes: regulatory T cells, helper T cells, and other CD4+ T cells. Similarly, CD8+ T cells were annotated into three subtypes: effective memory T cells, exhausted T cells, and other CD8+ T cells. The relative proportions of EO and LO conditions within each immune cell subtype cluster are depicted in Figure 1 (d). The results suggest a potential correlation between cell subtypes and onset conditions, which was formally tested using Fisher's exact test. We observed a highly significant correlation (p=0.0005), indicating that the data contains sufficient information for the desired comparative immune profiling. However, an omnibus test of independence does not provide information about which subtypes are shared or characteristic for each condition. We also implemented k-Means clustering [25] (results not shown), which, however, also stops short of detecting interpretable immune cell sub-populations for the comparison between EO and LO conditions. In summary, these preliminary analyses indicate the feasibility of inferring disparities in immune cell sub-populations between EO and LO conditions and demonstrate

the need for an inference approach specifically aiming at identifying comparative immune subtypes between conditions.

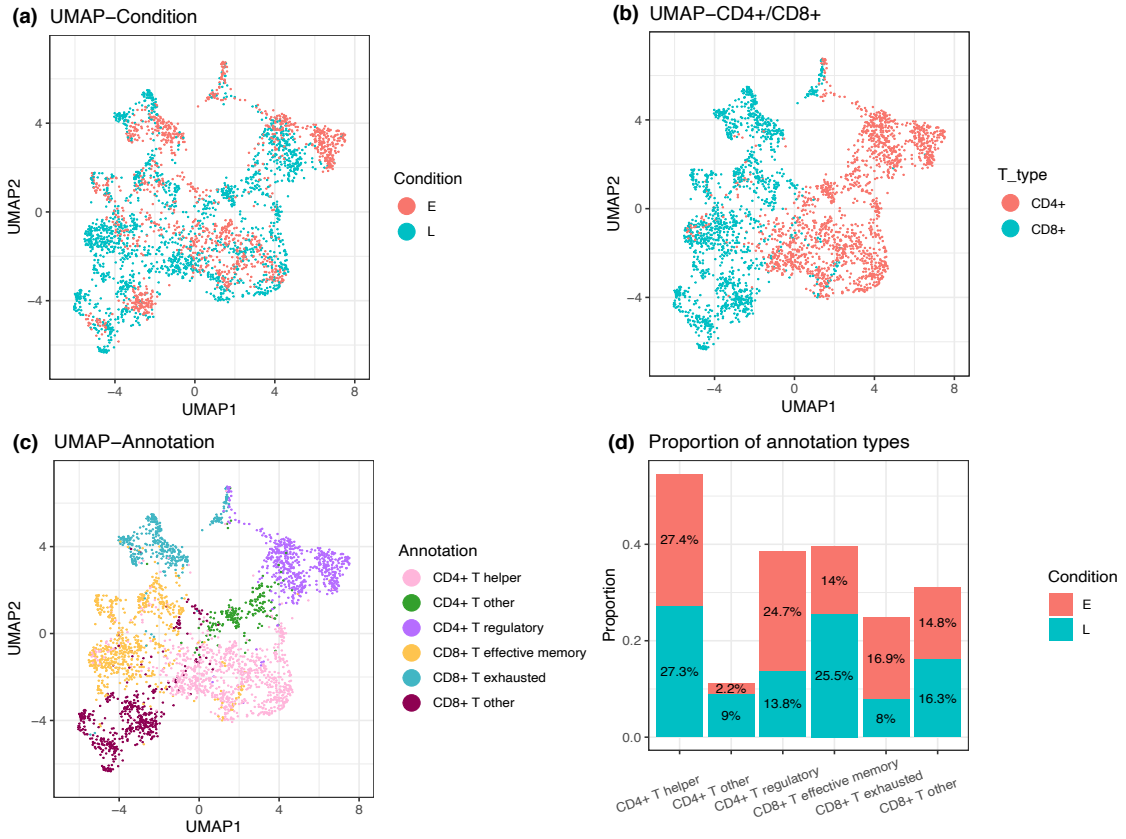

Figure 1: Single-cell RNA-seq data from CRC samples in UMAP plots, colored by EO/LO conditions (panel a), T cell CD4+/CD8+ types (panel b), and cell type annotations (panel c). Panel (d) shows the proportions of annotated cell types under EO and LO condition.

3 Dependent finite mixture model

3.1 888 model

We construct a dependent finite mixture model for comparative immune cell profiling to identify homogeneous immune cell sub-populations. Let y_{ji} denote gene expression features observed for cell $i = 1, \ldots, N_j$, under condition $j = 1, \ldots, J$. We develop the model for a more general case with $J > 1$ conditions, keeping in mind that we have $J = 2$ for the motivating application. The proposed probability model is defined as follows. The features y_{ji} are generated from finite mixtures $\int f(\cdot;\ \theta)\ dG_j(\theta)$ with respect to a finite, discrete mixing measure G_j, $j = 1, \ldots, J$, using, for example, a normal kernel $f(\cdot;\ \theta)$. Writing the mixture as an equivalent hierarchical

model with latent variables θ_{ji} we have:

$$y_{ji} \mid \theta_{ji} \sim f(y_{ji}; \theta_{ji}), \text{ and } \theta_{ji} \mid G_j \sim G_j, \tag{1}$$

$i = 1, \ldots, N_j$ and $j = 1, \ldots, J$. The desired dependence across j is induced by a hierarchical prior on G_j

$$G_j = \epsilon F_0 + (1 - \epsilon) F_j, \quad j = 1, \ldots, J, \tag{2}$$

In words, the probability measure G_j for subjects under condition j is decomposed into a shared component F_0 and condition-specific components F_j. In anticipation of the comparative immune profiling problem, the latter is constructed to include some parts that are interpretable as sub-populations enriched under condition j, and others that are interpretable as enriched under conditions $\ell \neq j$. This is implemented by representing F_j as a mixture of distributions $F_\ell^{j\star}$, $\ell = 1, \ldots, J$, including distributions that characterize the subtypes that are enriched under condition j itself ($F_j^{j\star}$), and in other conditions $\ell \neq j$ ($F_\ell^{j\star}$), respectively:

$$F_j = \sum_{\ell=1}^{J} w_\ell^j F_\ell^{j\star}, \quad F_\ell^{j\star}(\theta_{ji}) = \sum_{h=1}^{H_\ell} u_{\ell h}^j \delta_{\phi_{\ell h}}.$$

The w_ℓ^j define the relative weights of components enriched under different conditions, and the $u_{\ell h}^j$ define the relative weights of different sub-populations that are specific to condition ℓ. A prior on the relative weights w_ℓ^j leads to this desired interpretation by imposing that (*a priori*) $w_j^j \gg w_\ell^j$, $\ell \neq j$. That is, all conditions share the same subtypes with the same atoms (cluster-specific parameters ϕ_{lh}) but at different levels of relative abundance (w_ℓ^j). *A priori* under condition j, components in $F_j^{j\star}$ have a significantly higher abundance than components in $F_\ell^{j\star}$, ($\ell \neq j$). Similar to the $F_\ell^{j\star}$, the shared measure F_0 is also assumed to be discrete with finitely many atoms:

$$F_0(\theta_{ji}) = \sum_{h=1}^{H_0} u_{0h} \delta_{\phi_{0h}},$$

now with common (relative) weights, shared across all conditions, $j = 1, \ldots, J$. Together the weights ϵ, w_ℓ^j, and $u_{\ell h}^j$ (u_{0h}) define a prior on the set of weights under conditional j for all unique cell sub-populations. The purpose of the hierarchical split into the three sets of weights is to allow the prior to favor a structure that anticipates the desired interpretations. In other words, the purpose of the structured hierarchical prior on the weights is prior shrinkage towards an interpretable structure with the split into sub-populations enriched under one or the other condition etc. However, such structure is only favored, not enforced. It is possible that *a posteriori* some of the idiosyncratic clusters are estimated with practically equal weight under different conditions, making them effectively shared clusters.

The mixture model induces a random partition of cells. Introducing latent cluster membership indicators c_{ji} and s_{ji} the model can alternatively be written as follows. First, there is

a mixture of $f(y_{ji};\phi_{lh})$ with respect to $F_\ell^{j\star}$ characterized by s_{ji}:

$$y_{ji} \mid c_{ji} = \ell, s_{ji} = h \sim f(y_{ji};\phi_{\ell h}), \quad \ell = 0, 1, \ldots, J,\ h = 1, \ldots, H_\ell, \tag{3}$$

$$p(s_{ji} = h \mid c_{ji} = \ell) = \begin{cases} u_{0h} & \text{if } \ell = 0 \\ u_{\ell h}^j & \text{if } \ell = 1, \ldots, J \end{cases} \tag{4}$$

representing the mixture with respect to the H_ℓ components of $F_\ell^{j\star}$ (when $c_{ji} = l > 0$), and with respect to the H_0 components of F_0 (if $c_{ji} = 0$). Another mixture occurs with the choice of the type of component ℓ in

$$p(c_{ji} = \ell) = \begin{cases} \epsilon, & \ell = 0, \\ (1-\epsilon)w_\ell^j, & \ell \neq 0, \end{cases} \tag{5}$$

where c_{ji} is the indicator for y_{ji} arising from a common cluster ($\ell = 0$) or an idiosyncratic cluster ($\ell \neq 0$). A third level of mixing could arise with possibly random H_ℓ and H_0.

To ensure the structure of the different terms in the mixture, we introduce an asymmetric Dirichlet prior on w_ℓ^j:

$$\begin{aligned} \boldsymbol{w}^j &= (w_1^j, \ldots, w_J^j) \sim \mathrm{Dir}(\gamma_1^j, \ldots, \gamma_J^j), \\ \gamma_\ell^j &= \begin{cases} \gamma^{(H)}, & \ell = j, \\ \gamma^{(L)}, & \ell \neq j, \end{cases} \end{aligned} \tag{6}$$

with $\gamma^{(H)} \gg \gamma^{(L)}$, allowing *a priori* $w_j^j \gg w_\ell^j$, $\ell \neq j$. A detailed explanation on how the prior model shrinks the abundance of clusters and introduces the desired structure for our inference goal appears in the specific example of immune profiling, and is included in Section 3.2. We assume symmetric Dirichlet and beta priors on other weight parameters.

$$\begin{aligned} \epsilon &\sim \mathrm{Beta}(a_0, b_0), \\ \boldsymbol{u}^0 &= (u_{01}, \ldots, u_{0H}) \sim \mathrm{Dir}(\alpha_0), \\ \boldsymbol{u}_\ell^j &= (u_{\ell 1}^j, \ldots, u_{\ell H}^j) \sim \mathrm{Dir}(\alpha_\ell^j), \end{aligned} \tag{7}$$

where $\mathrm{Dir}(\cdot)$ with one parameter denotes a symmetric Dirichlet prior. Together (6) and (7) define for each condition j a prior on the set of weights for all unique atoms. The prior is defined hierarchically to allow the desired structure across conditions.

We refer to the model as "888 model", keeping in mind the case with 2 conditions and 8 clusters in each of the three cluster types, i.e., $H_0 = H_1 = H_2 = 8$. Note that the number of clusters H_j's can be random, or any fixed number in the general "888 model", despite the choice of the name.

3.2 Comparative immune profiling for EO- versus LOCRC

We discuss specific model details in the implementation for the motivating application to immune profiling. For easier notation and without loss of generality, assume $H_0 = H_\ell = H$, and consider the $J = 2$ conditions as in the CRC immune subtypes example, with $j = 1, 2$ denoting the LOCRC and EOCRC condition, respectively. The dependent mixture model can be expressed in the following form,

$$\begin{aligned} & \qquad\qquad \text{Common clusters} \qquad\qquad\qquad \text{LO clusters} \qquad\qquad\qquad \text{EO clusters} \\ \text{LOCRC: } y_{1i} &\sim \epsilon \sum_{h=1}^{H} u_{0h} N(y_{1i}; \phi_{0h}) + (1-\epsilon) \left[w_1^1 \sum_{h=1}^{H} u_{1h}^1 N(y_{1i}; \phi_{1h}) + w_2^1 \sum_{h=1}^{H} u_{2h}^1 N(y_{1i}; \phi_{2h}) \right], \\ \text{EOCRC: } y_{2i} &\sim \epsilon \sum_{h=1}^{H} u_{0h} N(y_{2i}; \phi_{0h}) + (1-\epsilon) \left[w_1^2 \sum_{h=1}^{H} u_{1h}^2 N(y_{2i}; \phi_{1h}) + w_2^2 \sum_{h=1}^{H} u_{2h}^2 N(y_{2i}; \phi_{2h}) \right], \end{aligned} \tag{8}$$

where $\phi_{\ell h} = \{\xi_{\ell h}, \Sigma_{\ell h}\}$, $l = 0, 1, 2$, are the cluster-specific parameters, i.e., the mean and covariance matrix of the Gaussian distributions. The first H clusters are the common immune subtypes, the second set of H clusters are the LO-enriched subtypes, and the last H clusters are the EO-enriched subtypes. The atoms (cluster-specific parameters) ϕ_{0h}, and $\phi_{\ell h}$ are shared between EOCRC and LOCRC conditions. Thus, the immune subtypes identified by the proposed mixture model share the same gene expression profile between EO and LO conditions, for both common and idiosyncratic clusters. The shared atoms in the idiosyncratic components in F_j are critical to allow for the desired comparison of immune cell profiles across conditions.

We assume that the common immune subtypes share common weights ϵu_{0h}, while condition-specific clusters are characterized by distinct weights $(1-\epsilon) w_\ell^j u_{\ell h}^j$, with *a priori* $w_j^j \gg w_\ell^j$, $\ell \neq j$, given by the prior. The EO/LO-enriched immune subtypes are prominent in samples of the matching condition, but not exclusively. In other words, EO-enriched clusters tend to have high weights in the EOCRC immune cell populations and low weights that are shrunk towards 0 in LOCRC immune cells, and vice versa for LO-enriched clusters.

To achieve such structure of comparative clusters in the model, an informative asymmetric beta prior is put on the idiosyncratic cluster weights w_ℓ^j for LO condition ($j = 1$) and EO ($j = 2$) condition, $\boldsymbol{w}^1 = (w_1^1, w_2^1) \sim Beta(\gamma^{(H)}, \gamma^{(L)})$, $\boldsymbol{w}^2 = (w_1^2, w_2^2) \sim Beta(\gamma^{(L)}, \gamma^{(H)})$, which is a special case of equation (6) when $J = 2$. We use a small value $\gamma^{(L)}$ for the beta parameters corresponding to the cluster weights to be shrunk to 0, and a large $\gamma^{(H)}$ for those weights to be shrunk away from 0. Symmetric priors for other relative weights, u_{0h}, u_{lh}, and ϵ, are specified in equation (7).

We use $\gamma^{(H)} = 10$, $\gamma^{(L)} = 1$, $\alpha^0 = \alpha_\ell^j = 1$, $j = 1, \ldots, J$, $a_0 = 10$, and $b_0 = 10$, in the simulation studies and data analysis, which serves as suitable prior hyperparameters for the desired structure. For example, under the EOCRC condition, the weights of LO-enriched clusters w_1^2 are shrunk to 0, and the sum of weights in EO-enriched clusters is close to 1. This is consistent with our prior assumption of the immune subtypes that we are interested in.

For other parameters in the model, we assume conditionally conjugate priors on the cluster-specific parameters, $\phi_{\ell h} \overset{\text{iid}}{\sim} \Phi(\phi_{\ell h})$. We use $\xi_{\ell h} \overset{\text{iid}}{\sim} N(\mu_0, \Psi_0)$ and $\Sigma_{\ell h} \overset{\text{iid}}{\sim} IW(\nu_0, \Phi_0)$ for $l = 1, 2$, $h = 1, \ldots, H$ in our examples. One could add a hyperprior on H if desired. In the application we choose to proceed with fixed H, to represent strong investigator preference for a moderate interpretable number of clusters for each condition.

3.3 Alternative models

Several similar inference approaches for dependent random probability models have been proposed in the recent literature, including the references that were already briefly cited in the introduction. We discuss some more details, including limitations that are relevant for the inference on immune cell subtypes.

Some alternative models for dependent random probability measures build on an additive structure where the discrete random measure G_j for condition j is a weighted mixture of DPs. The probability model for G_j is composed of a common component H_0 and an idiosyncratic random measure H_j,

$$G_j = \epsilon H_0 + (1 - \epsilon) H_j. \tag{9}$$

The model is completed with independent DP priors for H_j, $j = 0, 1, \ldots, J$. Such models were considered in [5] and further developed in [6] by restricting the prior probability model for ϵ to ensure marginal DP priors for G_j, and in [7] by assuming that H_0 and H_j arise by normalizing compeletly random measures (CRMs), including the DP as a special case. The common characteristics of these additive structure models is that the models across conditions share a common discrete measure H_0 implying shared atoms and proportional weights for the common components across all conditions, as desired for the motivating application. However, the idiosyncratic components are entirely different, with no notion of shared atoms. This is not consistent with the structure of homogeneous immune cell sub-populations that are shared across conditions, even if they are characteristic and enriched under one or the other condition.

The hierarchical DP (HDP) model proposed by Teh et al. [8] defines the random probability measures G_j as DP random measures, with a shared base measure G, which in turn is generated by another DP hyperprior:

$$G_j \mid M_j, G \overset{\text{iid}}{\sim} \text{DP}(M_j, G) \text{ and } G \mid M_0, G_0 \sim \text{DP}(M_0, G_0).$$

Camerlenghi et al. [10] generalize the HDP to hierarchical normalized random measures (NRM), with G_j and G defined as NRMs,

$$G_j \mid G \overset{\text{iid}}{\sim} \text{NRM}(G) \text{ and } G \mid G_0 \sim \text{NRM}(G_0). \tag{10}$$

including the HDP as a special case with a particular choice of the CRM. This type of hierarchical construction generates discrete random measures with all shared atoms across conditions

as in our proposed 888 model. However, although all atoms are shared across G_j, actual allocated clusters, i.e., atoms that are (latently) linked with data, need not be shared. Also, none of the components in the model share weights, and the models with exchangeable components lack the structure of the weights that identifies the common and condition-specific immune subtypes.

The nested DP (NDP) proposed by Rodriguez et al. [9] assumes the random measures G_j are samples from a distribution Q of distributions, that is $Q = \sum_{h=1}^{\infty} \pi_h \delta_{G_h^\star}(\cdot)$ with the weights generated by stick breaking with total mass parameter α, and the $G_h^\star$ being generated by a DP prior, $G_h^\star \sim \mathrm{DP}(M, H)$. In short,

$$G_j \mid Q \overset{\text{iid}}{\sim} Q = \sum_{h=1}^{\infty} \pi_h \delta_{G_h^\star}(\cdot) \text{ and } Q \sim \mathrm{DP}\{\alpha, \mathrm{DP}(M, H)\}.$$

The nested process (NP) introduced in Camerlenghi et al. [14] constructed the same nested structure with CRM's. Both NDP and NP generate random measures G_j that are either identical, sharing all the weights and atoms, or with all distinct weights and atoms. If θ_{ji} generated from G_j share at least one value across two conditions $j_1 \neq j_2$, then the posterior degenerates on the case $G_{j_1} = G_{j_2}$. This would be inappropriate for an application with common immune subtypes across all conditions. Recognizing this posterior degeneracy, Camerlenghi et al. [14] introduced the latent nested process (LNP), which is an additive structure, composed of a common component from a CRM, and an idiosyncratic component from a NP. This model tackles the degeneration problem of NDP and NP mentioned above, and allows for some shared clusters across conditions. But the idiosyncratic components can only be either identical or have all distinct weights and atoms. The model does not allow for clusters with common atoms and distinct weights, and therefore fails to accommodate inference for differential abundance of shared sub-populations.

The hidden hierarchical Dirichlet process (HHDP) model introduced by Lijoi et al. [16] builds on the NDP structure, but assumes an HDP prior for the $G_h^\star$. Replacing independent sampling from $\mathrm{DP}(M, H)$ with non-atomic H by the HDP prior allows for shared atoms across different $G_h^\star$.

$$G_j \mid Q \overset{\text{iid}}{\sim} Q = \sum_{h=1}^{\infty} \pi_h \delta_{G_h^\star}(\cdot),$$
$$(\pi_h)_{h\geq 1} \sim \mathrm{GEM}(\alpha), \quad (G_h^\star)_{h\geq 1} \sim \mathrm{HDP}(M, M_0, G_0),$$

Here GEM stands for the distribution of probability weights after Griffiths, Engen, and McCloskey, which characterizes the stick-breaking construction in the DP prior. The distinct distributions $G_h^\star$ are dependent and have a common base measure leading to shared atoms across $G_h^\star$'s. It overcomes the degeneracy problem of the NDP but introduces no structure

to distinguish the clusters by the characteristics of the weights in HHDP, so that the clusters could not be identified as common or specific to a certain condition by the construction of the model.

4 Properties of the 888 model

4.1 Random partition probability function and large sample properties

The mixture model in equation (1) induces a random partition of the experimental units – in our case, the immune cells. Given the latent cluster assignment indicator $\boldsymbol{c} = (c_{ji})_{j=1,\ldots,J,\ i=1,\ldots,N_j}$ for common ($c_{ji} = 0$), or condition ℓ enriched clusters ($c_{ji} = \ell \neq 0$), the probability function of the random partition for the 888 model is characterized by a partially exchangeable partition probability function (pEPPF) [10]. The pEPPF is a function of cluster sizes that is symmetric with respect to all clusters for each of the cluster types defined by $\boldsymbol{c}$ and satisfies a coherence condition across sample sizes. To derive the probability function, note that $F_\ell^{j\star}$ and F_0 are the discrete random mixing measures in an MFM. Let $\text{Dir}_K(\gamma)$ denote a symmetric $K-$dimensional Dirichlet $\text{Dir}(\gamma, \ldots, \gamma)$ distribution. Consider a generic MFM model,

$$y_i \overset{\text{iid}}{\sim} \sum_{k=1}^{K} \eta_k f(y_i;\ \phi_k), \tag{11}$$

$i = 1, \ldots, N$, with $p(\eta_1, \ldots, \eta_K) = \text{Dir}_K(\gamma)$ and a prior $p(K)$ on the size of the mixture. Let $\boldsymbol{n} = (n_1, \ldots, n_{K^+})$ denote the cardinalities of $K^+ \leq K$ clusters implied by the ties of $N = \sum_k n_k$ i.i.d. samples from (11). The random partition of the N experimental units (for example, cells), is characterized by the EPPF [19, 21],

$$\Pi_{K_+}^{(N)}(\boldsymbol{n}; \gamma) = V(N, K_+, \gamma) \prod_{j=1}^{K_+} \frac{\Gamma(n_j + \gamma)}{\Gamma(\gamma)}. \tag{12}$$

where

$$V(N, k, \gamma) = \sum_{K=k}^{\infty} p(K) \frac{K!\Gamma(\gamma K)}{(K-k)!\Gamma(\gamma K + N)}.$$

The random partition probability function for the 888 model can be derived based on (12) for each of the implied MFM models, and marginalizing out the ϵ and w weight parameters. We need some notation. For given data, let $H_{\ell+}^j$ (H_{0+}) denote the number of clusters created by ties of sampling latent parameters from $F_\ell^{j\star}$ (F_0). Let $n_{\ell h}^j$ be the number of units (cells, in our case) from condition j allocated to the hth cluster enriched under condition ℓ, $j = 1, \ldots, J$, $\ell = 1, \ldots, J$, $h = 1, \ldots, H_{\ell+}^j$, and $n_{\ell\bullet}^j = \sum_{h=1}^{H_{\ell+}^j} n_{\ell h}^j$ be the total number of units from condition j allocated to clusters enriched under condition ℓ. Similarly, let n_{0h} denote the number of units

from all conditions being allocated to the $h-$th cluster of the common clusters, $h = 1, \ldots, H_{0+}$, and let $n_{0\bullet}$ be the total number of units allocated to common clusters.

Proposition 1 *The pEPPF given the cluster type indicator* $\boldsymbol{c}$ *is*

$$\Pi(\boldsymbol{n}_0, \{\boldsymbol{n}_\ell^j\}_{\ell=1,\ldots,J,\ j=1,\ldots,J} \mid \boldsymbol{c}) = \Pi_{H_{0+}}^{n_{0\bullet}}(\boldsymbol{n}_0; \alpha_0) \cdot \prod_{j=1}^{J}\prod_{\ell=1}^{J}\left\{\Pi_{H_{\ell+}^j}^{n_{\ell\bullet}^j}(\boldsymbol{n}_\ell^j; \alpha_\ell^j)\right\}.$$

The random partition probability function for the 888 model is

$$\begin{aligned}\Pi(\boldsymbol{n}_0, \{\boldsymbol{n}_\ell^j\}_{\ell=1,\ldots,J,\ j=1,\ldots,J}) =& BetaBin(n_{0\bullet}; N, a_0, b_0) \cdot \Pi_{H_{0+}}^{n_{0\bullet}}(\boldsymbol{n}_0; \alpha_0) \\ & \cdot \prod_{j=1}^{J}\left\{DirMult(\boldsymbol{n}_\bullet{}^j; \boldsymbol{\gamma}^j) \cdot \prod_{\ell=1}^{J}\Pi_{H_{\ell+}^j}^{n_{\ell\bullet}^j}(\boldsymbol{n}_\ell^j; \alpha_\ell^j)\right\},\end{aligned}$$

where $\boldsymbol{n}_\bullet{}^j = (n_{1\bullet}^j, \ldots, n_{J\bullet}^j)$, $\boldsymbol{n}_\ell^j = (n_{\ell 1}^j, \ldots, n_{\ell H_{\ell+}^j}^j)$, $\boldsymbol{n}_0 = (n_{01}, \ldots, n_{0H_{0+}})$, $\boldsymbol{\gamma}^j = (\gamma_1^j, \ldots, \gamma_J^j)$. $\Pi_{H_{0+}}^{n_{0\bullet}}(\boldsymbol{n}_0; \alpha_0)$ *and* $\Pi_{H_{\ell+}^j}^{n_{\ell\bullet}^j}(\boldsymbol{n}_\ell^j; \alpha_\ell^j)$ *are the EPPFs of the static MFM defined in equation* (12). *BetaBin*$(\cdot)$, *DirMult*$(\cdot)$ *denote the probability functions of a beta-binomial and Dirichlet-multinomial distribution, respectively.*

Proof: See the appendix.

An alternative representation of the random partition is given by the predictive probability function [26]. Let s_{new}^j, c_{new}^j denote latent cluster membership indicators for a future $i = (N_j + 1)-$st sample from G_j. The predictive distribution for s_{new}^j, c_{new}^j given $\boldsymbol{c} = (c_{ji})_{j=1,\ldots,J,\ i=1,\ldots,N_j}$ and $\boldsymbol{s} = (s_{ji})_{j=1,\ldots,J,\ i=1,\ldots,N_j}$ takes a form similar to the Pólya-urn under a DP prior [27]. The following expression is useful for constructing posterior MCMC transition probabilities.

Proposition 2 *The Pólya-urn representation of the predictive probability function under the 888 model is*

$$p(s_{new}^j, c_{new}^j \mid \boldsymbol{c}, \boldsymbol{s}) = \frac{N+1}{a_0 + b_0 + N} \cdot \begin{cases} \frac{a_0 + n_{0\bullet}}{n_{0\bullet}+1} \begin{cases} \frac{V(n_{0\bullet}+1, 1+H_{0+}, \alpha_0)}{V(n_{0\bullet}, H_{0+}, \alpha_0)} \cdot \alpha_0, & c_{new}^j = 0, s_{new}^j = 1 + H_{0+} \\ \frac{V(n_{0\bullet}+1, H_{0+}, \alpha_0)}{V(n_{0\bullet}, H_{0+}, \alpha_0)} \cdot (n_{0h} + \alpha_0), & s_{new}^j = h \leq H_{0+} \end{cases} \\ \frac{b_0 + N - n_{0\bullet}}{N - n_{0\bullet}+1} \begin{cases} \frac{(\gamma_\ell^j + n_{\ell\bullet}^j)(\sum_\ell n_{\ell\bullet}^j)}{(\sum_\ell \gamma_\ell^j + \sum_\ell n_{\ell\bullet}^j)(n_{\ell\bullet}^j+1)} \frac{V(n_{\ell\bullet}^j+1, 1+H_{\ell+}^j, \alpha_\ell^j)}{V(n_{\ell\bullet}^j, H_{\ell+}^j, \alpha_\ell^j)} \cdot \alpha_\ell^j, & c_{new}^j = \ell, s_{new}^j = 1 + H_{\ell+}^j \\ \frac{(\gamma_\ell^j + n_{\ell\bullet}^j)(\sum_\ell n_{\ell\bullet}^j)}{(\sum_\ell \gamma_\ell^j + \sum_\ell n_{\ell\bullet}^j)(n_{\ell\bullet}^j+1)} \frac{V(n_{\ell\bullet}^j+1, H_{\ell+}^j, \alpha_\ell^j)}{V(n_{\ell\bullet}^j, H_{\ell+}^j, \alpha_\ell^j)} \cdot (n_{\ell h}^j + \alpha_\ell^j), & s_{new}^j = h \leq H_{\ell+}^j \end{cases} \end{cases}$$

where s_{new}^j, c_{new}^j *denote the allocation of a new object in condition* j*, indicating the cluster and the type of cluster, respectively, as defined in equation* (3).

Proof: See the appendix.

To study large sample properties of posterior inference under the 888 model it is convenient to focus on the implied marginal model for condition j. The marginal model for y_{ji}, $i = 1, \ldots, N_j$, can be written as a simple finite mixture model with all $J \times H$ clusters, and an asymmetric Dirichlet prior on the weights by using the aggregation property of the Dirichlet distribution, if assuming the following constraints on the prior model:

Lemma 1 *Assume the hyperparameters in the priors of the weights satisfy*

$$\begin{aligned} H\alpha_j^j &= \gamma^{(H)}, \quad \forall\, j = 1, \ldots, J, \\ H\alpha_\ell^j &= \gamma^{(L)}, \quad \forall\, \ell \neq j,\ l, j = 1, \ldots, J, \\ a_0 = H\alpha_0, &\quad b_0 = \gamma^{(H)} + (J-1)\gamma^{(L)}. \end{aligned} \tag{13}$$

Then the marginal model for y_j under the 888 model is equivalent to a MFM model with asymmetric Dirichlet prior:

$$\begin{aligned} & y_{ji} \sim p(y_{ji}) = \int f(y_{ji} \mid \theta) dG_j(\theta), \\ & G_j = \sum_{m=1}^{(J+1)H} \pi_m^j \delta_{\phi_m}, \quad j = 1, \ldots, J, \\ & \boldsymbol{\pi}^j \sim Dir(\boldsymbol{\alpha}^j) = Dir(\alpha_0, \ldots, \alpha_0, \alpha_1^j, \ldots, \alpha_1^j, \ldots, \alpha_J^j, \ldots, \alpha_J^j), \\ & \phi_m \overset{iid}{\sim} \Phi(\phi_m). \end{aligned} \tag{14}$$

Proof: The claim is implied by the aggregation property of the Dirichlet distribution [28, Chapter 49].

In this case, the 888 model can be considered as a variation of the MFM model with an asymmetric Dirichlet prior that incorporates the desired structure.

We briefly discuss large sample properties of the posterior for the mixing measure G_j. Assume that data y_{ji} are generated from a mixture with true mixing measure G_j^0. We show that for any condition j, the posterior distribution, $\Pi(G_j \mid y_{j1}, \ldots, y_{jN_j})$ converges with increasing N_j to the true mixing measure G_j^0, and the posterior contraction rate is given as follows.

Proposition 3 *Under model* (14)*, for* $\forall j = 1, \ldots, J$*, we have*

$$\Pi\left(G_j \in \bar{\mathcal{G}}(\Theta) : W_1(G_j, G_j^0) \lesssim (logN_j/N_j)^{1/2} \mid y_{j1}, \ldots, y_{jN_j}\right) \to 1$$

in probability under G_j^0. Here, $\bar{\mathcal{G}}(\Theta)$ denotes the space of all discrete measures, and W_1 denotes the L_1 Wasserstein distance.

Proof: We give an outline of the proof. The claim follows from Theorem 3.1 in Guha et al. [29] which gives posterior contraction rates for a general MFM. Replacing the symmetric

Dirichlet prior with the Dirichlet prior in equation (14), the condition on the prior probability of the parameters in a neighborhood of the truth used in the proof still holds. The assumptions (P.1) - (P.4) in [29] hold under our proposed model. Therefore, we can prove the posterior contraction rate following the scheme of proof in Guha et al. [29]. See the Appendix for details.

If we consider the finite case where the number of clusters $H < \infty$ is fixed, the posterior distribution of the mixing measure can also be proven to be consistent, and obtain a posterior contraction rate of $(\log N_j/N_j)^{1/2}$. This follows from Guha et al. [29] and Ho and Nguyen [30].

In summary, for large enough sample size (under each condition j) inference under the proposed model can recover the underlying sub-populations close to the truth. Therefore, under large enough sample sizes in all conditions, inference on the comparison between conditions recovers the truth.

4.2 Posterior inference

In our implementation (in the simulation studies and the data analysis) we use mixtures with fixed H. This choice is motivated by the fact that the number of immune subtypes that we are interested in for biological interpretation is limited. The model is sufficiently flexible as long as H is chosen as an upper bound beyond the number of immune subtypes of interest. We summarize posterior inference by first reporting a point estimate of the partition by minimizing the variation of information (VI) loss as proposed in Dahl et al. [31]. To maintain the structure of the model, the search for the optimal partition is restricted to the posterior Monte Carlo samples of imputed partitions. The *salso* R package is used to calculate VI loss. Minimizing VI loss, we obtain an optimal cluster assignment $\hat{c}_i$, and estimated cluster specific parameters $\hat{\phi}_{\ell h}$, including the cluster centroids $\hat{\xi}_{\ell h}$ and the covariance matrices $\hat{\Sigma}_{\ell h}$.

We report uncertainty on cluster allocation of the cells given the cluster specific parameters. Considering the posterior mean of the log cluster assignment probability, we define a metric $q_{ji}(\ell, h)$ as the similarity of cell i under condition j to cluster h among the type ℓ clusters,

$$q_{ji}(\ell, h) \propto E[\log p(c_{ji} = \ell, s_{ji} = h) \mid \hat{\boldsymbol{\phi}}].$$

For a given cell, the higher the similarity to cluster h, the higher the probability that it is assigned to cluster h. Here $\hat{\boldsymbol{\phi}}$ is a point estimate of the cluster specific parameters.

5 Simulation Study

We carry out a simulation study to evaluate the performance of the 888 model under realistic finite sample sizes and with single-cell sequencing data. The simulation shows how the

model successfully identifies the true clusters and provides the desired inference on comparative abundances of cell type sub-populations between two conditions.

The simulated data is generated from a publically available dataset of Peripheral Blood Mononuclear Cells (PBMC) collected on a 10X platform [32]. The cell types given in the dataset are used as hypothetical simulation truth, and are denoted as types 1-4 in Figure 2 (a). We assume the cells are collected under two conditions A and B. We generate cell-specific conditions with a defined true probability $P(\text{cell } i \in \text{ condition B} \mid \text{cell } i \text{ is of type } k) = p_k$. In our simulation setting, $p_1 = 0.01$, $p_2 = p_3 = 0.99$, $p_4 = 0.52$. Therefore, under the simulation truth, cell type 1 is dominant in condition A, cell types 2 and 3 are enriched under condition B, and cell type 4 shares a similar, but slightly different proportion between two conditions. Figure 2 (b) plots the cells color coded by simulated (true) conditions using UMAP scores. Figure 2 (c) shows the proportion of the four true cell types under condition A and B, respectively. In the analysis we assume $H = 2$, that is, we assume *a priori* that the first two clusters are shared, the second two clusters are specific to conditions B, and the last two clusters are specific to conditions A.

The inference results are consistent with the simulation truth. Figures 3 (a)-(c) show posterior summaries for the weights and the weight differences between the two conditions. Under the assumed model structure, the differences highlight which clusters are common and which are condition-specific. Figures 3 (c) and (d) indicate that cells of type 1 are mainly assigned to clusters specific to condition A, while cells of type 2 and 3 are mainly assigned to clusters specific to B, and type 4 cells are mostly assigned to the common clusters. Compared to the simulation truth that type 1 cells are enriched under A, type 2 and 3 cells are enriched under B, and type 4 cells are common, the accuracy of classifying cells to common, condition A enriched, and condition B enriched clusters is 91.2%. The balanced accuracy for the three types of clusters is summarized in Table 1.

Cluster assignments are shown in the panels in the first row of Figure 4. The figure plots UMAP scores colored by cluster assignments. The subfigures include a plot for all cells and for condition A and condition B only, respectively. Panels in rows 2 through 4 of Figure 4 show condition A cells and condition B cells in three types of clusters, common, specific to A, and specific to B. The clusters have the intended structures as desired. That is, the clusters specific to A have significantly higher proportion of cells from condition A, and very few cells from condition B, and vice versa for clusters specific to B.

	Common	Condition A enriched	Condition B enriched
Balanced accuracy	0.95	0.98	0.92

Table 1: Balanced accuracy of classifying cells as Common/Condition A enriched/Condition B enriched.

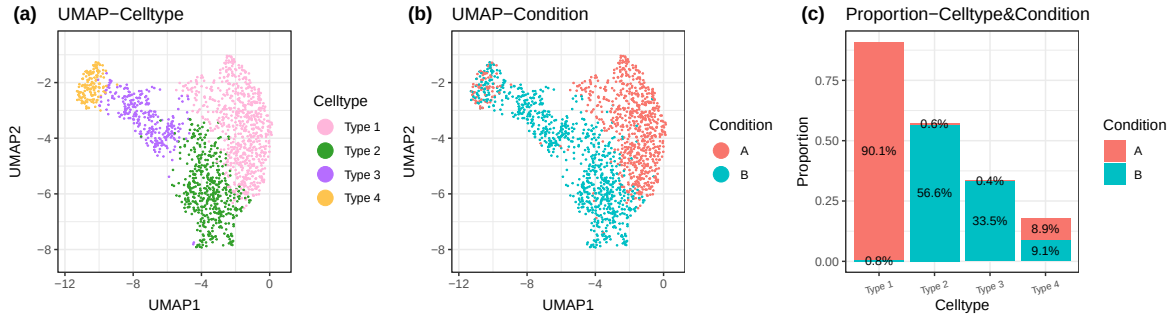

Figure 2: Simulation truth. UMAP plot of the single cell data colored by true cell type (panel a), and condition (b). Panel (c) shows the proportion of true cell types within the two conditions, respectively.

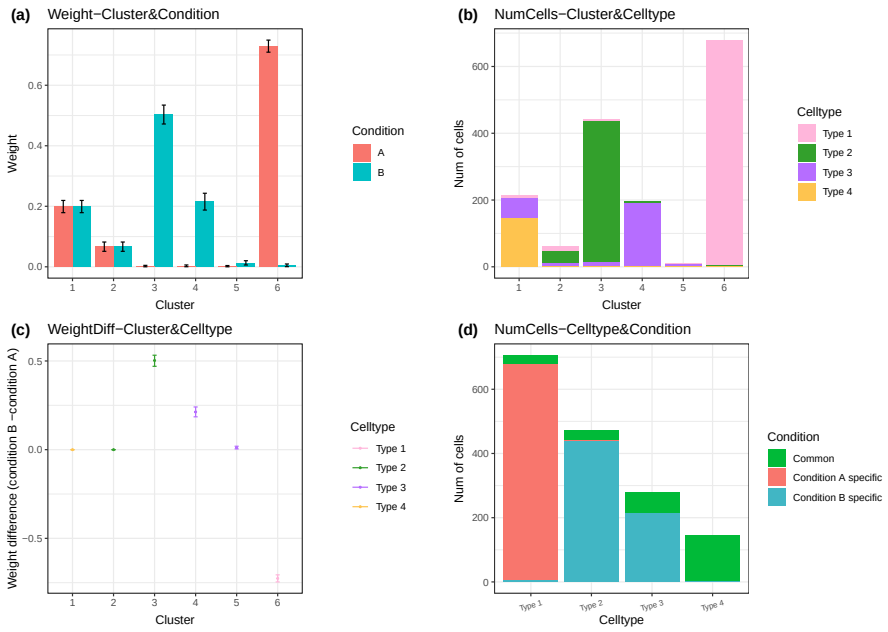

Figure 3: Inference results in the simulation study. (a) Posterior means and 95% credible intervals of the weights under the two conditions. (b) The number of cells from each true cell type in the reported clusters. (c) Posterior summaries of the differences for each cluster of the weights under condition A versus B. The plot shows the posterior means and credible intervals. (d) The number of cells in each true cell type being assigned into the three types of clusters.

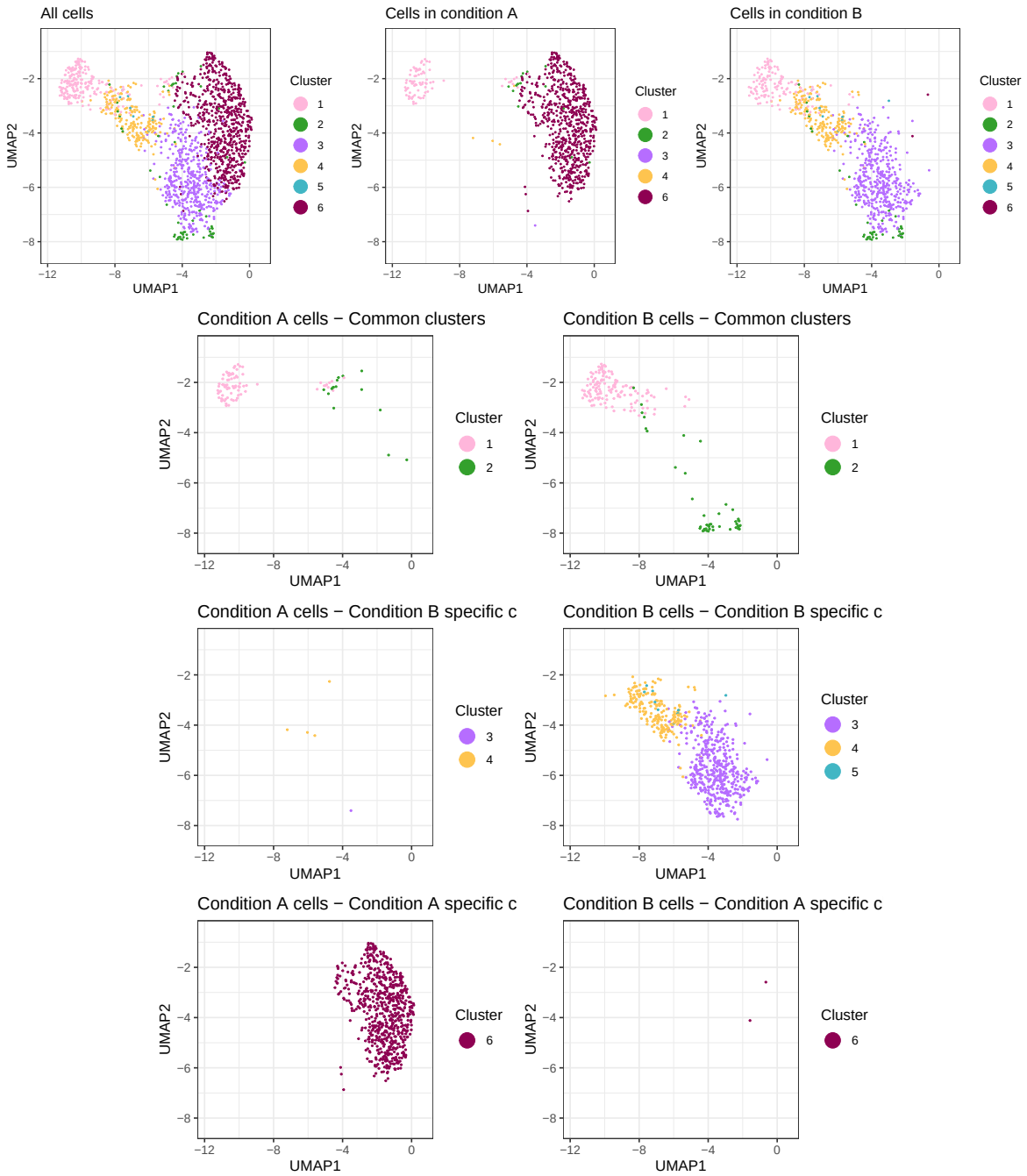

Figure 4: Simulation results (plotted as UMAP scores). The first row shows the clusters for all cells, cells under condition A, and cells under condition B, respectively. The second row shows condition A/B cells in the common clusters. The third row shows condition A/B cells in clusters that are enriched under condition B. The fourth row shows condition A/B cells in clusters enriched under condition A.

6 Results

We implemented inference under the 888 model on CRC data from Zhang et al. [22]. The features y_{ji} are the PCA scores obtained from the pre-processing pipeline outlined in Section 2. To prevent the PCA process from excluding representative immune cell biomarkers, we supplemented the features in y_{ji} with differentially expressed genes based on cell type annotation. Consequently, y_{ji} comprises the first n_1 PCA scores and the top n_2 differentially expressed (DE) genes based on false discovery rates. In our specific case, we choose the value of n_1 based on an elbow plot of the proportion of variance explained versus the number of PCA components, suggesting $n_1 = 25$, and we choose $n_2 = 20$. We carry out separate inference for CD4+ and CD8+ cells due to the comprehensive research on the biomarkers specific to these cell types and the clear separation achieved through CD4+ and CD8+ annotation. In both, the CD4+ and CD8+ models, we set H to 2 as an upper bound on the resolution level of features that we can realistically expect in this data.

Posterior inference on the partitions reports the following homogeneous cell sub-populations. A point estimate of the partition, i.e., a clustering assignment, is obtained by minimizing variance-of-information (VI) loss [31]. The allocation of cells to these clusters under different conditions can be seen in Figure 6. Figures 5 (a) and (b) display the posterior means and credible intervals of the condition-specific weights and weight differences, respectively. Using the 888 model, we identified six clusters within the CD4+ T cell condition and five clusters within the CD8+ T cell condition. Specifically, within the CD4+ T cells, clusters 1, 2, and 3 were classified as common clusters, while clusters 4 and 5 exhibited enrichment in LOCRC. The $(H + 1)$-st common cluster here is an EO-specific cluster under the prior model for which we found *a posteriori* negligible difference of the condition-specific weights. Cluster 6, on the other hand, presented EO enrichment. Within CD8+ T cells, clusters 7 and 8 were classified as common clusters, cluster 9 showed LO enrichment, and Clusters 10 and 11 displayed EO enrichment. Another LO enriched CD8+ cluster has negligible weights under both conditions, and we combine it with the closest cluster.

Next we studied the relation of the discovered clusters with known T cell subtypes, hoping to identify new variations or refined subtypes. Figure 5 (c) summarizes the composition of different immune cell types within the clusters. By comparing the 888 clustering results with the cell type annotation based on known markers, we made the following observations: (1) a specific subpopulation of regulatory T (Treg) cells demonstrated enrichment in the EOCRC samples; (2) a subpopulation of CD8+ effective memory T (Tem) cells displayed enrichment in the LOCRC samples; and (3) we identified a subpopulation of CD8+ exhausted T (Tex) cells that exhibited enrichment specifically in the LOCRC samples. Lastly, to illustrate the uncertainty in cluster assignment, Figure 5 (d) depicts the similarity of cells to a cluster, using

cluster 5 as an example. This summary highlights the variability and ambiguity in cluster assignments.

Given the identified common, LO-enriched, and EO-enriched T cell subtypes, we look further into their gene expression profiles, hoping to annotate these clusters with known T cell states in cancer [33] (Figure 7). We identify some immune subtypes with differentially expressed genes matching the known markers. In the LO-enriched CD4+ T helper cells, we observed a high expression of CCR7 and TCF7, which are markers for naïve T cells, suggesting a self-renewal characteristic [34, 35]. Conversely, the EO-enriched CD4+ T helper cells exhibited high expression of CCL5, which was linked to various mechanisms of tumor progression, including extracellular matrix remodeling, enhanced tumor cell migration, and the induction of immunosuppressive polarization in macrophages [36]. In EO-enriched CD4+ Treg cells, we found higher expression of co-stimulatory molecules TNFRSF4 and TNFRSF18. These molecules are known to enhance the proliferation and activation of Treg cells [37]. Additionally, the EO-enriched Treg cells displayed higher expression of IL2RB, indicating potentially stronger immunosuppressive activity [38]. In the CD8+ T cell population, including Tem and Tex cells, the EO-enriched Tem cells exhibited higher expression of CXCL13. Previous studies have linked the presence of intratumoral CXCL13+ CD8+ T cells to unfavorable clinical outcomes and an immunoevasive environment. [39]. Additionally, the EO-enriched Tex cells showed higher expression of TNFRSF4, which contributes to immunosuppression and pro-angiogenesis in the tumor microenvironment [37].

To summarize, we conducted comparative inference between two CRC conditions under the 888 model. Our findings revealed significant differences in immune cell subpopulations between EOCRC and LOCRC patients. The EOCRC samples, compared to LOCRC samples, exhibited an enrichment of immune-suppressive T cells, indicating a more immune-suppressive tumor microenvironment (TME) in EOCRC patients. These findings have the potential to inform novel approaches for the treatment and management of CRC, considering the distinct characteristics of the immune cell subpopulations in each condition.

7 Discussion

We introduced dependent mixture models with multi-layer structured priors on the weights to identify homogeneous subpopulations that highlight differential abundance across conditions. The proposed model allows us to evaluate comparative characteristics of T cell subtypes for EO versus LO colorectal cancer. We discussed limitations of existing models for dependent probability distributions in the context of the desired inference goals and showed how the asymmetric Dirichlet priors in the proposed hierarchically structured mixture models allow for the required comparison of immune profiles. Finite sample inference under the proposed approach

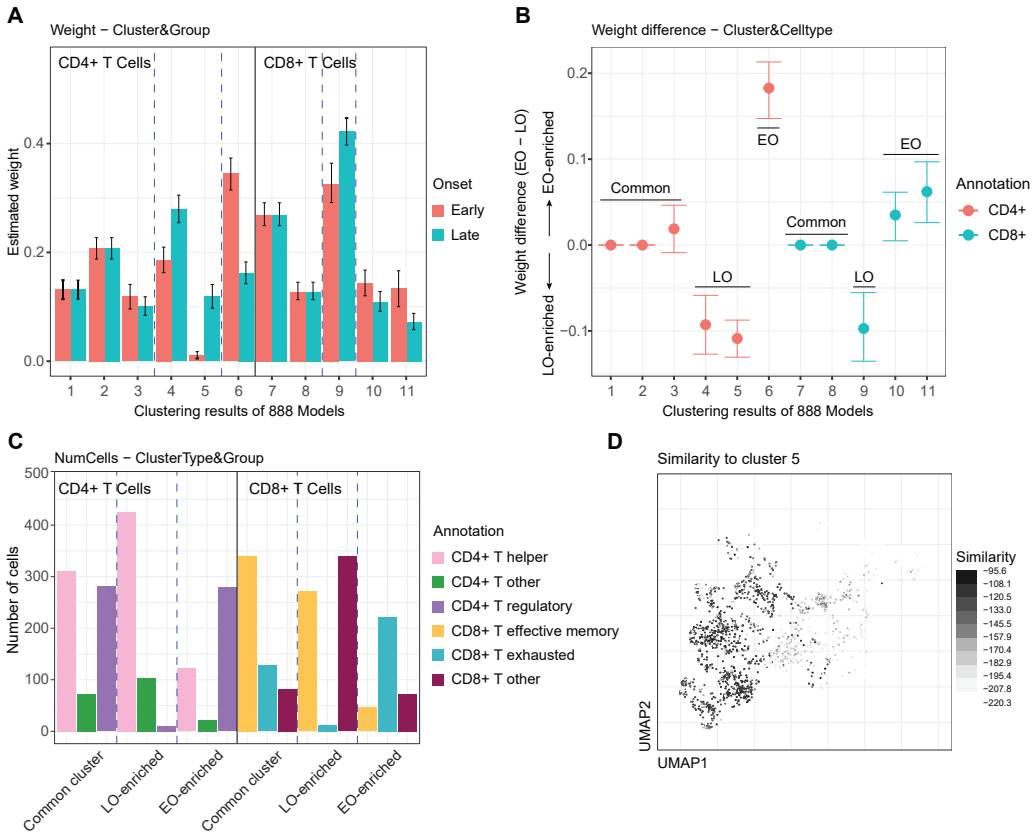

Figure 5: Results of the immune profile comparison between early-onset versus late-onset CRC. (a) Posterior mean and 95% credible interval of the cluster weights under EO/LO condition. (b) Posterior mean and 95% credible interval of the cluster weight differences colored by the dominant annotated cell type. (c) Number of cells from the annotated cell types in Common, LO-enriched, and EO-enriched clusters. (d) The similarity of each cell to cluster 5.

was explored in simulation studies. Inference results with the immune profile comparison data, as intended, highlight immune subtypes that are enriched in EOCRC/LOCRC, and subtypes that are shared across the two conditions.

We only recommend full model-based inference using the proposed model when a quantitative comparison of subpopulations across conditions is desired. The proposed model is not needed and unnecessarily complex when the aim is only to report a cluster arrangement of experimental units – in that case any ad-hoc clustering algorithm would do.

Several extensions of the proposed model could be considered. In this implementation we used a fixed number of components H in the mixture. But the model could easily be extended to random H, with an informative prior. This generalization would give rise to additional computational challenges when implementing trans-dimensional posterior inference [19]. Another

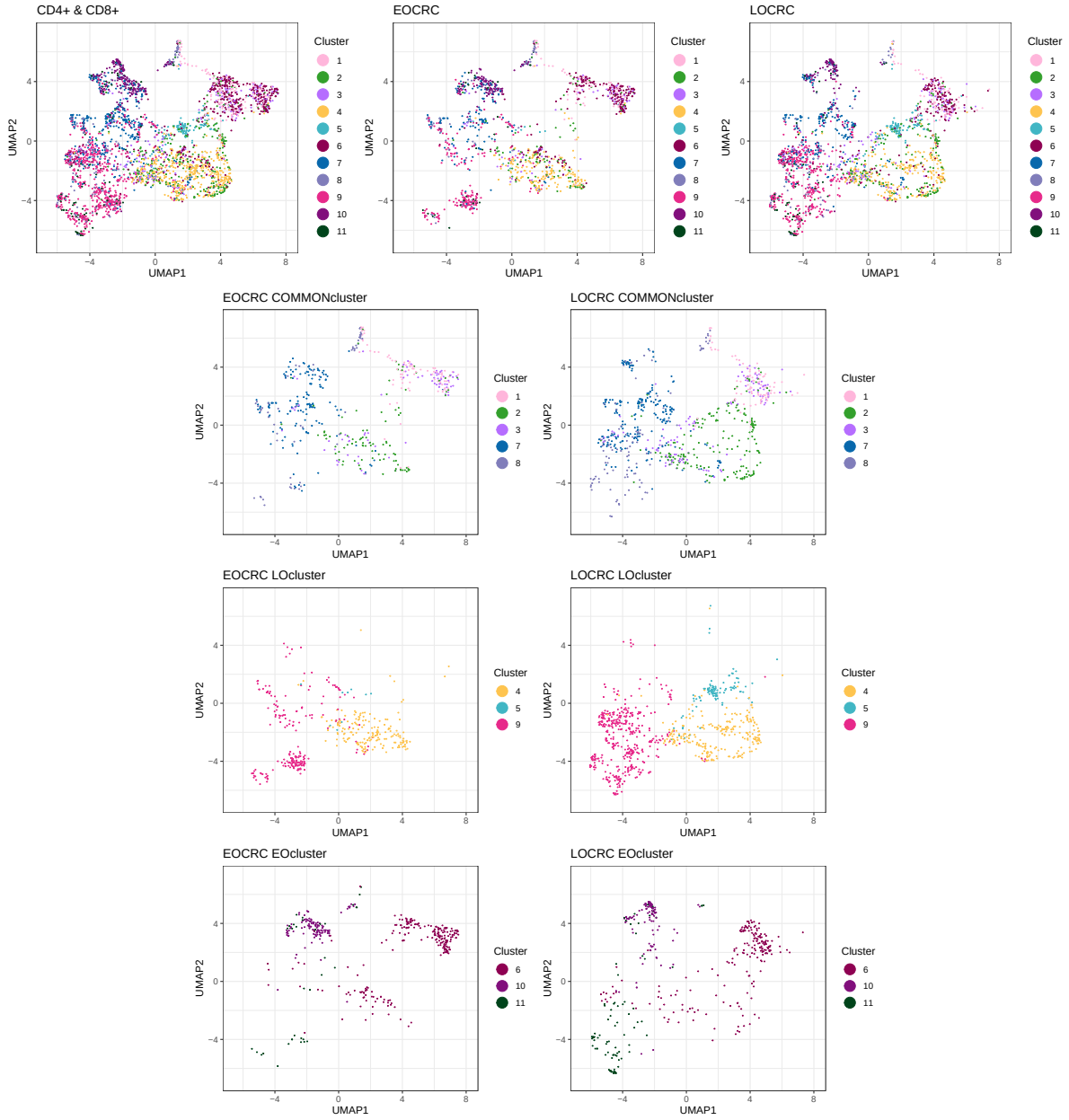

Figure 6: Results plotted in UMAP scores. The first row shows the clusters for all cells, EO cells, and LO cells. The second row shows EO/LO cells in the common clusters. The third row shows EO/LO cells in LO-enriched clusters. The fourth row shows EO/LO cells in EO-enriched clusters.

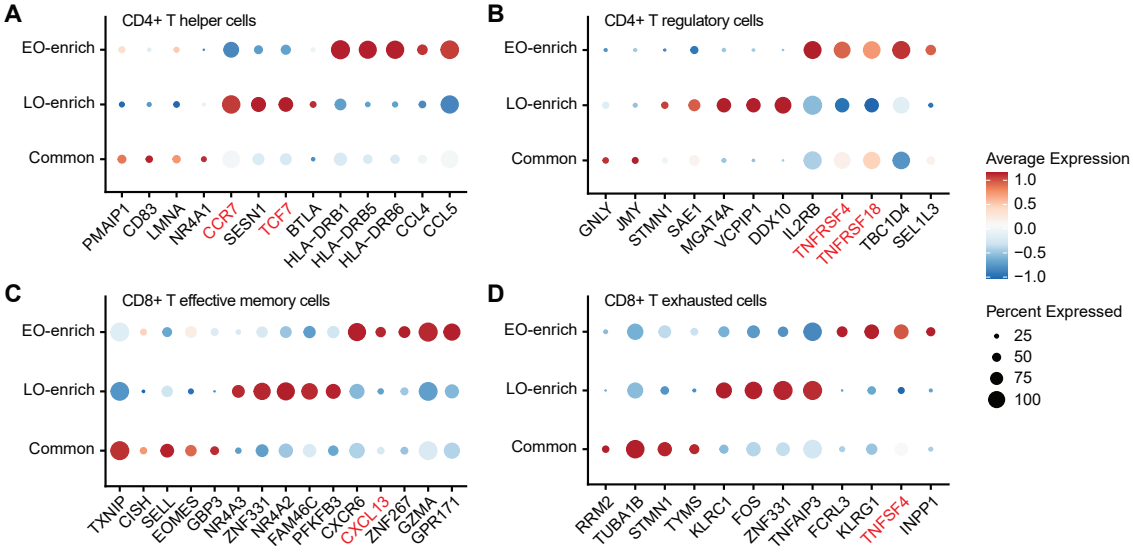

Figure 7: Differentially expressed genes in common, EO-enriched, LO-enriched clusters for CD4+ T helper, CD4+ T regulatory, CD8+ effective memory, and CD8+ T exhausted cells, respectively.

important generalization would be the inclusion of patient-specific random effects and possibly patient-specific covariates. One could introduce additive random effects for the normal location parameters in (7), or alternatively in the Dirichlet prior for the weights. In the latter case one would represent the Dirichlet random weights as normalized gamma random variables and use multiplicative random effects on those. The use of additive random effects formalizes the notion that the same homogeneous cell subpopulation might express slightly differently across patients or samples. The use of multiplicative random effects on the un-normalized weights corresponds to varying proportions of the same homogeneous subpopulations across patients. Incorporating covariates would naturally be formalized by multiplicative effects on the un-normalized weights.

The proposed model for comparative immune profiling might be used as a submodel for immune cells in a larger study of single cell data. Incorporating this submodel in the overall study pipeline, one can address the analysis on the immune subtypes that are distinctively enriched in some conditions, which helps explain the difference of micro-environment across disease subtypes.

References

[1] Manon CW Spaander, Ann G Zauber, Sapna Syngal, Martin J Blaser, Joseph J Sung, Y Nancy You, and Ernst J Kuipers. Young-onset colorectal cancer. *Nature Reviews Disease Primers*, 9(1):21, 2023.

[2] Olivia Marx, Marc Mankarious, and Gregory Yochum. Molecular genetics of early-onset colorectal cancer. *World Journal of Biological Chemistry*, 14(2):13, 2023.

[3] John Frank Charles Kingman. The representation of partition structures. *J. London Math. Soc. (2)*, 18(2):374–380, 1978. ISSN 0024-6107. doi: 10.1112/jlms/s2-18.2.374.

[4] Subhashis Ghosal and Aad Van der Vaart. *Fundamentals of nonparametric Bayesian inference*, volume 44. Cambridge University Press, 2017.

[5] Peter Müller, Fernando Quintana, and Gary Rosner. A method for combining inference across related nonparametric Bayesian models. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 66(3):735–749, 2004.

[6] Michalis Kolossiatis, Jim E Griffin, and Mark FJ Steel. On Bayesian nonparametric modelling of two correlated distributions. *Statistics and Computing*, 23:1–15, 2013.

[7] Antonio Lijoi, Bernardo Nipoti, and Igor Prünster. Bayesian inference with dependent normalized completely random measures. *Bernoulli*, 20:1260–1291, 8 2014.

[8] Yee Whye Teh, Michael I. Jordan, Matthew J. Beal, and David M. Blei. Hierarchical Dirichlet processes. *Journal of the American Statistical Association*, 101(476):1566–1581, 2006.

[9] Abel Rodriguez, David B Dunson, and Alan E Gelfand. The nested Dirichlet process. *Journal of the American Statistical Association*, 103(483):1131–1154, 2008.

[10] Federico Camerlenghi, Antonio Lijoi, Peter Orbanz, and Igor Prünster. Distribution theory for hierarchical processes. *The Annals of Statistics*, 47(1):67–92, 2019.

[11] Raffaele Argiento, Andrea Cremaschi, and Marina Vannucci. Hierarchical normalized completely random measures to cluster grouped data. *Journal of the American Statistical Association*, 115:318–333, 1 2020.

[12] Federico Camerlenghi, Antonio Lijoi, and Igor Prünster. Bayesian nonparametric inference beyond the Gibbs-type framework. *Scandinavian Journal of Statistics*, 45:1062–1091, 12 2018.

[13] John Frank Charles Kingman. *Poisson processes*, volume 3 of *Oxford Studies in Probability*. The Clarendon Press Oxford University Press, New York, 1993. ISBN 0-19-853693-3. Oxford Science Publications.

[14] Federico Camerlenghi, David B Dunson, Antonio Lijoi, Igor Prünster, and Abel Rodríguez. Latent nested nonparametric priors (with discussion). *Bayesian Analysis*, 14(4):1303, 2019.

[15] Francesco Denti, Federico Camerlenghi, Michele Guindani, and Antonietta Mira. A common atoms model for the Bayesian nonparametric analysis of nested data. *Journal of the American Statistical Association*, 2021:1–12, 7 2021.

[16] Antonio Lijoi, Igor Prünster, and Giovanni Rebaudo. Flexible clustering via hidden hierarchical Dirichlet priors. *Scandinavian Journal of Statistics*, 50(1):213–234, 2023.

[17] Juhee Lee, Peter Müller, Yitan Zhu, and Yuan Ji. A nonparametric Bayesian model for local clustering with application to proteomics. *Journal of the American Statistical Association*, 108(503):775–788, 2013.

[18] Qiaohui Lin, Giovanni Rebaudo, and Peter Mueller. Separate exchangeability as modeling principle in Bayesian nonparametrics. *arXiv preprint arXiv:2112.07755*, 2021.

[19] Sylvia Frühwirth-Schnatter, Gertraud Malsiner-Walli, and Bettina Grün. Generalized mixtures of finite mixtures and telescoping sampling. *Bayesian Analysis*, 16(4):1279–1307, 2021.

[20] Jan Greve, Bettina Grün, Gertraud Malsiner-Walli, and Sylvia Frühwirth-Schnatter. Spying on the prior of the number of data clusters and the partition distribution in Bayesian cluster analysis. *Australian & New Zealand Journal of Statistics*, 64(2):205–229, 2022.

[21] Jeffrey W Miller and Matthew T Harrison. Mixture models with a prior on the number of components. *Journal of the American Statistical Association*, 113(521):340–356, 2018.

[22] Lei Zhang, Xin Yu, Liangtao Zheng, Yuanyuan Zhang, Yansen Li, Qiao Fang, Ranran Gao, Boxi Kang, Qiming Zhang, Julie Y Huang, et al. Lineage tracking reveals dynamic relationships of T cells in colorectal cancer. *Nature*, 564(7735):268–272, 2018.

[23] Lei Zhang, Ziyi Li, Katarzyna M Skrzypczynska, Qiao Fang, Wei Zhang, Sarah A O'Brien, Yao He, Lynn Wang, Qiming Zhang, Aeryon Kim, et al. Single-cell analyses inform mechanisms of myeloid-targeted therapies in colon cancer. *Cell*, 181(2):442–459, 2020.

[24] Thierry André, Kai-Keen Shiu, Tae Won Kim, Benny Vittrup Jensen, Lars Henrik Jensen, Cornelis Punt, Denis Smith, Rocio Garcia-Carbonero, Manuel Benavides, Peter Gibbs,

et al. Pembrolizumab in microsatellite-instability–high advanced colorectal cancer. *New England Journal of Medicine*, 383(23):2207–2218, 2020.

[25] James B MacQueen. Some methods for classification and analysis of multivariate observations. In *Proc. 5th Berkeley Symposium on Math., Stat., and Prob*, page 281, 1965.

[26] Jaeyong Lee, Fernando A Quintana, Peter Müller, and Lorenzo Trippa. Defining predictive probability functions for species sampling models. *Statistical science*, 28(2):209, 2013.

[27] David Blackwell and James B MacQueen. Ferguson distributions via Pólya urn schemes. *The Annals of Statistics*, 1(2):353–355, 1973.

[28] Samuel Kotz, Narayanaswamy Balakrishnan, and Norman L Johnson. *Continuous multivariate distributions, Volume 1: Models and applications*, volume 1. John Wiley & Sons, 2004.

[29] Aritra Guha, Nhat Ho, and XuanLong Nguyen. On posterior contraction of parameters and interpretability in Bayesian mixture modeling. *Bernoulli*, 27(4):2159–2188, 2021.

[30] Nhat Ho and XuanLong Nguyen. On strong identifiability and convergence rates of parameter estimation in finite mixtures. *Electronic Journal of Statistics*, 10(1):271 – 307, 2016.

[31] David B Dahl, Devin J Johnson, and Peter Müller. Search algorithms and loss functions for Bayesian clustering. *Journal of Computational and Graphical Statistics*, 31(4):1189–1201, 2022.

[32] Satija Lab. *pbmc3k.SeuratData: 3k PBMCs from 10X Genomics*, 2020. R package version 3.1.4.

[33] Massimo Andreatta, Jesus Corria-Osorio, Sören Müller, Rafael Cubas, George Coukos, and Santiago J Carmona. Interpretation of T cell states from single-cell transcriptomics data using reference atlases. *Nature Communications*, 12(1):2965, 2021.

[34] Shannon K Bromley, Seddon Y Thomas, and Andrew D Luster. Chemokine receptor CCR7 guides T cell exit from peripheral tissues and entry into afferent lymphatics. *Nature Immunology*, 6(9):895–901, 2005.

[35] Giulia Escobar, Davide Mangani, and Ana C Anderson. T cell factor 1: A master regulator of the T cell response in disease. *Science Immunology*, 5(53):eabb9726, 2020.

[36] Donatella Aldinucci, Cinzia Borghese, and Naike Casagrande. The CCL5/CCR5 axis in cancer progression. *Cancers*, 12(7):1765, 2020.

[37] Michael Croft, Takanori So, Wei Duan, and Pejman Soroosh. The significance of OX40 and OX40L to T-cell biology and immune disease. *Immunological Reviews*, 229(1):173–191, 2009.

[38] Takatoshi Chinen, Arun K Kannan, Andrew G Levine, Xiying Fan, Ulf Klein, Ye Zheng, Georg Gasteiger, Yongqiang Feng, Jason D Fontenot, and Alexander Y Rudensky. An essential role for the IL-2 receptor in Treg cell function. *Nature Immunology*, 17(11): 1322–1333, 2016.

[39] Kaifeng Jin, Yifan Cao, Yun Gu, Hanji Fang, Yuchao Fei, Jieti Wang, Xin Liu, Kunpeng Lv, Xudong He, Chao Lin, et al. Poor clinical outcomes and immunoevasive contexture in CXCL13+ CD8+ T cells enriched gastric cancer patients. *Oncoimmunology*, 10(1): 1915560, 2021.

Appendix A

Clinical characteristics of patients with CRC

Characteristics	P0701	P1012	P1207	P1212	P1228	P0215	P0411
Age	68	35	66	42	77	75	75
Gender	Female	Female	Female	Female	Female	Male	Male
Histological type[a]	Rectum ADC	Colon ADC	Colon ADC	Colon ADC	Colon ADC	Colon ADC	Rectum ADC
pTNM: T	1	4	4	4	4	4	4
pTNM: N	0	2	0	0	0	2	0
pTNM: M	0	0	0	0	0	1	0
Stage	Ⅰ	ⅢC	Ⅱ	Ⅱ	Ⅱ	Ⅳ	ⅡB
Tumour size	1×0.8 cm	7×6 cm	6×6 cm	6×4 cm	4.5×4 cm	6.5×4 cm	6.5×3.5 cm
Grade	Well-differentiated	Low-differentiated	Moderate-differentiated	Low or moderate differentiated	Low or moderate differentiated	Low-differentiated	Moderate-differentiated
MSI status[b]	MSS	MSS	MSS	MSS	MSS	MSS	MSS

[a] ADC, adenocarcinoma.

[b] MSS, microsatellite stability; MSI, microsatellite instability.

Table 2: Clinical characteristics of 7 patients with CRC involved in this example.

Appendix B

Proof of Proposition 1

Proof: We have introduced the cluster assignment indicators s_{ji}, c_{ji} for unit i under condition j, where c_{ji} characterizes the type of cluster being common or certain condition-specific, s_{ji} is the cluster within a certain type. The probability function for the cluster assignment is given in equation (3). Assume we have a total number of N subjects under conditions $j = 1, \ldots, J$. Let $\mathcal{S}$ and $\mathcal{C}$ denote the partition of $[N] = \{1, \ldots, N\}$ induced by s_{ji} and c_{ji}, respectively. Suppose $\boldsymbol{w} = (w_\ell^j)_{\ell, j=1,\ldots,J}$ is a vector containing all the w_ℓ^j weight parameters. Marginalizing out ϵ and $\boldsymbol{w}$, we have

$$
\begin{aligned}
p(\mathcal{S}, \mathcal{C}) &= \int\int p(\mathcal{S} \mid \mathcal{C}) p(\mathcal{C} \mid \epsilon, \boldsymbol{w}) p(\epsilon) p(\boldsymbol{w}) d\epsilon d\boldsymbol{w}, \\
&= p(\mathcal{S} \mid \mathcal{C}) \int\int p(\mathcal{C} \mid \epsilon, \boldsymbol{w}) p(\epsilon) p(\boldsymbol{w}) d\epsilon d\boldsymbol{w}, \\
&= p(\mathcal{S} \mid \mathcal{C}) \int\int \mathrm{Bin}(n_{0\bullet}; N, \epsilon) \mathrm{Beta}(\epsilon; a_0, b_0) \prod_{j=1}^{J} \left\{ \mathrm{Mult}(\boldsymbol{n}_\bullet^j; \boldsymbol{w}^j) \mathrm{Dir}(\boldsymbol{w}^j; \boldsymbol{\gamma}^j) \right\} d\epsilon d\boldsymbol{w}, \\
&= p(\mathcal{S} \mid \mathcal{C}) \cdot \mathrm{BetaBin}(n_{0\bullet}; N, a_0, b_0) \prod_{j=1}^{J} \left\{ \mathrm{DirMult}(\boldsymbol{n}_\bullet^j; \boldsymbol{\gamma}^j) \right\}.
\end{aligned}
$$

The second equation is true because $p(\mathcal{S} \mid \mathcal{C})$ is independent of ϵ and $\boldsymbol{w}$. The distribution of $p(\mathcal{S} \mid \mathcal{C})$ is a pEPPF induced by the EPPF of the MFMs $F_\ell^{j\star}$ and F_0. Each of them is given by

equation (12), and the elements are conditionally independent given cluster type assignment $c_{ji} = 0$ or ℓ. Therefore, the pEPPF given the cluster type indicators $\{c_{ji}\}$ is

$$p(\mathcal{S} \mid \mathcal{C}) = \Pi_{H_{0+}}^{n_{0\bullet}}(\boldsymbol{n}_0; \alpha_0) \cdot \prod_{j=1}^{J} \prod_{\ell=1}^{J} \left\{ \Pi_{H_{\ell+}^j}^{n_{\ell\bullet}^j}(\boldsymbol{n}_\ell^j; \alpha_\ell^j) \right\}.$$

□

Proof of Proposition 2

Proof: The Pólya-urn representation of the predictive probability function is derived from the random partition probability function for the model. The proof follows the notations in Section 4.

When a new unit under condition j is allocated to a new cluster in the set of common type clusters, such that $c_{new}^j = 0, s_{new}^j = 1 + H_{0+}$, the predictive probability function is given by

$$p(s_{new}^j, c_{new}^j \mid \boldsymbol{c}, \boldsymbol{s}) = \frac{\text{BetaBin}(n_{0\bullet} + 1; N + 1, a_0, b_0)\Pi_{H_{0+}+1}^{n_{0\bullet}+1}(\tilde{\boldsymbol{n}}_0; \alpha_0) \cdot C}{\text{BetaBin}(n_{0\bullet}; N, a_0, b_0)\Pi_{H_{0+}}^{n_{0\bullet}}(\boldsymbol{n}_0; \alpha_0) \cdot C},$$

where $\tilde{\boldsymbol{n}}_0 = (n_{01}, \ldots, n_{0H_{0+}+1}) = (n_{01}, \ldots, n_{0H_{0+}}, 1)$, and

$$C = \prod_{j=1}^{J} \left\{ \text{DirMult}(\boldsymbol{n}_\bullet{}^j; \boldsymbol{\gamma}^j) \cdot \prod_{\ell=1}^{J} \Pi_{H_{\ell+}^j}^{n_{\ell\bullet}^j}(\boldsymbol{n}_\ell^j; \alpha_\ell^j) \right\}.$$

Substituting the beta-binomial probability function, the ratio of the $BetaBin()$ factors is

$$\frac{\binom{N+1}{n_{0\bullet}+1} \frac{\Gamma(n_{0\bullet} + 1 + a_0)\Gamma(N - n_{0\bullet} + b_0)}{\Gamma(N + 1 + a_0 + b_0)B(a_0, b_0)}}{\binom{N}{n_{0\bullet}} \frac{\Gamma(n_{0\bullet} + a_0)\Gamma(N - n_{0\bullet} + b_0)}{\Gamma(N + a_0 + b_0)B(a_0, b_0)}} = \frac{(N+1)(n_{0\bullet} + a_0)}{(n_{0\bullet} + 1)(N + a_0 + b_0)}$$

Plugging in the definition of $\Pi()$, the ratio of the $\Pi()$ term is

$$\frac{V(n_{0\bullet} + 1, 1 + H_{0+}, \alpha_0) \cdot \prod_{j=1}^{H_{0+}} \frac{\Gamma(n_{0j} + \alpha_0)}{\Gamma(\alpha_0)} \cdot \frac{\Gamma(1 + \alpha_0)}{\Gamma(\alpha_0)}}{V(n_{0\bullet}, H_{0+}, \alpha_0) \cdot \prod_{j=1}^{H_{0+}} \frac{\Gamma(n_{0j} + \alpha_0)}{\Gamma(\alpha_0)}} = \frac{V(n_{0\bullet} + 1, 1 + H_{0+}, \alpha_0)}{V(n_{0\bullet}, H_{0+}, \alpha_0)} \cdot \alpha_0$$

Therefore, when $c_{new}^j = 0, s_{new}^j = 1 + H_{0+}$,

$$p(s_{new}^j, c_{new}^j \mid \boldsymbol{c}, \boldsymbol{s}) = \frac{(a_0 + n_{0\bullet})(N+1)}{(a_0 + b_0 + N)(n_{0\bullet} + 1)} \frac{V(n_{0\bullet} + 1, 1 + H_{0+}, \alpha_0)}{V(n_{0\bullet}, H_{0+}, \alpha_0)} \cdot \alpha_0$$

Similarly, when $c^j_{new} = 0, s^j_{new} = h \leq H_{0+}$, let $\tilde{\boldsymbol{n}}_0 = (n_{01}, \ldots, n_{0h} + 1, \ldots, n_{0H_{0+}})$. The predictive probability function is

$$
\begin{aligned}
p(s^j_{new}, c^j_{new} \mid \boldsymbol{c}, \boldsymbol{s}) &= \frac{\mathrm{BetaBin}(n_{0\bullet} + 1; N + 1, a_0, b_0)\Pi^{n_{0\bullet}+1}_{H_{0+}}(\tilde{\boldsymbol{n}}^h_0; \alpha_0) \cdot C}{\mathrm{BetaBin}(n_{0\bullet}; N, a_0, b_0)\Pi^{n_{0\bullet}}_{H_{0+}}(\boldsymbol{n}_0; \alpha_0) \cdot C} \\
&= \frac{(a_0 + n_{0\bullet})(N+1)}{(a_0 + b_0 + N)(n_{0\bullet} + 1)} \frac{V(n_{0\bullet} + 1, H_{0+}, \alpha_0)}{V(n_{0\bullet}, H_{0+}, \alpha_0)} \cdot (n_{0h} + \alpha_0)
\end{aligned}
$$

When $c^j_{new} = \ell, s^j_{new} = 1 + H^j_{\ell+}$, let $\tilde{\boldsymbol{n}}^j_\bullet = (n^j_{1\bullet}, \ldots, n^j_{\ell\bullet} + 1, \ldots, n^j_{J\bullet})$ and $\tilde{\boldsymbol{n}}^j_\ell = (n^j_{\ell 1}, \ldots, n^j_{\ell H^j_{\ell+}}, 1)$.

$$
\begin{aligned}
p(s^j_{new}, c^j_{new} \mid \boldsymbol{c}, \boldsymbol{s}) &= \frac{\mathrm{BetaBin}(n_{0\bullet}; N + 1, a_0, b_0) \cdot \mathrm{DirMult}(\tilde{\boldsymbol{n}}^j_\bullet; \boldsymbol{\gamma}^j) \cdot \Pi^{n^j_{\ell\bullet}+1}_{1+H^j_{\ell+}}(\tilde{\boldsymbol{n}}^j_\ell; \alpha^j_\ell)}{\mathrm{BetaBin}(n_{0\bullet}; N, a_0, b_0) \cdot \mathrm{DirMult}(\boldsymbol{n}_\bullet{}^j; \boldsymbol{\gamma}^j) \cdot \Pi^{n^j_{\ell\bullet}}_{H^j_{\ell+}}(\boldsymbol{n}^j_\ell; \alpha^j_\ell)} \\
&= \frac{(b_0 + N - n_{0\bullet})(N+1)}{(a_0 + b_0 + N)(N - n_{0\bullet} + 1)} \frac{(\gamma^j_\ell + n^j_{\ell\bullet})(\sum_\ell n^j_{\ell\bullet})}{(\sum_\ell \gamma^j_\ell + \sum_\ell n^j_{\ell\bullet})(n^j_{\ell\bullet} + 1)} \frac{V(n^j_{\ell\bullet} + 1, 1 + H^j_{\ell+}, \alpha^j_\ell)}{V(n^j_{\ell\bullet}, H^j_{\ell+}, \alpha^j_\ell)} \cdot \alpha^j_\ell.
\end{aligned}
$$

When $c^j_{new} = \ell, s^j_{new} = h \leq H^j_{\ell+}$, let $\tilde{\boldsymbol{n}}^j_\bullet = (n^j_{1\bullet}, \ldots, n^j_{\ell\bullet} + 1, \ldots, n^j_{J\bullet})$ and $\tilde{\boldsymbol{n}}^j_\ell = (n^j_{\ell 1}, \ldots, n^j_{\ell h} + 1, \ldots, n^j_{\ell H^j_{\ell+}})$.

$$
\begin{aligned}
p(s^j_{new}, c^j_{new} \mid \boldsymbol{c}, \boldsymbol{s}) &= \frac{\mathrm{BetaBin}(n_{0\bullet}; N + 1, a_0, b_0) \cdot \mathrm{DirMult}(\tilde{\boldsymbol{n}}^j_\bullet; \boldsymbol{\gamma}^j) \cdot \Pi^{n^j_{\ell\bullet}+1}_{1+H^j_{\ell+}}(\tilde{\boldsymbol{n}}^j_\ell; \alpha^j_\ell)}{\mathrm{BetaBin}(n_{0\bullet}; N, a_0, b_0) \cdot \mathrm{DirMult}(\boldsymbol{n}_\bullet{}^j; \boldsymbol{\gamma}^j) \cdot \Pi^{n^j_{\ell\bullet}}_{H^j_{\ell+}}(\boldsymbol{n}^j_\ell; \alpha^j_\ell)} \\
&= \frac{(b_0 + N - n_{0\bullet})(N+1)}{(a_0 + b_0 + N)(N - n_{0\bullet} + 1)} \frac{(\gamma^j_\ell + n^j_{\ell\bullet})(\sum_\ell n^j_{\ell\bullet})}{(\sum_\ell \gamma^j_\ell + \sum_\ell n^j_{\ell\bullet})(n^j_{\ell\bullet} + 1)} \frac{V(n^j_{\ell\bullet} + 1, H^j_{\ell+}, \alpha^j_\ell)}{V(n^j_{\ell\bullet}, H^j_{\ell+}, \alpha^j_\ell)} \cdot (n^j_{\ell h} + \alpha^j_\ell)
\end{aligned}
$$

□

Proof of Proposition 3

Proof: To prove Proposition 3 for any given $j \in \{1, \ldots, J\}$, we consider the posterior contraction rate of the mixing measure G_j given observations $y_{j1}, \ldots, y_{jN_j}$. For the moment we drop all subscripts and superscripts j in the parameters. Then the 888 model under conditions (13) in Lemma 1 can be written as,

$$
\begin{aligned}
y_i \sim \int f(y_i \mid \theta) dG(\theta), \quad G = \sum_{m=1}^{K} \pi_m \delta_{\phi_m}, \\
\boldsymbol{\pi} \sim \mathrm{Dir}(\boldsymbol{\alpha}), \quad \phi_m \sim \Phi, \quad K \sim p_K(K)
\end{aligned}
\tag{15}
$$

where $\boldsymbol{\alpha} = (\alpha_0, \ldots, \alpha_0, \alpha^j_1, \ldots, \alpha^j_1, \ldots, \alpha^j_J, \ldots, \alpha^j_J) = \{\alpha_\ell\}_{\ell=1,\ldots,K}$.

Model (15) is a static MFM model if we replace the value of $\boldsymbol{\alpha}$ by $\boldsymbol{\alpha} = (\gamma/K, \ldots, \gamma/K)$. Based on Guha et al. [29], under the following assumptions on the MFM model,

(P.1) The parameter space Θ is compact, while kernel density function f is first-order identifiable and admits the uniform Lipschitz property up to the first order.

(P.2) The base distribution Φ is absolutely continuous with respect to Lebesgue measure μ on $\mathbb{R}$ and admits a density function $g(\cdot)$. Additionally, Φ is approximately uniform: $\min_{\theta\in\Theta} g(\theta) > c_0 > 0$.

(P.3) There exists $\epsilon_0 > 0$ such that $\int (p_{G^0}(x))^2/p_G(x)d\mu(x) \leq M(\epsilon_0)$ as long as $W_1(G, G^0) \leq \epsilon_0$ for any $G \in \mathbb{O}_{k_0}$ where $M(\epsilon_0)$ depends only on ϵ_0, G^0, and Θ.

(P.4) The prior p_K places positive mass on the set of natural numbers, i.e., $p_K(k) > 0$ for all $k \in \mathbb{N}$.

we have that

$$\Pi\Big(G \in \bar{\mathcal{G}}(\Theta) : W_1(G, G^0) \lesssim (logN/N)^{1/2} \mid y_1, \ldots, y_N\Big) \to 1$$

in probability under the true mixing measure G^0.

We first prove that the model in equation (15) with asymmetric Dirichlet prior $\boldsymbol{\alpha} = \{\alpha_\ell\}_{\ell=1,\ldots,K}$ has the same posterior contraction rate under conditions (P.1), (P.2), (P.3), and (P.4). Following the steps in the proof in Section 6 in [29], we claim that equation (18) in the paper is also true under the prior $Q = \mathrm{Dir}(\boldsymbol{\alpha})$ when $\boldsymbol{\alpha}$ is asymmetric, such that

$$\epsilon^{\gamma'} \lesssim Q(|\pi_m - \pi_m^0| \leq \epsilon,\ m = 1, \ldots, k_0)$$

where k_0 is the number of clusters in use, $\gamma' > 0$ is a positive constant and ϵ is sufficiently small. Therefore, the conclusion on the contraction rate is proved by following the same steps as in Guha et al. [29]. We give the proof of the last claim:

$$Q(|\pi_m - \pi_m^0| \leq \epsilon, m = 1, \ldots, k_0) = \frac{\Gamma(\sum_{m=1}^{k_0} \alpha_m)}{\prod_{m=1}^{k_0} \Gamma(\alpha_m)} \cdot \int_{|\pi_m - \pi_m^0| \leq \epsilon,\ 1 \leq m \leq k_0} \prod_{m=1}^{k_0-1} \pi_m^{\alpha_m - 1}(1 - \sum_{m=1}^{k_0-1} \pi_m)^{\alpha_{k_0}-1} d\pi_1 \ldots d\pi_{k_0-1}$$

If $\alpha_{k_0} \leq 1$, we have

$$\begin{aligned} Q(|\pi_m - \pi_m^0| \leq \epsilon, m = 1, \ldots, k_0) &\geq \frac{\Gamma(\sum_{m=1}^{k_0} \alpha_m)}{\prod_{m=1}^{k_0} \Gamma(\alpha_m)} \int_{|\pi_m - \pi_m^0| \leq \epsilon,\ 1 \leq m \leq k_0} \prod_{m=1}^{k_0-1} \pi_m^{\alpha_m - 1} d\pi_1 \ldots d\pi_{k_0-1} \\ &\gtrsim \frac{\Gamma(\sum_{m=1}^{k_0} \alpha_m)}{\prod_{m=1}^{k_0} \Gamma(\alpha_m)} \frac{1}{\prod_{m=1}^{k_0-1} \alpha_m} \epsilon^{\sum_{m=1}^{k_0-1} \alpha_m} \end{aligned}$$

where the first inequality due to the fact that $1 - \sum_{m=1}^{k_0-1} \pi_m \leq 1$.

If $\alpha_{k_0} > 1$, we have

$$
\begin{aligned}
Q(|\pi_m - \pi_m^0| \le \epsilon, m = 1, \dots, k_0) &\ge \frac{\Gamma(\sum_{m=1}^{k_0} \alpha_m)}{\prod_{m=1}^{k_0} \Gamma(\alpha_m)} \left(\frac{1 - \sum_{m=1}^{k_0-1} \pi_m^0}{2} \right)^{\alpha_{k_0}-1} \cdot \\
&\int_{|\pi_m - \pi_m^0| \le \epsilon,\, 1 \le m \le k_0} \prod_{m=1}^{k_0-1} \pi_m^{\alpha_m - 1} d\pi_1 \dots d\pi_{k_0-1} \\
&\gtrsim \frac{\Gamma(\sum_{m=1}^{k_0} \alpha_m)}{\prod_{m=1}^{k_0} \Gamma(\alpha_m)} \left(\frac{1 - \sum_{m=1}^{k_0-1} \pi_m^0}{2} \right)^{\alpha_{k_0}-1} \frac{1}{\prod_{m=1}^{k_0-1} \alpha_m} \epsilon^{\sum_{m=1}^{k_0-1} \alpha_m}
\end{aligned}
$$

The first inequality in the above display is true because when $\pi_m^0 > 0$ for all $m \in \{1, \dots, k_0\}$ for sufficient small $\epsilon > 0$ such that $|\pi_m - \pi_m^0| \le \epsilon$, we have $\pi_{k_0} \ge \pi_{k_0}^0/2$.

We are now ready to show that the proposed 888 model satisfies the condition for consistency. We can restrict the parameters to a compact space, the kernel density f is a location-scale multivariate Gaussian distribution so that it is first order identifiable and admits the uniform Lipschitz property up to the first order (P.1), and the assumption (P.3) which controls the growing rate of KL neighborhood also holds for multivariate Gaussian. The prior distribution on the atoms in our model is continuous and approximately uniform (P.2). A prior $p(H)$ for the number of clusters that places positive mass on the set of natural number is assumed so that (P.4) is satisfied.

□