Global-Local Dirichlet Processes for Clustering Grouped Data in the Presence of Group-Specific Idiosyncratic Variables

Arhit Chakrabarti[1], Yang Ni[1], Debdeep Pati[2], and Bani K. Mallick[1]

[1] Department of Statistics, Texas A&M University

[2] Department of Statistics, University of Wisconsin-Madison

Forty-Second International Conference on Machine Learning (ICML), 2025

Motivation

- Cancer is not a single disease
- Classification of cancer – a step towards personalized treatment
- Traditional classification of cancer restricted to tumor site of origin
- Recent studies have shown different cancers have significant clinical and molecular similarities
- Classification is done through genomic similarities
- Clinical variables/prognostic biomarkers not used in the classification

Pan-Gastrointestinal Cancer

- Cancers of the GI Tract:
  - Esophageal and stomach $\rightarrow$ upper GI tract
  - Colon and rectal $\rightarrow$ lower GI tract
- Cluster patients within each cancer while allowing clusters to be shared across the cancers
  - Incorporate cancer-specific clinical variables and/or and prognostic biomarkers
  - **Challenge**: clinical variables and prognostic biomarkers are cancer-specific
- Existing grouped clustering methods consider only shared genomic information across cancers
  - Hierarchical Dirichlet Process (HDP, [Teh et al., 2006])
  - Nested Dirichlet Process [Rodríguez et al., 2008]
  - Hidden Hierarchical Dirichlet Process [Lijoi et al., 2023]
  - Common Atoms Model [Denti et al., 2023]
- Disregards valuable information from clinical variables and/or prognostic biomarkers

Proposed Solution: Global-Local (GLocal) Dirichlet Process

- Incorporate cancer-specific clinical (local) variables in the clustering of grouped data
- For cancer j and patient i
  - $\boldsymbol{x}_{ji}^{L} \rightarrow$ cancer-specific clinical/biomarker data (not available for all cancers) $\rightarrow$ local variables
  - $\boldsymbol{x}_{ji}^{G} \rightarrow$ genomic data on common genes (shared across cancers) $\rightarrow$ global variables
  - $\boldsymbol{x}_{ji} = \left(\boldsymbol{x}_{ji}^{L}, \boldsymbol{x}_{ji}^{G}\right)$
  - Consider the sampling distribution

$$F(\boldsymbol{x}_{ji} \mid \boldsymbol{\theta}_{ji}) = F_1(\boldsymbol{x}_{ji}^{L} \mid \boldsymbol{\theta}_{ji}^{L})\, F_2(\boldsymbol{x}_{ji}^{G} \mid \boldsymbol{\theta}_{ji}^{G}). \tag{1}$$

GLocal Dirichlet Process

- GLocal DP defined by

$$
\begin{aligned}
(\boldsymbol{\theta}_{ji}^{L}, \boldsymbol{\theta}_{ji}^{G}) = \boldsymbol{\theta}_{ji} \mid G_j &\sim G_j, \\
G_j \mid \alpha, V &\sim \mathrm{DP}(\alpha, U_j \otimes V), \\
V \mid \gamma &\sim \mathrm{DP}(\gamma, H).
\end{aligned} \tag{2}
$$

- Since U_j is group-specific, G_j does not share atoms
- Global atoms are marginally shared through V

GLocal DP: Stick Breaking Representation

- Since $V \mid \gamma \sim \text{DP}(\gamma, H)$

$$\begin{aligned} V &= \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \\ \boldsymbol{\beta} &= (\beta_k)_{k=1}^{\infty} \mid \gamma \sim \text{GEM}(\gamma), \\ \phi_k &\overset{iid}{\sim} H. \end{aligned} \tag{3}$$

- Since $G_j \mid \alpha, V \sim \text{DP}(\alpha, U_j \otimes V)$

$$\begin{aligned} G_j &= \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}, \\ \boldsymbol{\pi}_j &= (\pi_{jt})_{t=1}^{\infty} \mid \alpha \sim \text{GEM}(\alpha), \\ \psi_{jt} &= \left(\psi_{jt}^{L}, \psi_{jt}^{G}\right) \mid V \overset{ind}{\sim} U_j \otimes V. \end{aligned} \tag{4}$$

- $\left(\psi_{jt}^{G}\right)_{t=1}^{\infty}$ is necessarily same as $(\phi_k)_{k=1}^{\infty}$
- In fact, ψ_{jt}^{G} takes the value ϕ_k with probability β_k

GLocal DP: Global and Local Clusters

$$\left(\boldsymbol{\theta}_{ji}^{L}, \boldsymbol{\theta}_{ji}^{G}\right) = \boldsymbol{\theta}_{ji} \mid G_j \sim G_j, \qquad G_j \mid \alpha, V \sim \mathrm{DP}(\alpha, U_j \otimes V),$$

$$G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}, \quad \boldsymbol{\pi}_j = (\pi_{jt})_{t=1}^{\infty} \mid \alpha \sim \mathrm{GEM}(\alpha), \quad \psi_{jt} = \left(\psi_{jt}^{L}, \psi_{jt}^{G}\right) \mid V \overset{ind}{\sim} U_j \otimes V, \tag{5}$$

$$V \mid \gamma \sim \mathrm{DP}(\gamma, H), \quad V = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \quad \boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \mid \gamma \sim \mathrm{GEM}(\gamma), \quad \phi_k \overset{iid}{\sim} H.$$

- Introduce

$$t_{ji} \mid \boldsymbol{\pi}_j \overset{ind}{\sim} \boldsymbol{\pi}_j \tag{6}$$

$$k_{jt} \mid \boldsymbol{\beta} \overset{ind}{\sim} \boldsymbol{\beta} \tag{7}$$

GLocal DP: Global and Local Clusters

$$\left(\theta_{ji}^{L}, \theta_{ji}^{G}\right) = \theta_{ji} \mid G_j \sim G_j, \qquad G_j \mid \alpha, V \sim \mathrm{DP}(\alpha, U_j \otimes V),$$

$$G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}, \quad \boldsymbol{\pi}_j = (\pi_{jt})_{t=1}^{\infty} \mid \alpha \sim \mathrm{GEM}(\alpha), \quad \psi_{jt} = \left(\psi_{jt}^{L}, \psi_{jt}^{G}\right) \mid V \overset{ind}{\sim} U_j \otimes V, \tag{5}$$

$$V \mid \gamma \sim \mathrm{DP}(\gamma, H), \quad V = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \quad \boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \mid \gamma \sim \mathrm{GEM}(\gamma), \quad \phi_k \overset{iid}{\sim} H.$$

- Introduce

$$t_{ji} \mid \boldsymbol{\pi}_j \overset{ind}{\sim} \boldsymbol{\pi}_j \tag{6}$$

$$k_{jt} \mid \boldsymbol{\beta} \overset{ind}{\sim} \boldsymbol{\beta} \tag{7}$$

  - t_{ji} is the *local-level* cluster label

GLocal DP: Global and Local Clusters

$$\left(\theta_{ji}^{L}, \theta_{ji}^{G}\right) = \theta_{ji} \mid G_j \sim G_j, \qquad G_j \mid \alpha, V \sim \mathrm{DP}(\alpha, U_j \otimes V),$$

$$G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}, \quad \boldsymbol{\pi}_j = (\pi_{jt})_{t=1}^{\infty} \mid \alpha \sim \mathrm{GEM}(\alpha), \quad \psi_{jt} = \left(\psi_{jt}^{L}, \psi_{jt}^{G}\right) \mid V \overset{ind}{\sim} U_j \otimes V, \tag{5}$$

$$V \mid \gamma \sim \mathrm{DP}(\gamma, H), \quad V = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \quad \boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \mid \gamma \sim \mathrm{GEM}(\gamma), \quad \phi_k \overset{iid}{\sim} H.$$

- Introduce

$$t_{ji} \mid \boldsymbol{\pi}_j \overset{ind}{\sim} \boldsymbol{\pi}_j \tag{6}$$

$$k_{jt} \mid \boldsymbol{\beta} \overset{ind}{\sim} \boldsymbol{\beta} \tag{7}$$

  - t_{ji} is the *local-level* cluster label
  - $k_{jt_{ji}}$ is the *global-level* cluster label

GLocal DP: Global and Local Clusters

$$\left(\boldsymbol{\theta}_{ji}^{L}, \boldsymbol{\theta}_{ji}^{G}\right) = \boldsymbol{\theta}_{ji} \mid G_j \sim G_j, \qquad G_j \mid \alpha, V \sim \mathrm{DP}(\alpha, U_j \otimes V),$$

$$G_j = \sum_{t=1}^{\infty} \pi_{jt} \delta_{\psi_{jt}}, \quad \boldsymbol{\pi}_j = (\pi_{jt})_{t=1}^{\infty} \mid \alpha \sim \mathrm{GEM}(\alpha), \quad \psi_{jt} = \left(\psi_{jt}^{L}, \psi_{jt}^{G}\right) \mid V \overset{ind}{\sim} U_j \otimes V, \tag{5}$$

$$V \mid \gamma \sim \mathrm{DP}(\gamma, H), \quad V = \sum_{k=1}^{\infty} \beta_k \delta_{\phi_k}, \quad \boldsymbol{\beta} = (\beta_k)_{k=1}^{\infty} \mid \gamma \sim \mathrm{GEM}(\gamma), \quad \phi_k \overset{iid}{\sim} H.$$

- Introduce

$$t_{ji} \mid \boldsymbol{\pi}_j \overset{ind}{\sim} \boldsymbol{\pi}_j \tag{6}$$

$$k_{jt} \mid \boldsymbol{\beta} \overset{ind}{\sim} \boldsymbol{\beta} \tag{7}$$

  - t_{ji} is the *local-level* cluster label
  - $k_{jt_{ji}}$ is the *global-level* cluster label
- Local clusters are *nested* within the global clusters
  - two individuals belonging to same local cluster share the same global cluster but not vice versa

GLocal DP Mixture Model and Finite Mixture Representation

GLocal DP Mixture model is given by,

$$
\begin{aligned}
&\beta \mid \gamma \sim \text{GEM}(\gamma), && k_{jt} \mid \beta \sim \beta, \\
&\pi_j \mid \alpha \sim \text{GEM}(\alpha), && t_{ji} \mid \pi_j \sim \pi_j, \\
&\phi_k \sim H, && \psi^L_{jt} \sim U_j, \\
&\boldsymbol{x}_{ji} \mid (\phi_k)_{k=1}^{\infty}, (\psi^L_{jt})_{t=1}^{\infty}, t_{ji}, (k_{jt})_{t=1}^{\infty} \sim F_1(\boldsymbol{x}^L_{ji} \mid \psi^L_{jt_{ji}}) F_2(\boldsymbol{x}^G_{ji} \mid \phi_{k_{jt_{ji}}}).
\end{aligned}
\tag{8}
$$

GLocal DP Mixture Model and Finite Mixture Representation

GLocal DP Mixture model is given by,

$$\begin{aligned}
&\beta \mid \gamma \sim \text{GEM}(\gamma), && k_{jt} \mid \beta \sim \beta, \\
&\pi_j \mid \alpha \sim \text{GEM}(\alpha), && t_{ji} \mid \pi_j \sim \pi_j, \\
&\phi_k \sim H, && \psi^L_{jt} \sim U_j, \\
&\boldsymbol{x}_{ji} \mid (\phi_k)_{k=1}^{\infty}, (\psi^L_{jt})_{t=1}^{\infty}, t_{ji}, (k_{jt})_{t=1}^{\infty} \sim F_1(\boldsymbol{x}^L_{ji} \mid \psi^L_{jt_{ji}}) F_2(\boldsymbol{x}^G_{ji} \mid \phi_{k_{jt_{ji}}}).
\end{aligned} \tag{8}$$

The finite mixture model representation is,

$$\begin{aligned}
&\beta \mid \gamma \sim \text{Dir}(\gamma/L, \ldots, \gamma/L), && k_{jt} \mid \beta \sim \beta, \\
&\pi_j \mid \alpha \sim \text{Dir}(\alpha/T, \ldots, \alpha/T), && t_{ji} \mid \pi_j \sim \pi_j, \\
&\phi_k \sim H, && \psi^L_{jt} \sim U_j, \\
&\boldsymbol{x}_{ji} \mid (\phi_k)_{k=1}^{L}, (\psi^L_{jt})_{t=1}^{T}, t_{ji}, (k_{jt})_{t=1}^{T} \sim && F_1(\boldsymbol{x}^L_{ji} \mid \psi^L_{jt_{ji}}) F_2(\boldsymbol{x}^G_{ji} \mid \phi_{k_{jt_{ji}}}),
\end{aligned} \tag{9}$$

with $L \leq T$. Then, as $L \to \infty$, model (9) converges to model (8)

Truncation Error Bounds I

Proposition 1. *Let $P^{\infty,\infty}(\boldsymbol{\theta})$ and $P^{T,K}(\boldsymbol{\theta})$ denote the prior distribution of the parameters $\boldsymbol{\theta}$ under the GLocal DP prior and its corresponding truncated version with the random measures integrated out. Furthermore, let $m^{\infty,\infty}(\boldsymbol{x})$ and $m^{T,K}(\boldsymbol{x})$ denote the marginal distribution of the data $\boldsymbol{x}$, derived from these priors. Then,*

$$\int_{\mathcal{X}^N} \left| m^{T,K}(\boldsymbol{x}) - m^{\infty,\infty}(\boldsymbol{x}) \right| d\boldsymbol{x} \leq \int_{\Xi^N} \left| P^{T,K}(\boldsymbol{\theta}) - P^{\infty,\infty}(\boldsymbol{\theta} \right| d\boldsymbol{\theta} \leq \epsilon^{T,K}(\alpha, \gamma),$$

where

$$\epsilon^{T,K}(\alpha,\gamma) = 4\left[1 - \left\{\left(1 - \left(\frac{\alpha}{1+\alpha}\right)^{T-1}\right)\right\}^N \left\{\left(1 - \left(\frac{\gamma}{1+\gamma}\right)^{K-1}\right)\right\}^N\right],$$

$N = n_1 + \cdots + n_J$, $\Xi^N = \prod_{j=1}^{J} (\Theta_j \times \Omega)^{n_j}$, and $\mathcal{X}^N$ denotes the sample space of observations $\boldsymbol{x}$.

Truncation Error Bounds II

Proposition 2. *The posterior distribution of the parameters $\boldsymbol{\theta}$ under the GLocal DP prior and its truncated version,*

$$\pi^{\infty,\infty}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})P^{\infty,\infty}(\boldsymbol{\theta})}{m^{\infty,\infty}(\boldsymbol{x})}, \qquad \pi^{T,K}(\boldsymbol{\theta}|\boldsymbol{x}) = \frac{f(\boldsymbol{x}|\boldsymbol{\theta})P^{T,K}(\boldsymbol{\theta})}{m^{T,K}(\boldsymbol{x})},$$

satisfies

$$\int_{\mathcal{X}^N}\int_{\Xi^N} \left|\pi^{T,K}(\boldsymbol{\theta}|\boldsymbol{x}) - \pi^{\infty,\infty}(\boldsymbol{\theta}|\boldsymbol{x})\right| m^{\infty,\infty}(\boldsymbol{x})\, d\boldsymbol{\theta}\, d\boldsymbol{x} = \mathcal{O}\left(\epsilon^{T,K}(\alpha,\gamma)\right).$$

Variational Posterior Inference

- Truncated variational distribution: truncation levels denoted by T and K.

$$
\begin{aligned}
& q(\boldsymbol{t}, \boldsymbol{k}, \boldsymbol{u}, \boldsymbol{v}, \boldsymbol{\phi}, \boldsymbol{\psi}, \alpha, \gamma; \boldsymbol{\lambda}) \\
&= \prod_{j=1}^{J} \prod_{i=1}^{n_j} q(t_{ji}; \{\xi_{jit}\}_{t=1}^{T}) \prod_{j=1}^{J} \prod_{t=1}^{T} q(k_{jt}; \{\rho_{jtl}\}_{l=1}^{K}) \prod_{j=1}^{J} \prod_{t=1}^{T-1} q(u_{jt}; \bar{a}_{jt}, \bar{b}_{jt}) \times \\
&\quad \times \prod_{k=1}^{K-1} q(v_k; \bar{a}_k, \bar{b}_k) \times \prod_{k=1}^{K} q(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k; \boldsymbol{m}_k, \lambda_k, c_k, \boldsymbol{D}_k)\, q(\gamma; r_1, r_2) \times \\
&\quad \times \prod_{j=1}^{J} \prod_{t=1}^{T} q(\boldsymbol{\mu}_{jt}, \boldsymbol{\Lambda}_{jt}; \boldsymbol{m}_{jt}, \lambda_{jt}, c_{jt}, \boldsymbol{D}_{jt})\, q(\alpha; s_1, s_2),
\end{aligned}
$$

- Optimal variational parameters: maximize the Evidence Lower Bound $\rightarrow$ coordinate ascent variational inference algorithm. [Blei and Jordan, 2006]

Simulations

- Three groups. $p = 3$, $p_1 = 2$, $p_2 = 3$, and $p_3 = 4$. $L_{\ell_1} = 3$, $L_{\ell_2} = 5$, $L_{\ell_3} = 4$, and $L_g = 6$. $n_j = 200$ samples. Data generated from:

$$\boldsymbol{x}_{ji} \sim \left\{ \sum_{t=1}^{L_{\ell_j}} \pi_{jt} \mathcal{N}_{p_j}(\boldsymbol{x}_{ji}^L \mid \boldsymbol{\mu}_{jt}, \Lambda_{jt}^{-1}) \right\} \left\{ \sum_{k=1}^{L_g} \beta_k \mathcal{N}_p(\boldsymbol{x}_{ji}^G \mid \boldsymbol{\mu}_k, \Lambda_k^{-1}) \right\}.$$

- Local parameters:

$$(\boldsymbol{\mu}_{jt}, \boldsymbol{\Lambda}_{jt}) \sim \text{NW}(\mathbf{0}, \lambda_L^{-1}, 5p_j, \mathbb{I}_{p_j}), \quad \alpha \sim \text{Gamma}(25, 1),$$
$$\boldsymbol{\pi}_j \sim \text{Dir}(\alpha/L_{\ell_j}, \ldots, \alpha/L_{\ell_j}).$$

- Global parameters:

$$(\boldsymbol{\mu}_k, \boldsymbol{\Lambda}_k) \sim \text{NW}(\mathbf{0}, \lambda_G^{-1}, 15, \mathbb{I}_3), \quad \gamma \sim \text{Gamma}(25, 1),$$
$$\boldsymbol{\beta} \sim \text{Dir}(\gamma/L_g, \ldots, \gamma/L_g).$$

- t_{ji} and k_{jt} drawn from multinomials with parameters $\boldsymbol{\pi}_j$ and $\boldsymbol{\beta}$.
- Fix $\lambda_G = 1$ (low global variable separation). Varied $\lambda_L = 0.1, 0.05, 0.01$ (control local variable separation across groups).

Simulation Results

- Compared GLocal DP (on global and local variables), HDP (on global only) at *global-level*.
- Compared GLocal DP with per-group GMMs and DPMs at *local-level*.
- Clustering performance measured by ARI. Results replicated 50 times.

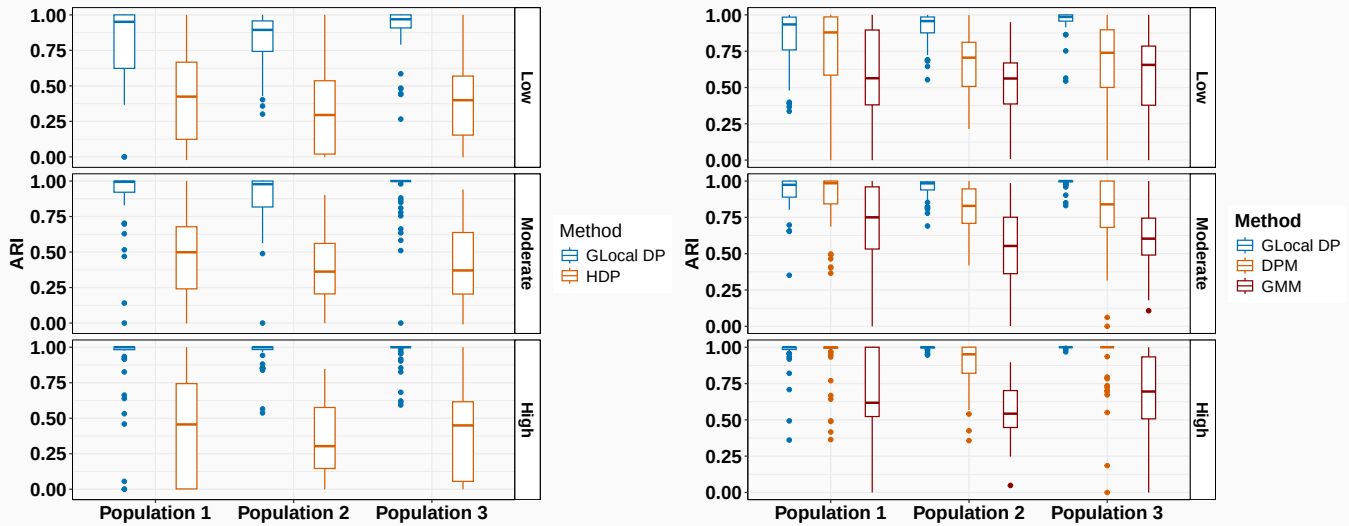
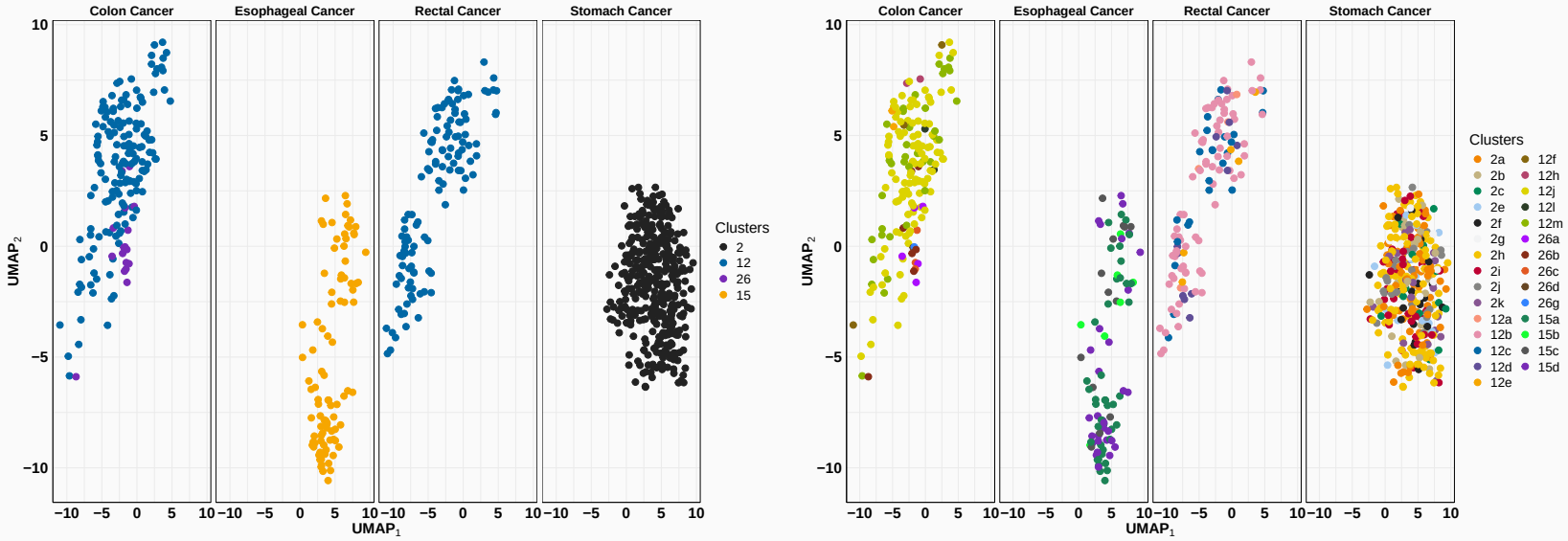


Figure 1: Comparison of clustering performance of GLocal DP with competing methods for varying separation of local variables.

Real Data Analysis: Pan-Gastrointestinal Cancer

- Esophageal cancer ($n = 92$)
  - Local variable: Number of cigarettes smoked per day
- Stomach cancer ($n = 363$)
  - Local variable: Number of positive lymph nodes
- Colon cancer ($n = 173$)
  - Local variable: Carcinoembryonic antigen level (CEA), body mass index (BMI)
- Rectal cancer ($n = 120$)
  - Local variable: Carcinoembryonic antigen level (CEA)
- Global variables: gene expression measured on a common gene set

Pan-GI Cancer: Global and Local Clustering



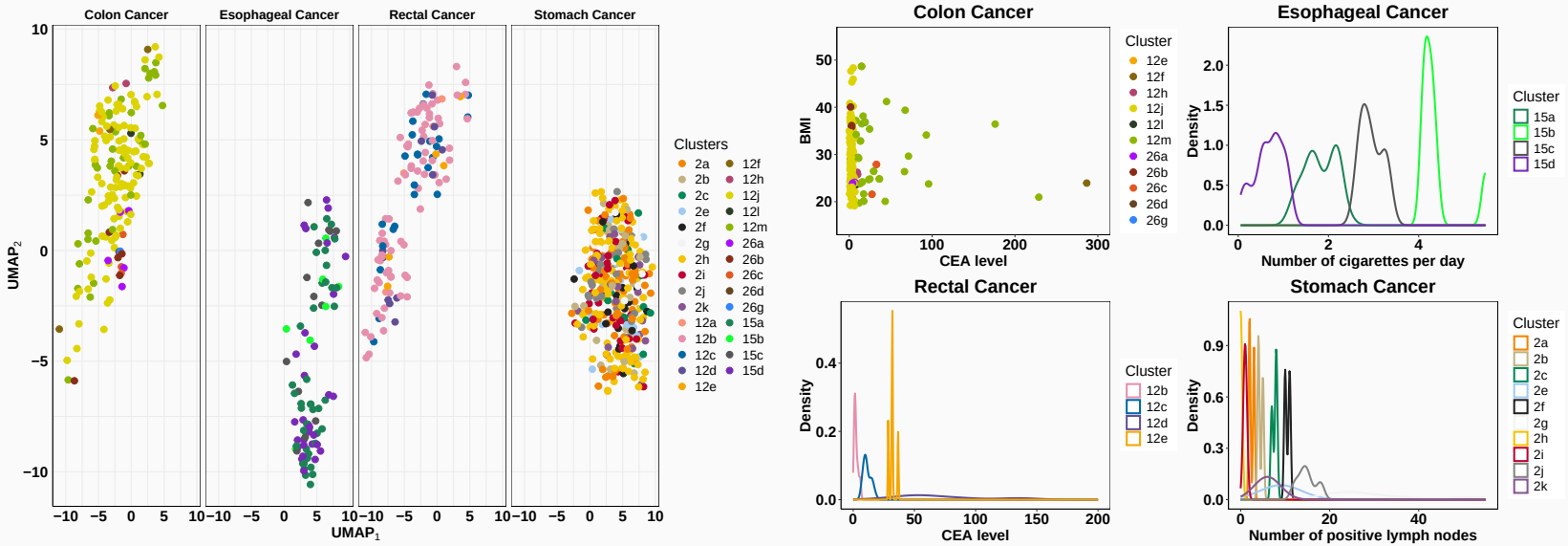
(a) Global-level clusters.

(b) Local-level clusters.

Figure 2: Global variables: Dimension reduced to 2 by UMAP [McInnes et al., 2018] for visualization. (a) The colors indicate global-level clusters estimated from GLocal DP. (b) The colors indicate the estimated local-level clusters.

- Rectal and colon cancers share global clusters $\rightarrow$ the cancers are similar
- Colon cancer has unique subpopulation (cluster 26) not found in rectal cancer $\rightarrow$ patients have high BMI and higher median CEA

Pan-GI Cancer: Local Clustering



(a) Local-level clusters.

(b) Kernel density/scatter plot.

Figure 3: (a) The estimated local-level clusters for global variables. (b) Plot of local variables colored by the estimated local-level clusters.

- Distinct clusters are observed among esophageal cancer patients based on smoking history
- Corresponding gene expression patterns for these subgroups show differences

Pan-GI Cancer: Local Survival

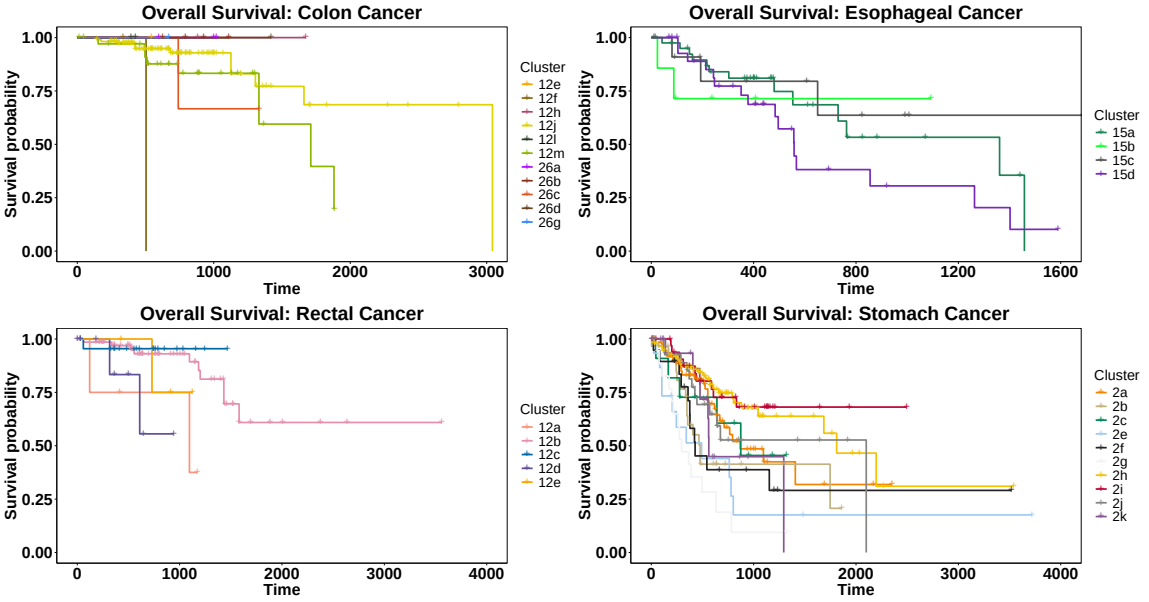


Figure 4: Kaplan-Meier survival curves by local-level sub-clusters induced by cancer-specific clinical variables estimated from GLocal DP.

- Esophageal cancer: survival curves show distinct trajectories associated with smoking history
- Stomach cancer: subcluster 2g characterized by patients with a very high number of positive lymph nodes.
  - exhibits poorer survival outcomes compared to other groups
  - compared to those with fewer positive lymph nodes.

Conclusion and Future Direction

Contributions:

- Introduced GLocal DP for modeling group of random measures that account for varying variable sets
- Presented different representations of GLocal DP
- Introduced GLocal DP finite mixture model (GLocal DP-MM)
- Proposed a highly scalable variational Bayes posterior inference algorithm
- Presented use of GLocal DP-MM for clustering pan-cancer data incorporating cancer-specific clinical variables

Future Work:

- Extend model to incorporate the group-clustering feature of the nested DP along with the cluster-sharing feature of the HDP
  - Possibly provide insights on similar cancer subtypes
  - Clustering shared observations across the tumor subtypes
  - Cancer-specific clinical variables refine the clusters shared across cancers

References i


Blei, D. M. and Jordan, M. I. (2006).
Variational inference for Dirichlet process mixtures.
Bayesian Analysis, 1(1):121 – 143.


Denti, F., Camerlenghi, F., Guindani, M., and Mira, A. (2023).
A Common Atoms Model for the Bayesian Nonparametric Analysis of Nested Data.
Journal of the American Statistical Association, 118(541):405–416.


Lijoi, A., Prünster, I., and Rebaudo, G. (2023).
Flexible Clustering via Hidden Hierarchical Dirichlet priors.
Scandinavian Journal of Statistics, 50(1):213–234.


McInnes, L., Healy, J., Saul, N., and Großberger, L. (2018).
UMAP: Uniform Manifold Approximation and Projection.
Journal of Open Source Software, 3(29):861.

References ii


Rodríguez, A., Dunson, D. B., and Gelfand, A. E. (2008).
The Nested Dirichlet Process.
Journal of the American Statistical Association,
103(483):1131–1154.


Teh, Y. W., Jordan, M. I., Beal, M. J., and Blei, D. M. (2006).
Hierarchical Dirichlet Processes.
Journal of the American Statistical Association,
101(476):1566–1581.