Bayesian Multi-Domain Learning for Cancer Subtype Discovery

Ehsan Hajireamezanali* Siamak Dadaneh* Alireza Karbalayghareh* Mingyuan Zhou** Xiaoning Qian*

*Department of Electrical and Computer Engineering, Texas A&M University
**McCombs School of Business, University of Texas at Austin

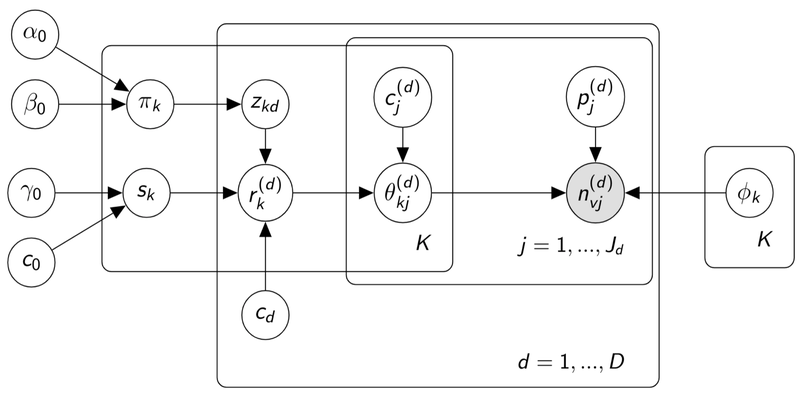
Introduction

The **Bayesian Multi-Domain Learning (BMDL)** is proposed to analysis overdispersed count data from next-generation sequencing (NGS) experiments, with the goal of enhancing cancer subtyping in the *target domain* with a limited number of NGS samples by leveraging surrogate data from other domains.

- The **BMDL** is a hierarchical negative binomial factorization model for NGS counts, to derive domain-dependent latent representations allowing both domain-specific and globally shared latent factors.
- The over-dispersion is appropriately modeled and *ad-hoc* pre-processing is not needed.
- Low-dimensional representations of counts in different domains can help achieve more robust subtyping results.
- The sample relevance across domains can be explicitly learned to guarantee the effectiveness of joint learning across multiple domains.

Model and Inference

❖ Probabilistic Modeling



❑ Hierarchical Negative Binomial Factorization:

$$n_{vj}^{(d)} = \sum_{k=1}^{K} n_{vjk}^{(d)}, \qquad n_{vjk}^{(d)} \sim \mathrm{NB}(\phi_{vk}\theta_{kj}^{(d)}, p_j^{(d)}),$$
$$p_j^{(d)} \sim \mathrm{Beta}(a_0, b_0),$$
$$(\phi_{1k}, \ldots, \phi_{Vk}) \sim \mathrm{Dir}(\eta, \ldots, \eta), \quad \eta \sim \mathrm{Gamma}(s_0, w_0),$$
$$\theta_{kj}^{(d)} \sim \mathrm{Gamma}(r_k^{(d)}, 1/c_j^{(d)}), \quad c_j^{(d)} \sim \mathrm{Gamma}(e_0, 1/f_0),$$
$$r_k^{(d)} \sim \mathrm{Gamma}(s_k z_{kd}, 1/c_d), \quad s_k \sim \mathrm{Gamma}(\gamma_0/K, 1/c_0),$$

❑ Domain-Dependent Selector:

$$z_{kd} \sim \mathrm{Bernoulli}(\pi_k), \qquad \pi_k \sim \mathrm{Beta}(\alpha_0, \beta_0),$$

- The domain-dependent binary variables enable sparse domain-dependent latent representations of count data that help explicitly establish the relevance across domains.
- Across domains, the shared loading factors help achieve more robust inference when the number of samples in target domain is low.
- BMDL automatically identifies both domain-specific and globally shared latent factors in different domains and avoid negative transferring across domains.
- The shrinkage property of the priors for the factor strength parameter s_k, facilitates learning of the number of factors K in practice.

❖ Gibbs Sampling via Data Augmentation and Marginalization

$$(\ell_{vj}^{(d)}|-) \sim \mathrm{CRT}\left(n_{vj}^{(d)}, \sum_{k=1}^{K}\phi_{vk}\theta_{kj}^{(d)}\right)$$
$$(\ell_{vj1}^{(d)}, \ldots, \ell_{vjK}^{(d)}|-) \sim \mathrm{Mult}\left(\ell_{vj}^{(d)}; \frac{\phi_{v1}\theta_{1j}^{(d)}}{\sum_{k=1}^{K}\phi_{vk}\theta_{kj}^{(d)}}, \ldots, \frac{\phi_{vK}\theta_{Kj}^{(d)}}{\sum_{k=1}^{K}\phi_{vk}\theta_{kj}^{(d)}}\right)$$
$$(\phi_{1k}, \ldots, \phi_{Vk}|-) \sim \mathrm{Dir}(\eta + \ell_{1.k}^{(.)}, \ldots, \eta + \ell_{V.k}^{(.)})$$
$$\theta_{kj}^{(d)} \sim \mathrm{Gamma}\left(r_k^{(d)} + \ell_{.jk}^{(d)}, \frac{1}{c_j^{(d)} - q_j^{(d)}}\right),$$
$$(\tilde{\ell}_{jk}^{(d)}|-) \sim \mathrm{CRT}(\ell_{.jk}^{(d)}, r_k^{(d)}),$$
$$(r_k^{(d)}|-) \sim \mathrm{Gamma}\left(z_{kd}s_k + \tilde{\ell}_{.k}^{(d)}, \frac{1}{c_k - \sum_j \ln(1-\tilde{p}_j^{(d)})}\right)$$
$$(\tilde{\tilde{\ell}}_k^{(d)}|-) \sim \mathrm{CRT}(\tilde{\ell}_{.k}^{(d)}, s_k),$$
$$(s_k|-) \sim \mathrm{Gamma}\left(\gamma_0/K + \sum_d \tilde{\tilde{\ell}}_k^{(d)}, \frac{1}{c_0 - \tilde{q}_k}\right),$$
$$(\acute{\ell}_k|-) \sim \mathrm{CRT}(\sum_d \tilde{\tilde{\ell}}_k^{(d)}, \gamma_0/K),$$
$$(\gamma_0|-) \sim \mathrm{Gamma}\left(a_0 + \acute{\ell}_{.}, \frac{1}{b_0 - \sum_k \ln(1-\tilde{q}_k)/K}\right),$$
$$(z_{kd}|-) \sim \delta(\tilde{\ell}_{.k}^{(d)} = 0)\mathrm{Bernoulli}\left(\frac{(\tilde{q}_k^{(d)})^{s_k}\pi_k}{(\tilde{q}_k^{(d)})^{s_k}\pi_k + (1-\pi_k)}\right) + \delta(\tilde{\ell}_{.k}^{(d)} > 0),$$
$$(q_k|-) \sim \mathrm{Beta}(\ell_{..k}^{(.)}, \eta V), \quad u_{vk} \sim \mathrm{CRT}(\ell_{v.k}^{(.)}, \eta),$$
$$(\eta|-) \sim \mathrm{Gamma}\left(s_0 + \sum_{k,v} u_{kv}, \frac{1}{w_0 - V\sum_k \ln(1-q_k)}\right),$$
$$(p_j^{(d)}|-) \sim \mathrm{Beta}(a_0 + \sum_v n_{vj}^{(d)}, b_0 + \sum_k \theta_{jk}^{(d)}).$$

❖ Scalability for Deeply Sequenced NGS data

Rather than sampling by $(\ell_{vj}^{(d)}|-) \sim \mathrm{CRT}\left(n_{vj}^{(d)}, \sum_{k=1}^{K}\phi_{vk}\theta_{kj}^{(d)}\right)$ we approximate it as follows:

$$\begin{aligned}\mathrm{CRT}(n,r) &= \sum_{i=1}^{n}\mathrm{Bernoulli}\left(\frac{r}{i-1+r}\right)\\ &= \sum_{i=1}^{m}\mathrm{Bernoulli}\left(\frac{r}{i-1+r}\right) + \sum_{i=m+1}^{n}\mathrm{Bernoulli}\left(\frac{r}{i-1+r}\right)\\ &= \mathrm{CRT}(m,r) + \mathrm{Pois}(\lambda),\\ \lambda &= \sum_{i=1}^{n}\frac{r}{i-1+r} = r[\psi(n+r) - \psi(m+r)].\end{aligned}$$

This approximation reduces the computational complexity for sampling all $\ell_{vjk}^{(d)}$ from $O[\sum_d\sum_v\sum_j n_{vj}^{(d)}K]$ to $O[\sum_d\sum_v\sum_j \min(n_{vj}^{(d)}, m)K]$ which can lead to significant computation saving for a large number of genes where large counts are abundant.

Results

❖ Experimental Setup

➢ Training Procedure

- Fix the truncation level $K = 100$.
- Consider 3,000 Gibbs sampling iterations.
- Retain the weights $\{r_k^d\}_{1,k}$ and $\{\phi_k\}_{1,k}$ as factors.
- Use the last MCMC sample for the test procedure.

➢ Test Procedure

- Apply 1, 000 blocked Gibbs sampling iterations.
- Collect the last 500 MCMC samples to estimate the posterior mean of the latent factor score $\theta_j^{(d_t)}$.

➢ Classification

- Train a linear support vector machine (SVM) classifier on all $\bar{\theta}_j^{(d_t)}$ in the training set.
- Classify each $\bar{\theta}_j^{(d_t)}$ in the test set.

❖ Compare BMDL to other Bayesian latent models

➢ NB-HDP [Zhou and Carin, 2012]

Under the NB process and integrated to HDP, NB-HDP employed a Dirichlet process (DP) to model the rate measure of a Poisson process ($p_j^{(d)} = 0.5$).

➢ HDP-NBFA

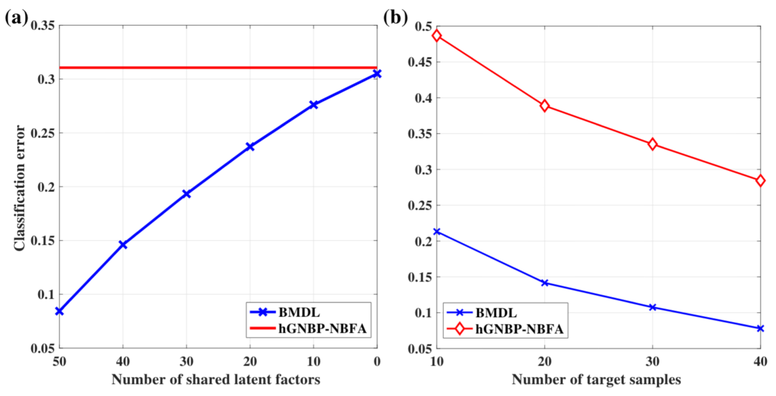
To have a fair comparison and make sure the superior performance of BMDL is not only due to the modeling of the sequencing depth variation across samples, we apply HDP to model latent scores in NB factorization as well. More specifically we model $\theta_{kj}^{(d)}$ as independent gamma random variables with scale parameter 1.

➢ hGNBP [Zhou et al., 2018]

To evaluate the advantages of the beta-Bernoulli modeling in BMDL, we compare the results with hGNBP, which z_{kd} is set to 1.

❖ Synthetic Data Experiments

The classification error of BMDL and hGNBP-NBFA as a function of (a) domain relevance, and (b) the number of target samples.



❖ Case study: Lung cancer

- Two subtypes of lung cancer, i.e. Lung Adenocarcinoma (LUAD) and Lung Squamous Cell Carcinoma (LUSC) from TCGA [The Cancer Genome Atlas Research Network et al., 2008]
- High-related: RNA-seqV2 of LUAD and LUSC
- Low-related: Head and Neck Squamous Cell Carcinoma (HNSC)

Lung cancer subtyping results (average accuracy (%) and STD)

Method	High-related (N_s)		Low-related (N_s)	
	25	100	25	100
NB-HDP	55.22±3.69	56.52±4.61	54.57±7.73	53.83±7.79
HDP-NBFA	63.48±1.23	65.65±4.22	54.89±7.38	51.83±8.32
hGNBP	74.13±7.07	77.61±3.54	72.94±1.70	74.55±8.84
BMDL	**78.46**±5.97	**81.49**±5.12	**78.85**±4.55	**78.10**±5.65
hGNBP-NBFA	73.38 ± 7.29			
Raw Counts	59.28 ± 5.54			

Conclusions

➢ By introducing the hierarchical Bayesian model with selector variables to flexibly assign both domain-specific and globally shared latent factors to different domains, the derived latent representations of NGS data preserves predictive information in corresponding domains so that accurate cancer subtyping is possible even with a limited number of samples.

➢ As BMDL learns domain relevance based on given samples across domains and enables the flexibly of sharing useful information through common latent factors (if any), BMDL performs consistently better than single-domain learning regardless of the domain relevance level.

➢ Across domains, the shared loading factors help achieve more robust inference when the number of samples in target domain is low, which often is the case when analyzing biomedical data.

➢ A new approximation methods are proposed in Gibbs sampling inference to significantly reduce the computational complexity.

➢ The results show that 1) using more domains with more samples helps subtyping in target domain; 2) BMDL avoids negative transfer even when adding low-related domains.

Acknowledgment

We thank Texas A&M High Performance Research Computing and Texas Advanced Computing Center for providing computational resources to perform experiments in this work. This work was supported in part by the NSF Awards CCF-1553281, IIS-1812641, and IIS-1812699.