BayReL: Bayesian Relational Learning for Multi-omics Data Integration

Ehsan Hajiramezanali *, Arman Hasanzadeh *,
Nick Duffield, Krishna R Narayanan,
Xiaoning Qian, * = co-first authors

TEXAS A&M UNIVERSITY

Introduction

BayReL is a novel Bayesian relational learning method:

- Integrating high-dimensional multi-omics data
- Taking advantage of a priori known relationships modeled as a graph at each corresponding view
- Inferring the relational interactions as a bipartite graph without any pre-known interactions across views
- Exploiting non-linear and deep transformations of data
- Enabling Bayesian interpretation

Method

Embedding nodes to the latent space

Independently parametrize the distribution over the adjacency matrix of each view

$$\int p_\theta(\mathcal{G}, \mathcal{U})\,\mathrm{d}\mathcal{U} = \prod_{v=1}^{V} \int p_\theta(\mathbf{A}^v, \mathbf{U}_v)\,\mathrm{d}\mathbf{U}_v = \prod_{v=1}^{V} \int p_\theta(\mathbf{A}^v \,|\, \mathbf{U}_v)\, p(\mathbf{U}_v)\,\mathrm{d}\mathbf{U}_v,$$

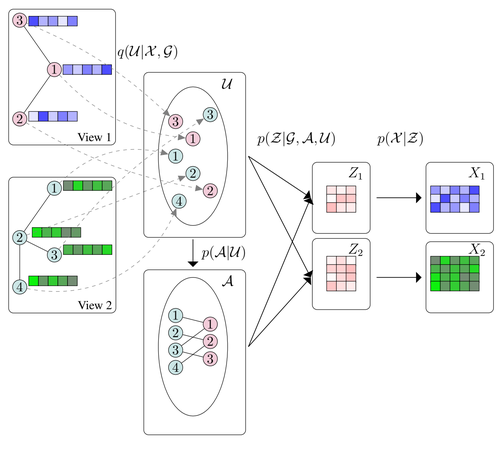
where the prior distribution of U_v's are standard Gaussian. We further approximate the distribution of $\mathcal{U}$ with a factorized posterior distribution

$$q(\mathcal{U} \,|\, \mathcal{X}, \mathcal{G}) = \prod_{v=1}^{V} q(\mathbf{U}_v \,|\, \mathbf{X}_v, G_v) = \prod_{v=1}^{V} \prod_{i=1}^{N_v} q(\mathbf{u}_{i,v} \,|\, \mathbf{X}_v, G_v),$$

$$q(\mathbf{u}_{i,v} \,|\, \mathbf{X}_v, G_v) = \mathcal{N}\left(\mathbf{u}_{i,v} \,|\, \boldsymbol{\mu}_{i,v}, \mathrm{diag}(\boldsymbol{\sigma}_{i,v}^2)\right),$$

$$\boldsymbol{\mu}_v = \varphi_v^{\mathrm{emb},\mu}(\mathbf{X}_v, G_v), \quad \log(\boldsymbol{\sigma}_v) = \varphi_v^{\mathrm{emb},\sigma}(\mathbf{X}_v, G_v)$$

ϕ_v^{emb}: variants of graph neural networks including GCN, GIN, GraphSAGE



Schematic illustration of BayReL

Constructing relational multi-partite graph

The distribution of bi-adjacency matrices are defined as follows

$$p(\mathbf{A}^{vv'} \,|\, \mathbf{U}_v, \mathbf{U}_{v'}) = \prod_{i=1}^{N_v} \prod_{j=1}^{N_{v'}} \mathrm{Bernoulli}\left(\mathbf{A}_{ij}^{vv'} \,|\, \varphi^{\mathrm{sim}}(\mathbf{u}_{i,v}, \mathbf{u}_{j,v'})\right),$$

where $\varphi_{\mathrm{sim}} = \sigma(\boldsymbol{u}_{i,v}\ \boldsymbol{u}_{j,v}^T)$ is a score function measuring the similarity between the latent representations of nodes.

Inferring view-specific latent variables

Parametrize the distributions for node attributes at each view independently

$$\int p_\theta(\mathcal{X}, \mathcal{Z} \,|\, \mathcal{G}, \mathcal{A}, \mathcal{U})\,\mathrm{d}\mathcal{Z} \;=\; \prod_{v=1}^{V} \prod_{i=1}^{N_v} \int p_\theta\left(\mathbf{z}_{i,v} \,|\, \mathcal{G}, \mathcal{A}, \mathcal{U}\right)\, p_\theta(\mathbf{x}_{i,v} \,|\, \mathbf{z}_{i,v})\,\mathrm{d}\mathbf{z}_{i,v}.$$

Prior construction

$$p_\theta(\mathcal{Z} \,|\, \mathcal{G}, \mathcal{A}, \mathcal{U}) = \prod_{v=1}^{V} \prod_{i=1}^{N_v} p_\theta(\mathbf{z}_{i,v} \,|\, \mathcal{G}, \mathcal{A}, \mathcal{U}), \qquad p_\theta(\mathbf{z}_{i,v} \,|\, \mathcal{G}, \mathcal{A}, \mathcal{U}) = \mathcal{N}(\boldsymbol{\mu}_{i,v}^{\mathrm{prior}}, \boldsymbol{\sigma}_{i,v}^{\mathrm{prior}}),$$

$$\boldsymbol{\mu}^{\mathrm{prior}} = \varphi^{\mathrm{prior},\mu}(\mathcal{A}, \mathcal{U}), \quad \boldsymbol{\sigma} = \varphi^{\mathrm{prior},\sigma}(\mathcal{A}, \mathcal{U})$$

Approximate the posterior

$$q(\mathcal{Z} \,|\, \mathcal{X}, \mathcal{G}) = \prod_{v=1}^{V} \prod_{i=1}^{N_v} q(\mathbf{z}_{i,v} \,|\, \mathbf{X}_v, G_v), \qquad q(\mathbf{z}_{i,v} \,|\, \mathbf{X}_v, G_v) = \mathcal{N}(\boldsymbol{\mu}_{i,v}^{\mathrm{post}}, \boldsymbol{\sigma}_{i,v}^{\mathrm{post}})$$

$$\boldsymbol{\mu}^{\mathrm{post}} = \{\varphi_v^{\mathrm{post},\mu}(\mathbf{X}_v, G_v)\}_{v=1}^{V}, \quad \boldsymbol{\sigma}^{\mathrm{post}} = \{\varphi_v^{\mathrm{post},\sigma}(\mathbf{X}_v, G_v)\}_{v=1}^{V}$$

Overall likelihood and learning

Marginal likelihood

$$p_\theta(\mathcal{X}, \mathcal{G}) = \int \prod_{v=1}^{V} p_\theta(\mathbf{X}_v \,|\, \mathbf{Z}_v)\, p_\theta(\mathbf{Z}_v \,|\, \mathcal{G}, \mathcal{A}, \mathcal{U})\, p(\mathcal{A} \,|\, \mathcal{U})\, p(\mathcal{G} \,|\, \mathcal{U})\, p(\mathcal{U})\,\mathrm{d}\mathbf{Z}_1 \ldots \mathrm{d}\mathbf{Z}_V\,\mathrm{d}\mathcal{A}\,\mathrm{d}\mathcal{U}.$$

Evidence Lower Bound (ELBO)

$$\mathcal{L} = \sum_{v=1}^{V} \Big[\mathbb{E}_{q_\phi(\mathbf{Z}_v \,|\, \mathcal{G}, \mathcal{X})} \log p_\theta(\mathbf{X}_v \,|\, \mathbf{Z}_v) + \mathbb{E}_{q_\phi(\mathbf{Z}_v, \mathcal{U} \,|\, \mathcal{G}, \mathcal{X})} \log p_\theta(\mathbf{Z}_v \,|\, \mathcal{G}, \mathcal{A}, \mathcal{U}) - \mathbb{E}_{q_\phi(\mathbf{Z}_v \,|\, \mathcal{G}, \mathcal{X})} q_\phi(\mathbf{Z}_v \,|\, \mathcal{G}, \mathcal{X})\Big] - \mathrm{KL}\left(q_\phi(\mathcal{U} \,|\, \mathcal{G}, \mathcal{X}) \,||\, p(\mathcal{U})\right),$$

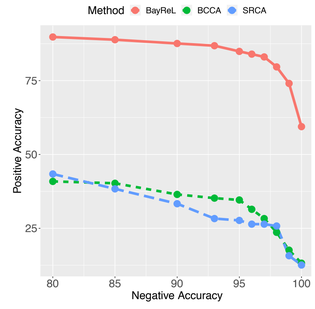
Experiments

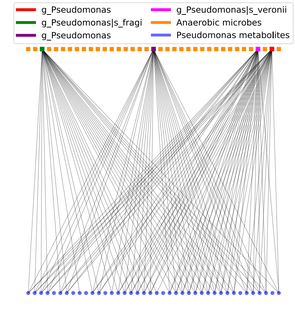
Microbiome-metabolome interactions in cystic fibrosis

Data description
- 172 patients with CF
- 138 unique microbial taxa
- 462 metabolite features

Graph density of input networks:
- Microbiome network: 0.102
- Metabolite network: 0.011

Evaluation metrics
- Positive accuracy: accuracy to identify the validated molecules interacting with *P. aeruginosa*
- Negative accuracy: accuracy of not detecting common targets between anaerobic microbes and notable pathogen





Left: Positive vs negative accuracy in CF dataset. **Right:** A sub-network of dependency graph consisting of *P. aeruginosa*.

miRNA-mRNA interactions in breast cancer

Data description
- 1156 patients with BRCA
- 11872 genes
- 432 miRNA

Networks:
- Gene regulatory networks (GRN): GENIE3 R package
- miRNA-miRNA: MISIM v2.0
- miRNA-mRNA validation: miRNet

Table 1: Prediction sensitivity (in %) in TCGA for different percentage of training samples.

Avg. degree	BCCA # of training samples 289 (25%)	BCCA # of training samples 578 (50%)	BCCA # of training samples 1156 (100%)	BayReL # of training samples 289 (25%)	BayReL # of training samples 578 (50%)	BayReL # of training samples 1156 (100%)
0.20	17.4 ± 0.8	17.6 ± 1.0	21.0 ± 0.0	31.9 ± 3.0	32.1 ± 1.0	34.0 ± 2.5
0.30	26.0 ± 0.8	26.4 ± 1.0	31.1 ± 0.7	45.8 ± 3.1	45.9 ± 1.5	47.4 ± 2.6
0.40	35.4 ± 0.8	35.5 ± 0.7	41.1 ± 0.2	57.6 ± 4.4	58.7 ± 1.3	59.5 ± 3.0

Precision medicine in acute myeloid leukemia

Data description
- 30 AML patients
- 53 drugs
- 9071 genes

Networks:
- GRN: 14 AML cell lines
- drug-drug: action mechanisms
- drug-gene validation: DGIdb

Evaluation metrics
- Prediction sensitivity of identifying reported drug-gene interactions

Table 2: Comparison of prediction sensitivity (in %) in AML dataset for different graph densities.

Avg. degree	0.10	0.15	0.20	0.25	0.30	0.40	0.50
SRCA	8.03	12.00	17.15	20.70	26.85	34.93	45.79
BCCA	9.65 ± 0.75	14.34 ± 0.06	18.96 ± 0.42	23.29 ± 0.52	28.22 ± 0.66	38.02 ± 2.15	46.88 ± 1.88
BayReL	$\mathbf{15.56} \pm 0.75$	$\mathbf{21.70} \pm 0.65$	$\mathbf{27.20} \pm 0.17$	$\mathbf{32.43} \pm 1.02$	$\mathbf{37.76} \pm 0.85$	$\mathbf{47.90} \pm 0.43$	$\mathbf{56.76} \pm 0.50$

- Consistency of significant gene-drug interactions in two different AML datasets with 30 patients and 14 cell lines: KL divergence between two inferred bipartite graphs are 0.38 and 0.66 for BayReL and BCCA, respectively.

References

Arto Klami et al. (2013). "Bayesian Canonical Correlation Analysis." In Journal of Machine Learning Research.

James T. Morton et al. (2019). "Learning representations of microbe–metabolite interactions." In Nature methods.

Su-In Lee et al. (2018). "A machine learning approach to integrate big data for precision medicine in acute myeloid leukemia." In Nature communications.