Dirichlet Process Mixture Model with Latent Matchings for Cross Species Expression Analysis

Hai-Son Le and Ziv Bar-Joseph

Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA

Overview

- Many applications need to match datapoints.
  - ▷ Orthologous genes (evolving from the same ancestry) across species.
  - ▷ Images and image captions.
  - ▷ Interactions between two groups of people.
- We would like to develop a model to:

1. Infer the matchings of datapoints.
2. Cluster datapoints into coherent groups.

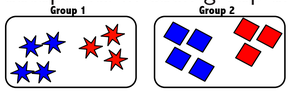
Results

- Non-parametric Bayesian model based on Dirichlet Process. The model can infer the posterior matching probability and the number of clusters.
- Inference algorithm: Variational Inference.
- Simulation and Real data results show good performance.

Methods

1. Some prior belief about potential matchings between two groups of datapoints = a sparse matrix.
2. Match datapoints of 2 groups and cluster them:

Clusters of datapoints in each group are coherent.

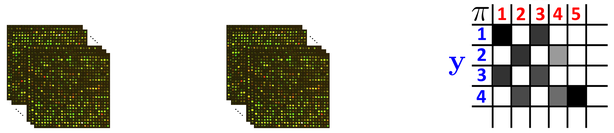


Good matchings → Clusters *across groups* are *also coherent*.

Joint likelihood of the matched datapoint clusters ∼ Quality of the matchings

Notations

Group 1 | Group 2 | A prior matching matrix π



$\mathbf{x} = [x_1, x_2, \ldots, x_{n_x}]$ $\mathbf{y} = [y_1, y_2, \ldots, y_{n_y}]$ *Indices are datapoints*

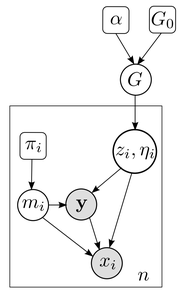
- π is a sparse matrix = the matching prior belief.

$$\pi_{i,j} = p(x_i \text{ matches to } y_j),\ \pi_{i,0} = p(x_i \text{ has no match in } \mathbf{y}),\ \sum_j \pi_{i,j} = 1$$

- **Assumption:** Matched datapoints are generated from a common distribution. Distributions of datapoints:
  - ▷ $F_Y(\eta)$: the marginal distribution of $Y = y_i$.
  - ▷ $F_{X|Y}(y, \eta)$: the conditional distribution of $X = x_i$ given $Y = y$.

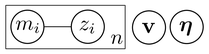
Graphical Model

- Latent matching variable: $m_i = \begin{cases} j, & x_i \text{ is matched to } y_j. \\ 0, & x_i \text{ has no match in } \mathbf{y}. \end{cases}$
- Mixture membership variable: z_i. Mixture parameters: η_i.



$$\begin{aligned} G &\sim \text{Dirichlet Process}(\alpha, G_0) \\ z_i, \eta_i \mid G &\sim G \\ m_i \mid \boldsymbol{\pi}_\mathbf{i} &\sim \text{Discrete}(\boldsymbol{\pi}_\mathbf{i}) \\ y_{m_i} \mid m_i, z_i, \eta_i &\sim F_Y(\eta_i), \text{if } m_i > 0 \\ x_i \mid m_i, z_i, \eta_i, \mathbf{y} &\sim \begin{cases} F_{X|Y}(y_{m_i}, \eta_i) & \text{if } m_i > 0 \\ F_X(\eta_i) & \text{otherwise} \end{cases} \end{aligned}$$

Inference

- Variational Inference using Stick-breaking variables $\mathbf{v}$: 

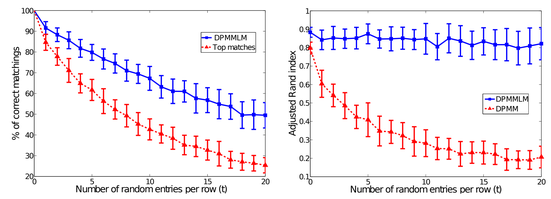
$$q(\mathbf{m}, \mathbf{z}, \mathbf{v}, \boldsymbol{\eta}) = \prod_{i=1}^{n_x} \{ q_{\phi_i}(m_i) \prod_{j=0}^{n_y} q_{\theta_{i,j}}(z_i)^{m_i^j} \} \prod_{k=1}^{K-1} q_{\gamma_k}(v_k) \prod_{k=1}^{K} q_{\lambda_k}(\eta_k)$$

- Extension to Variational Inference of DPMM by Blei and Jordan (2003).
- **Note:**
  - ▷ Need to compute the posterior mixture parameters given the conditional and marginal likelihood of datapoints.
  - ▷ *In our application:* Distributions of datapoints are Gaussians with unknown mean and covariance matrix.

  There is no closed form update rules for exact computation of mixture parameters → approximation.

Simulations

- Generate 120 datapoints ∼ Mixture of three 5-dimensional Gaussians (with high correlation between dimensions).
- We use the first 3 dimensions to create $x_{1,\ldots,120}$ and the last 2 dimensions to create $y_{1,\ldots,100}$.
- True π matrix is a *diagonal* matrix. Add $t \in [0, 20]$ random noise entries on each row of $\boldsymbol{\pi}$.
  - ▷ noise $\sim \frac{1}{2}\chi^2_1$. Repeat 40 times.



(a) Correct matchings. (b) Clustering evaluation.

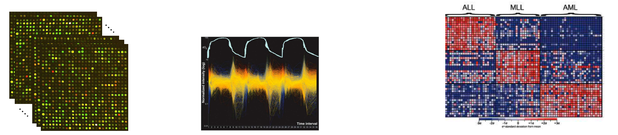
Baseline: DPMM on datapoint pairs inferred by top matches in π only.

Application

- Systematic discovery of orthology in many species from *(DNA) Sequences* and *Gene Expression data*.
- Discovery of novel groups of coherently changing genes across species.

DNA Sequence and Gene Expression data

- DNA Sequence is static.
- Gene Expression data measures gene dynamics.



Time series data Conditions: treatment, ...

- Sequences are well conserved. Many possible ortholog matches.
- *Find genes that are similar in sequence but different in expression.*
- *Genes that are similar in both sequence and expression, ...*

Immune Response data

Human Mouse

Salmonella infected Yersinia infected

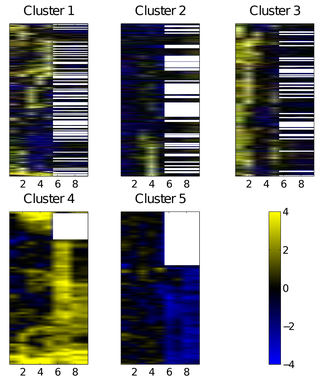


Figure: Clusters of genes.

x-axis: human(1-5)/mouse samples.
y-axis: human/matched mouse genes.

P value	Corrected P	GO term description
2.86216e-10	<0.001	regulation of apoptosis
4.97408e-10	<0.001	regulation of cell death
7.82427e-10	<0.001	protein binding
4.14320e-10	<0.001	regulation of programmed cell death
4.49332e-09	<0.001	positive regulation of cellular process
4.77653e-09	<0.001	positive regulation of biological process
8.27313e-09	<0.001	response to chemical stimulus
1.17013e-07	0.001	cytoplasm
1.28299e-07	0.001	response to stress
2.20104e-07	0.001	cell proliferation

Table: The GO enrichment result for cluster 1.

- Genes in cluster 1 are associated with immune and stress responses.
- Genes in cluster 3 are strongly upregulated in human cells while not changing in mouse: enriched for ribosomal proteins.
- Cluster 4 contains the most coherent set of upregulated genes across the two species.

Conclusion

Non-parametric Bayesian method based on DPMM to:

1. Infer the matchings of datapoints.
2. Cluster datapoints into coherent groups.

Systematic and rigorous approach to:

- Identify gene matchings.
- Infer groups of coherently-changing genes.

http://www.sb.cmu.edu/ {hple,zivbj}@cs.cmu.edu