Grade of Membership (GoM) models for counts data

Kushal K Dey

Overview

In biological applications, one often encounters counts data

	$feature_1$	$feature_2$	$\cdots$	$feature_G$
$samp_1$	c_{11}	c_{12}	$\cdots$	c_{1G}
$samp_2$	c_{12}	c_{22}	$\cdots$	c_{2G}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$samp_N$	c_{N1}	c_{N2}	$\cdots$	c_{NG}

Overview

In biological applications, one often encounters counts data

	$feature_1$	$feature_2$	$\cdots$	$feature_G$
$samp_1$	c_{11}	c_{12}	$\cdots$	c_{1G}
$samp_2$	c_{12}	c_{22}	$\cdots$	c_{2G}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$samp_N$	c_{N1}	c_{N2}	$\cdots$	c_{NG}

Aim: To cluster the samples based on the data across the features, usually $N < G$.

Overview

In biological applications, one often encounters counts data

	$feature_1$	$feature_2$	$\cdots$	$feature_G$
$samp_1$	c_{11}	c_{12}	$\cdots$	c_{1G}
$samp_2$	c_{12}	c_{22}	$\cdots$	c_{2G}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$samp_N$	c_{N1}	c_{N2}	$\cdots$	c_{NG}

Aim: To cluster the samples based on the data across the features, usually $N < G$.

We calculate grades of membership in different clusters (which we call topics) for each sample.

Overview

In biological applications, one often encounters counts data

	$feature_1$	$feature_2$	$\cdots$	$feature_G$
$samp_1$	c_{11}	c_{12}	$\cdots$	c_{1G}
$samp_2$	c_{12}	c_{22}	$\cdots$	c_{2G}
$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$samp_N$	c_{N1}	c_{N2}	$\cdots$	c_{NG}

Aim: To cluster the samples based on the data across the features, usually $N < G$.

We calculate grades of membership in different clusters (which we call topics) for each sample.

Model

We assume for sample n,

$$(c_{n1}, c_{n2}, \cdots, c_{nG}) \sim Mult\left(c_{n+}, p_{n1}, p_{n2}, \cdots, p_{nG}\right)$$

where $c_{n+} = \sum_{g=1}^{G} c_{ng}$.
Assuming number of clusters to be K, write p_{ng} as

$$p_{ng} = \sum_{k=1}^{K} \omega_{nk}\theta_{kg} \qquad \sum_{k=1}^{K} \omega_{nk} = 1 \qquad \sum_{g=1}^{G} \theta_{kg} = 1$$

Assume the priors

$$(\omega_{n1}, \omega_{n2}, \cdots, \omega_{nK}) \sim Dir\left(\frac{1}{K}, \frac{1}{K}, \cdots, \frac{1}{K}\right)$$

$$(\theta_{k1}, \theta_{k2}, \cdots, \theta_{kG}) \sim Dir\left(\frac{1}{KG}, \frac{1}{KG}, \cdots, \frac{1}{KG}\right)$$

Model Intuition

In the context of RNA-seq, for each sample n and feature g, we assume that some proportion of reads come from one of the K clusters or topics.

Model Intuition

In the context of RNA-seq, for each sample n and feature g, we assume that some proportion of reads come from one of the K clusters or topics.

The proportion of reads coming from k th cluster for sample n is given by ω_{nk}.

Model Intuition

In the context of RNA-seq, for each sample n and feature g, we assume that some proportion of reads come from one of the K clusters or topics.

The proportion of reads coming from k th cluster for sample n is given by ω_{nk}.

For cluster k, $\theta_{k.}$ is a G length vector of cluster probability distribution or relative expression pattern.

Model Intuition

In the context of RNA-seq, for each sample n and feature g, we assume that some proportion of reads come from one of the K clusters or topics.

The proportion of reads coming from k th cluster for sample n is given by ω_{nk}.

For cluster k, $\theta_{k.}$ is a G length vector of cluster probability distribution or relative expression pattern.

Model Intuition

We have shown for RNA-seq data across multiple tissues that GoM model performs better than the standard hierarchical models in separating the samples from different tissues.

Model Intuition

We have shown for RNA-seq data across multiple tissues that GoM model performs better than the standard hierarchical models in separating the samples from different tissues.

In cases, where there the topic proportions $\omega_{n.}$ vary continuously across samples, GoM models can pick such variations well, compared to other competing methods like PCA and t-SNE.

Model Intuition

We have shown for RNA-seq data across multiple tissues that GoM model performs better than the standard hierarchical models in separating the samples from different tissues.

In cases, where there the topic proportions $\omega_{n.}$ vary continuously across samples, GoM models can pick such variations well, compared to other competing methods like PCA and t-SNE.

The Structure plot provides a nice visualization of the clustering patterns across samples, and one can extract the top features driving each cluster.

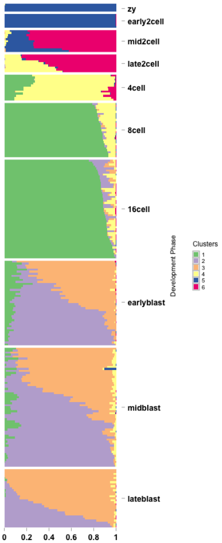
Model Intuition

We have shown for RNA-seq data across multiple tissues that GoM model performs better than the standard hierarchical models in separating the samples from different tissues.

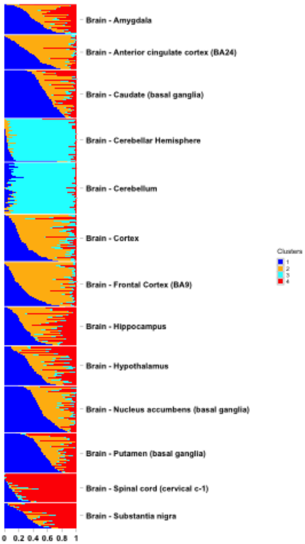
In cases, where there the topic proportions $\omega_{n.}$ vary continuously across samples, GoM models can pick such variations well, compared to other competing methods like PCA and t-SNE.

The Structure plot provides a nice visualization of the clustering patterns across samples, and one can extract the top features driving each cluster.

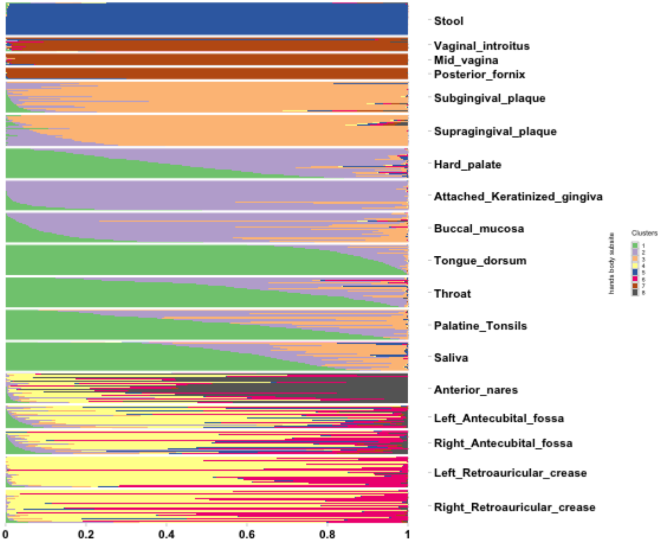
Example: Single cell development



GTEx v6 Brain clustering



Metagenomics HMP data



Possible routes

There is a hierarchy in the features in metagenomics data (OTU, species, genus). We may want to either incorporate this hierarchy in the model or perform a post-processing step to determine the effect of each level in driving clusters.

Possible routes

There is a hierarchy in the features in metagenomics data (OTU, species, genus). We may want to either incorporate this hierarchy in the model or perform a post-processing step to determine the effect of each level in driving clusters.

We may want to compare different clustering methods like PCA, t-SNE, diffusion maps with the GoM model and build a pipeline for metagenomics data that performs all these exploration and generates effective visualizations to interpret the data

Possible routes

There is a hierarchy in the features in metagenomics data (OTU, species, genus). We may want to either incorporate this hierarchy in the model or perform a post-processing step to determine the effect of each level in driving clusters.

We may want to compare different clustering methods like PCA, t-SNE, diffusion maps with the GoM model and build a pipeline for metagenomics data that performs all these exploration and generates effective visualizations to interpret the data

We may want to build networks over the OTUs and compare these network method results with the GoM output.

The End