Tissue tree specific gene biclustering

Avinash Das, Fabian Müller, Peter Ebert

August 24, 2013

1 Background

[**?**] gave a probabilistic generative model framework to find hierarchical organized biclusters in gene expression matrices (Gene x condition). The hierarchical biclustering method can be seen as a special instance of biclustering. The biclusters near the roots are coarse group of condition tied by a small subset of genes with homogeneous expression across the group wheres nodes deeper in the tree correspond to samller sets of conditions that exhibit homogeneous expression levels in a large subset of genes.

As other generative modeling, they start by presenting a generative process to observe a gene expression matrix. The generative process first samples the tree hierarchy of conditions (i.e. the condition are partitioned into tree structure), next it positions the genes into the sampled tree structure and finally it samples expression of genes for each condition (which are observed variables).

Sampling of tree structure: In order to partition the condition, they used the Chinese restaurant process (CRP). The root of the tree is initialized with all the conditions in the data and are partitioned using CRP. Each of these groups are recursively partitioned till maximum tree height is reached.

Length assignment to edges: Next, they used feature activation model to find active genes at each node, defined by latent variable $z_{j,u}$ for gene j and node u. The length l_{uv} of each edge (u, v) is sampled with $\sim \text{Beta}(\alpha, \beta)$. The $z_{j,u}$ is then sampled with probability of gene to be active equal to l_{uv} (i.e. $P(z_{j,u} = 1) = l_{uv}$). Further, once a gene is activated at a node, it remains active in all its descendants.

Generation of gene expression Given the gene states at each node of the tree, the gene expression Y is generated using:

$$\begin{aligned} Y_{ju}|z_{ju} = 1 &\sim N(\mu_{ju}, \sigma^2) \\ Y_{ju}|z_{ju} = 0 &\sim N(\mu_0 = 1, \sigma_0^2 = 1) \end{aligned} \tag{1}$$

The generative process is shown in the fig. **??**.

¡++¿ Next, they define joint distribution over the random variables and perform inferencing using the Gibbs sampling.

We make the following modification to the presented hierarchical biclustering method in order to adapt for the given genes $\times$ tissue expression matrix. Observations:

Figure 1: Illustration of the generative process for a fictitious noise-free data set with three genes (A, B, and C) and four samples (healthy 1, healthy 2, leukemia, and melanoma). The healthy samples share the same path assignment, while each of the cancer samples has its own unique path. The rounded rectangles represent nodes and indicate the current feature activation state. Gene A becomes active at both of the root's child nodes, leading to homogeneous expression for the healthy samples as well as for the cancer samples, although the between-group difference in expression is significant. Gene C exhibits homogeneous expression under both cancer samples, but not under the healthy samples. Gene B has a specific expression pattern for each of the samples (fig taken from [**?**]).

- We already have a tissue hierarchy.
- Our expression data are discrete values (no expression, weak expression, medium expression and high expression).
- In our tissue tree internal nodes have observation.
- We want, in addition to find gene clusters, to categorize genes based on its expression pattern across tissue.

2 Method

Taking cues from the generative method defined in [**?**], we can define a generative process in context of our problem. We will start with the tissue tree $\mathsf{T}(\mathsf{V},\mathsf{E})$ which can be constructed by assigning to each node, all leaf-tissues that are included by the subtree induced by that node. We will determine the activation states of genes at each tree node. Finally, given the state of gene we will sample its observed expression level

Latent variable Z: We define latent state variable $\mathsf{Z} = \{z_{ju} \in \{0,1,2\} : \forall j \in \mathsf{G}, u \in \mathsf{V}\}$ corresponding to each gene at each node with following definition:

$z_{ju} = 0 \implies$ *gene* j *is homogenously inactive in subtree with* u *as root.*
$z_{ju} = 1 \implies$ *gene* j *is homogenously active with high expression in subtree with* u *as root.*
$z_{ju} = 2 \implies$ *gene* j *have inhomogenously activity in subtree with* u *as root.*

In addition, once a node z_{ju} takes a value either 0 or 1 corresponding to states homogeneously active or in-active all its children node in subtree stays in same homogeneous state. Z can be then sampled from multinomial distribution with

a Dirichlet prior.

$$\begin{aligned} z &\sim \text{mult}(\phi) \\ \phi &\sim \text{Dir}(\alpha, \beta, \gamma) \end{aligned}$$

Generation of expression level: Lets assume we have binary data for expression level i.e. only 2 state expressed or not expressed. Given the state of the genes at the nodes of the tissue tree, we can generate samples of expression level using a binomial distribution.

$$\begin{aligned} Y|Z = i &\sim \text{Bin}(p_i) \\ p_i &\sim \text{Beta}(\alpha_i, \beta_i) \forall i \in \{1, 2, 3\} \end{aligned} \tag{2}$$

The beta priors can be chosen so to reflect our expected behaviour of expression level in different gene state. One example of prior that can be taken is shown in fig. **??**.

Figure 2: Beta priors for homogeneous active, homogeneous inactive and inhomogeneous gene state. The priors are consistent with the interaction matrix data; e.g. $E(p_2)$ = average that any gene is expressed in interaction matrix.

We can extend the method for the discrete values of expression level by using multinomial distribution instead of binomial distribution in order to represent the different expression stages (weak, intermediate, high and no expression).

2.1 Inference

TBD.

3 Expected outcomes

- Tissue specific gene active and inactive gene bicluster: We will get state of gene Z for each tissue subtree, that can be used to get information like gene clusters with tissue specific activity. We will also get information whether a gene cluster is active or inactive homogeneously at that node.
- We can quickly identify gene which are only expressed only at a node (or small subtree) and are not expressed in rest of the tree.
- We can quantify distance between different node of tissue tree in term of gene activity and expression level i.e we will get a distance between different tissues.

- We will obtain a noise resistant model because we are modelling the output observation as distribution.
- the method will be computationally less demanding as the CRP does not have to be sampled.