Chapter 2 Preview: Variable Selection

Kushal K Dey

December 3, 2015

Why Variable Selection?

The topic model due to Matt Taddy's **maptpx** package ran for almost 2 days on the 8555 GTEX samples with roughly 56000 genes. For the cellcyleR method on single cells as well, the computation time blows up as number of samples or number of genes considered increases (though in terms of run time, it is less expensive than admixture). The main goal going forward for us is to reduce the run time and make these methods faster, without compensating for the performance.

In most cases, we have a set of genes that are informative about clustering (topic model) or cell ordering (cellcycleR). For GTEX data for example, the genes mostly driving the clusters are the ones that vary in expression across tissues. Now, if we remove the noisy genes which are non-informative for clustering or in this case, do not differ much across tissues, then we can significantly reduced the computational time and also obtain higher predictive power.

Sometimes focussing on a particular biologically meaningful set of genes can show us additional patterns. For Yoav's data, mostly ribosomal genes and mitochondrial genes kept getting picked up when the analysis (topic model/ cellcyleR) was carried on all genes. When we focus just on the genes with supposedly cell phase specific expression patterns, then we seem to get more biologically meaningful ordering of the single cells based on cell phases.

In summary, the three main advantages of variable selection are

- Computational and time expenditure are less
- Removes noisy genes and leads to higher predictive power
- if variable selection is forced to be driven by some biologically interesting variable (like cell cycle), then it may throw up additional patterns in data.

What has been done

People have studied variable selection in clustering and regression models for ease of computation and better predicitve accuracy. Some of these approaches are pretty ad-hoc, others that involve model based approaches are not too feasible for big data like the GTEx data. The package **clustvarsel** implements the variable selection algorithm due to Dean and Raftery (2006). However the number of parameters that requires estimation is often quadratic in data dimensionality, this approach is pretty slow in high dimensions. A related approach is due to Maugis, Celeux, and Martin-Magniette (2009) who developed a package **selvarclust** and relaxed some assumptions on the role of variables from the method due to Dean and Raftery. Besides these, there are a few implicit approaches such as mixtures of factor analyzers (Ghahramani and Hinton 1997, Montanari and Viroli 2010; Viroli 2010, Andrews and McNicholas 2011a,b) and for latent class models (Brendan Murphy).

My feeling was that most of these methods were decorated and would be time expensive in high dimensions. One of the major motivations for us to do this selection is to reduce time consumption as this is essentially a pre-processing step to topic model or cellcyleR and these methods (I checked **clustvarsel**, **selvarclust** papers, Murphy, Viroli 2010 and Andrews and McNicholas 2011a) do not quite tick this box.

Types of variable selection

In most biological experiments, we are given sample metadata information and each metadata group may be thought of as a small class/cluster that is known. Biological clusters often correspond to these metadata already provided which also helps in validating the clusters. Keeping this in mind, we may have two scenarios to consider for variable selection.

- **metadata-driven approach**
- **fully unsupervised scenario**

An example of **metadata-driven approach** would be genes for GTEx analysis, where we know that they would be pretty much homogeneous within each tissue type and would vary between tissue types. A **fully unsupervised scenario** would be a scenario where we do not have metadata for the samples.

Thoughts about models

I will present sketches of my ideas about how to go about this issue. The models can definitely be improved and some other model may also get suggested.

metadata driven approach

For *metadata-driven* scenario, I think we can use the relevant metadata to extract a set of meaningful genes. This also implies the clustering then would be driven by the metadata. For example, for GTEX data, we can use the metadata given by **tissue/subtissue labels**.

To illustrate, let Y_{ng} be some normalized expression for sample n and gene g, obtained from the reads in the GTEx data.

$$H_1 : Y_{ng} = \mu_g + \beta_{t(n):g} + \epsilon_{ng}, \qquad \epsilon_{ng} \sim N(0, \sigma_g^2)$$

$$H_0 : Y_{ng} = \mu_g + \epsilon_{ng}, \qquad \epsilon_{ng} \sim N(0, \sigma_g^2)$$

where $t(n)$ represents the tissue level information of the sample n. While discussing on STAN implementation of **ash**, we did discuss estimating this model together with adaptive shrinkage on β values. One can also consider the generalized version of this using the Poisson model applied directly on the read counts data, without transforming the data first.

An important assumption for the topic model and the cellcycleR model is that the genes are all independent. We can compute some score for each gene, for example the LRT test statistic $-2log(L_1/L_0)$. This statistic under null should have a χ^2 distribution, but we can look for an appropriate variance stabilizing transformation to make the resulting statistic behave as asymptotically normal.

Then we have G values of these LRT statistics, say LR_g for g th gene. We may want to apply **ash** type shrinkage to these LRT values or some transformation of the LRT values that can be assumed to be normal. We may also apply **ash** type shrinkage to the β values and compare the posterior distribution of H_1 and H_0 instead of taking the likelihood route.

fully unsupervised scenario

When we do not have any metadata, then it becomes pretty hard to do the variable screening. The above method does not work as we have no metadata or we do not want to force the metadata available on the clusters. For this problem, we can adaptively learn suitable genes that can drive the clustering. The algorithm I am thinking of is as follows.

- We partition the genes into say B blocks.
- Run a topic model on the original data or thinned version of the original data for each of these B blocks parallely (thinned data will run faster). Choose your K to be same you want to apply on the full data or just a slightly larger K to account for more variation. For cellcycleR, do the same blockwise application without the thinning part and without choosing K.
- From the run in block b, where b varies from 1 to B, we get some expression profiles for topic model, and the SNR values for all the genes used in that block.
- For topic model, weight the genes based on KL divergence score that we had been using for cluster annotation. For cellcyleR, use the SNR values instead (higher SNR leads to higher weights).
- Run this method several times (say R times) and record the weights $w_{g,r}$ for the gene g after each run r.
- For each gene, calculate
$$w_g = \frac{1}{R}\sum_{r=1}^{R} w_{g,r}$$
- We remove genes with weights w_g below a threshold, suitably chosen, and then run the actual topic model or cellcyleR on the remaining genes.

This approach of partitioning the data and assigning weights on variables based on clustering results is analogous to adaptive boosting approach for classification.

Note that the fully unsupervised scenario is supposed to give similar results to the topic model on the full data without variable selection but the metadata-driven approach may not. So point (3) in advantages of variable selection is only applicable for metadata-driven approach.

For example, for Yoav data, if we use cell cycle phases as the metadata, then we expect to mainly pick up genes that have cell cycle phase related information and then the clustering will be driven by cell cycle phases. However, the original topic model on all genes will cluster based on the ribosomal protein related genes etc, so will be pretty different from the metadata-driven approach.