Voom topic models

Kushal K Dey

February 18, 2016

Introduction

The central concept of `voom` sis to transform the RNA-seq read counts data such that the transformed data follows an approximately normal distribution, with the mean and the variance structure appropriately taken account of to reflect the central tendency and the variation structure at the counts level of the data.

The idea in this script is to use a normal approximation version of the topic model. The topic model fits a multinomial distribution to the RNA-seq read counts for each sample separately, conditional on the library size. This is similar to assuming a Poisson model for the counts for the unconditional data. The idea then is to convert from the Poisson model to a suitable normal approximation model using the concepts similar to `voom`.

Aim

A major reasoning behind performing a voom- based topic model are

- **Overdispersion**: The Poisson model assumes that the mean and the variance are same. This is in general not true, although it is true that higher mean of counts often leads to higher dispersion as well.
- **Sparsity**: We would like to impose sparsity assumptions on the θ matrix of the topic model, and may be sometimes on both ω and θ and compare the results under different sparsity assumptions. Sparsity is easier to implement in normal models compared to the Poisson models.

Voom Transformation

We start with the counts c_{ng} for sample n and gene g. Then, we can write

$$(c_{n1}, c_{n2}, \cdots, c_{nG}) \sim Mult(c_{n+}, p_{n1}, p_{n2}, \cdots, p_{nG}) \tag{1}$$

where

$$p_{ng} = \sum_{k=1}^{K} \omega_{nk}\theta_{kg} \qquad \sum_{k=1}^{K} \omega_{nk} = 1 \qquad \forall n \qquad \sum_{g=1}^{G} \theta_{kg} = 1 \qquad \forall g \tag{2}$$

We define the log-cpm as follows

$$y_{ng} = log_2\left(\frac{c_{ng} + 0.5}{c_{n+} + 1} \times 10^6\right) \tag{3}$$

Let us define

$$\lambda_{ng} := E(c_{ng}) = c_{n+} \sum_{k=1}^{K} \omega_{nk}\theta_{kg} \tag{4}$$

$$var(c_{ng}) := \lambda_{ng} + \phi_{ng}\lambda_{ng}^2 \tag{5}$$

where ϕ_{ng} is the overdispersion parameter.

We can show that when λ_{ng} is large.

$$var(y_{ng}) = \frac{1}{\lambda_{ng}} + \phi_{ng} \tag{6}$$

using Taylor series expansion.

Topic model preprocessing

As a preprocessing step, we first run a topic model on the c_{ng} and obtain a rough estimate of $\omega_{nk}^{(0)}$ and $\theta_{kg}^{(0)}$. Then we can estimate

$$\lambda_{ng}^{(0)} = c_{n+} \sum_{k=1}^{K} \omega_{nk}^{(0)} \theta_{kg}^{(0)} \tag{7}$$

Let us define

$$\mu_{ng} := E(y_{ng}) \tag{8}$$

We can write

$$\mu_{ng} = \lambda_{ng} - log_2(c_{n+} + 1) + 6log_2(10) \tag{9}$$

Next we define

$$c_g := \bar{y}_g + log_2(GM(c_{n+})) - 6log_2(10) \tag{10}$$

where $GM(.)$ represents the geometric mean.

The fitted value from the preprocessing step

$$\mu_{ng}^{(0)} = \lambda_{ng}^{(0)} - log_2(c_{n+} + 1) + 6log_2(10) \tag{11}$$

We define $e_{ng} = y_{ng} - \mu_{ng}^{(0)}$, and then compute the standard deviation for the gth gene across all samples s_g. To obtain a smooth mean-variance trend, a LOESS curve is fitted to square root of standard deviations as a function of mean log counts. We define by $lo(\lambda_{ng})$ the predicted value of square root standard deviation of y_{ng}.

The voom precision weights are the inverse variances $w_{ng} = (lo(\lambda_{ng}))^{-4}$.

Factor model

We input the y_{ng} and and the precision weights w_{ng} into a voom normal topic model scenario. We now fit the model

$$y_{ng} = \mu_{ng} + e_{ng} \tag{12}$$

$$= c_{n+} \sum_{k=1} \omega_{nk}\theta_{kg} - log_2(c_{n+} + 1) + 6log_2(10) + e_{ng} \tag{13}$$

where $e_{ng} \sim N(0, w_{ng}^2)$ where w is estimated as above.

But this model may not be easy to fit given the constraints on ω and θ. So, instead we fit a factor analysis model

$$y_{ng} + log_2(c_{n+} + 1) = \sum_{k=1}^{K} \lambda_{nk} f_{kg} + e_{ng} \tag{14}$$

where λ and f comprise of all non-negative entries. Next we can impose sparsity constraints on λ and f as done in `pmd` or in `flash`.

The other way of handling it would be to fit

$$y_{ng} = \sum_{k=1}^{K} \lambda_{nk} f_{kg} + e_{ng} \tag{15}$$

where we impose no constarints on λ and f. But the previous model seems more meaningful to me.

Challenges

The main challenge is to incorporate the fact that the precision weights are different across samples and across genes and athey are known. Right now, PMD can handle non-negative λ and f and also can impose various sparsity constraints (non-adaptive). But it does not take into account the fixed precision weights varying across samples and genes as here.

As for applying `flash`, we need both the non-negativity constraint and the unequal precision weights flexibility.