Likelihood ratio tests for differential expression across pseudotime

Kieran Campbell

kieran.campbell@sjc.ox.ac.uk

January 7, 2016

1 Introduction

Many experiments involve cells progressing through a biological process such as differentiation or apoptosis (cell death). Using RNA sequencing the gene expression in individual cells can be measured. As a result the expression profile of each cell represents a distinct time point through the process, known as the *pseudotime*, and algorithms have been developed to order cells and assign pseudotimes to them.

Once cells have been ordered in pseudotime it is useful to find which genes are *differentially expressed* across pseudotime (in other words, what genes change more than expected at random across pseudotime?). A simple way of doing this involves model selection: we construct a null model that relates to no differential expression and an alternative model that relates expression to some function of pseudotime, then pick the model that best explains the differential expression. If this model is the alternative model, we designate the gene as differentially expressed.

1.1 Model

Let y_{ij} denote the $\log_2$ gene expression of gene i in cell j at pseudotime t_j then

$$y_{ij}(t_j) \sim \mathcal{N}(\mu_i(t_j),\, \sigma_i^2) \tag{1}$$

where

$$\mu_i(t_j) = \begin{cases} \mu_i^{(0)}, & \text{if gene } i \text{ not differentially expressed,} \\ \frac{2\mu_i^{(0)}}{1+\exp\left(-k_i(t_j - t_i^{(0)})\right)}, & \text{if gene } i \text{ differentially expressed.} \end{cases} \tag{2}$$

2 Assignment

1. Why is using a likelihood ratio test suitable for comparing these models?

2. How many degrees of freedom does the null model have? How many does the alternative model have?

Interlude Recall that for a likelihood ratio test the statistic

$$D = -2\ln(\text{likelihood for null model}) + 2\ln(\text{likelihood for alternative model})$$

follows a χ^2 distribution with $\text{df}_a - \text{df}_n$ degrees of freedom (where df_a and df_n are the degrees of freedom for the alternative and null models respectively).

The rest of this assignment will follow in `R`. It is recommended you work through each command listed here to set yourself up for the assignment questions. If you would prefer to continue in Python you'll have to be comfortable with numerical optimisation of the log-likelihood of the above model.

First install the `R` package `modelselectionworkshop`:

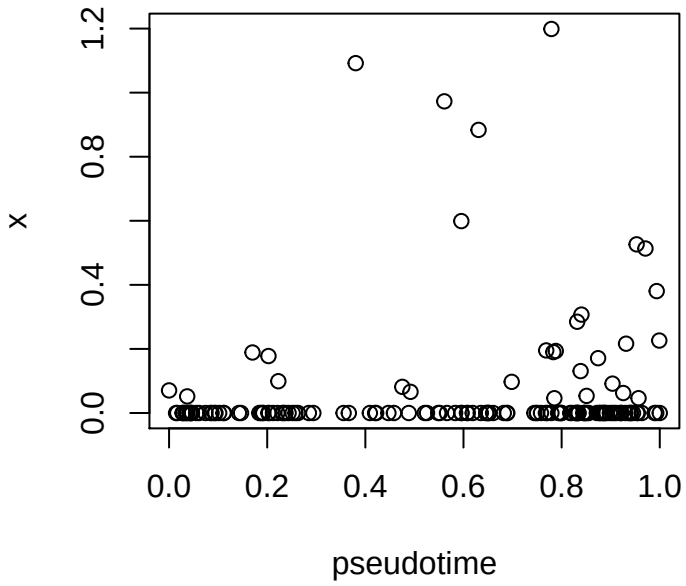
```
install.packages("devtools")
devtools::install_github("kieranrcampbell/modelselectionworkshop")
```

This includes a 500 by 155 matrix of gene expression measurements (for 500 genes and 155 cells) `X` and a 155 element vector `pseudotime` of pseudotimes for each cell. In `r` we can load data associated with a package by calling `data` and view the structure of any items using `str`:

```
library(modelselectionworkshop) # load package
data(X)
data(pseudotime)
str(X)

##  num [1:500, 1:155] 0 0.55 0 0.572 0.785 ...
##  - attr(*, "dimnames")=List of 2
##   ..$ : chr [1:500] "ENSG00000233750.3" "ENSG00000176595.3" "ENSG00000102098.13" "ENSG00000124222.16" .
##   ..$ : chr [1:155] "T0_CT_A01" "T0_CT_A03" "T0_CT_A05" "T0_CT_A06" ...

str(pseudotime)

##  num [1:155] 0.0742 0.1119 0.17 0.034 0.0168 ...
```

To get a feel for the data let's look at the first gene. In `R` we index matrices using square brackets, with the first value corresponding to subsetting on the row and second on the column. The `plot` command comes in handy for quickly looking at data:

```
x <- X[1,]
plot(pseudotime, x)
```



The functions `fit_null_model` and `fit_alt_model` provided by the package fit the two models and provide the log likelihoods at the maximum likelihood estimates. These return `list` objects in `R`, with two entries: `par` corresponding to the parameter vector and `loglik` corresponding to the log-likelihood. (Hint: in `R` an element named `e` in the list `L` can be accessed using the dollar sign like `L$e`.)

```
null_model <- fit_null_model(x)
alt_model <- fit_alt_model(x, pseudotime)
print(null_model)

## $par
## [1] 0.05945317 0.03512770
##
## $loglik
## [1] -40.09383

print(alt_model)

## $par
##            L            k           t0       sig_sq
## 8.138547e-02 1.898407e+03 3.697681e-01 3.390793e-02
##
## $loglik
## [1] -42.33715
```

3. What is a p-value? What *isn't* a p-value?

4. Find a p-value for the first gene being differentially expressed. (Hint: `pchisq` will come in handy.)

5. Find p-values for all genes in the dataset. (Hints: `apply` can be used to apply a function over a matrix. `myfunc <- function(arg1, arg2)` defines a function in `R`).

6. What subsequent analysis needs done before we can decide if a gene is differentially expressed? (Hint: `p.adjust` will come in handy.)

7. What are the other weaknesses of this method?