Modeling gene expression temporal variation

Guilherme de Sena Brandine

September 29, 2016

1 Introduction

Define $T = \{t_1, \dots t_n\} \subset [0, 1]$ a set of time points in which we have samples of the gene values in that time, that is, for each time point t_j we have a corresponding set $X_j = \{x_{j1} \dots x_{jn}\}$ of observations for this particular time. The goal is to find a function $\mu(t)$ that best characterizes the temporal evolution of the gene values.

2 Modeling the behavior of μ and σ

We define $\mu(t)$ as a family of functions that simulate how genes can biologically behave throughout time. We define three types of functions:

- Oscillatory genes (eg, cell cycle genes), given by: $f_o(t) = \alpha_o \sin(\beta_o t + \delta_o)$ for some parameters $\alpha_o, \beta_o, \delta_o$.
- Continuously increasing/decreasing genes, herein given by: $f_{id}(t) = \alpha_{id} x^{\beta_{id}} + \delta_{id}$ (note that this function can be constant if $\alpha_{id} = 0$)
- Swich genes modelled by heaveside functions: $f_s(t) = \mathbb{I}_{t>\alpha_s}$, which can also be used to model transitions between oscillation/increase-decrease behaviors:

We model $\mu(t)$ as a combination of these parameters:

$$\mu(t) = f_o(t) f_s(t) + f_{id}(t)(1 - f_s(t))$$

We can model the samples as either a normal distribution (for the case where normalization gives continuous values, such as RPKM or DESeq), or a Negative Binomial (for the discrete case):

$$x_{ji} \sim \begin{cases} \mathcal{N}(\mu(t_j), \sigma^2(t_j)) \text{ if x is continuous} \\ NB(\mu(t_j), \sigma^2(t_j)) \text{ if x is discrete} \end{cases}$$

Where $\sigma^2(t) = \mu(t) + \epsilon\mu^2(t)$ is the well-characterized overdispersion that represents technical and biological noise in gene expression reads.

The values of $\hat{\mu}(t_j)$ can thus be estimated by the UMVUEs of the Normal/NB distributions (both are the same):

$$\hat{\mu}(t_j) = \frac{\sum_{x \in X_j} x}{n}$$

$$\hat{\sigma^2}(t_j) = \frac{\sum_{x \in X_j} (x - \hat{\mu}(t))^2}{n-1}$$

3 Fitting the gene function

To predict the behavior of $\mu(t)$ all we need to do is find the parameters $S = \{\alpha_o, \beta_o, \delta_o, \alpha_{id}, \beta_{id}, \delta_{id}, \alpha_s\}$ that minimize the squared error :

$$E(S) = \sum_{i=1}^{n} (\hat{\mu}(t_j) - \mu(t_j))^2 + (\hat{\sigma^2}(t_j) - \sigma^2(t_j))^2$$

Which is a simple convex function optimization problem.