

1 Background and Objectives

We assume there are S single cells in our dataset. We do not know which cell cycle phase they come from and have only recorded the gene expression of these cells. Not all of these genes bear cell cycle specific information, so we constrain our focus to only the cell cycle genes. Before we try to model the data we have, let's have a glimpse at what we have and what we want to achieve.

- Objective: We have some cells, we do not know which phase they are in. We want to find out which phase they are in and the relative order
- We know a set of genes are expressed in these phases. Is that the full set is not known. Lengths of the phases - not known. Also very likely not all of the genes in the list are actually expressive or oscillatory. Should we do some preprocessing to filter them (aka Macosko approach of correlation thresholding). May be look at qtlCurves plot to get a sense.
- Researchers assume sinusoidal wave function for modeling the oscillatory genes. Is that a good model? What does qtlCurves say? Can we do Fourier or wavelet fitting instead (that's just because I am learning these :)).
- We want a hierarchical ordering. First we want to group the single cells into the higher level order of cell cycle phases, namely G1.S, S, G2.M, M and M.G1. But then once we have done that, we also want to reorder the cells inside each of these phases. That is a little extra than just finding the ordering of the cells, because that does not tell us which phase it is from and does not fully draw from phase specific cell cycle genes.

2 Model

Let the vector of time orders for the S cells is given by t_S and for gene g and cell s , we can write down the model

$$Y_{sg} = \alpha_g \sin(t_s + \phi_g) + \epsilon_{sg} \quad \epsilon_{sg} \sim N(0, \sigma_g^2)$$

Note that the frequency is 1 because the reference frame is the cell cycle which is a circle and the period is thus assumed to be 2π . This model is assumed for all g which are sinusoidal. Without loss of generality, we assume that all the genes are such. So, we can write down the model for cell s as

$$\mathcal{L}_s \propto \prod_{g=1}^G N(\alpha_g \sin(t_s + \phi_g), \sigma_g^2)$$

The full model over all the S cells is given by

$$\mathcal{L} \propto \prod_{s=1}^S \prod_{g=1}^G N(\alpha_g \sin(t_s + \phi_g), \sigma_g^2)$$

Here t_s are the cell specific parameters and α_g , ϕ_g and σ_g^2 are gene specific parameters. I do not think this model is identifiable.

Given the vector t_s , we can write

$$\begin{aligned} Y_{sg} &= \alpha_g \cos(\phi_g) \sin(t_s) + \alpha_g \sin(\phi_g) \cos(t_s) + \epsilon_{sg} \\ &= \beta_{1g} \sin(t_s) + \beta_{2g} \cos(t_s) + \epsilon_{sg} \end{aligned}$$

There is a bijective mapping from (α, ϕ) to (β_1, β_2) and so it is enough to find the ML estimates of β_1 and β_2 instead of α and ϕ . The bijective map is given by

$$\alpha_g = \sqrt{\beta_{1g}^2 + \beta_{2g}^2}$$

$$\phi_g = \tan^{-1} \left(\frac{\beta_{2g}}{\beta_{1g}} \right)$$

We can thus write down the model in matrix notation as

$$X_{S \times 1}^g = M_{S \times 2}^s \beta_{2 \times 1}^g + \epsilon_{S \times 1}^g$$

Assume flat prior for β^g

$$P(\beta^g) \propto 1$$

Then the posterior for β^g is given by

$$P(\beta^g | \sigma_g^2, t_S, X) \propto N(\beta^g | \hat{\beta}^g, \sigma^2 (M^T M)^{-1})$$

We assume the prior for σ_g^2 to be

$$P(\sigma_g^2) \propto \frac{1}{\sigma_g^2}$$

Then the posterior is given by

$$\begin{aligned} P(\sigma_g^2 | \beta^g, t_S, X) &\propto \frac{1}{\sigma_g^2} N(X_g | M \beta_g, \sigma_g^2 I) \\ &\propto \text{InvGamma} \left(\sigma_g^2 | \frac{S}{2}, \frac{1}{2} (X^g - M^s \beta^g)^T (X^g - M^s \beta^g) \right) \end{aligned}$$

Now the most important part is finding the posterior

$$P(t_S | \sigma_g^2, \beta^g, X)$$

. One suggestion was to use t_s for each s as a discrete variable that takes values G1.S, S, G2.M, M and M.G1. However then how should we fix the relative order within each group? The hierarchy of the ordering is important as mentioned in previous section. So, just finding a posterior on t_S which is a S dimensional vector and pretty big may not be the best idea as we are missing out on cell phase information.

Even if we want to find the latter posterior, then also we may have to resort to MCMC.