MASSACHVSETTS INSTITVTE OF TECHNOLOGY

Problem Set 3
Due March 8

In this problem set, we will review a simple maximum likelihood method to estimate per-chip scaling factors for normalization of expression array data, using spiked-in controls with known quantities. There is only one problem, but we expect an answer for each sentence or clause that begins with the words, "Write down..." Please include the full derivations of all work formulae written, not just the final answer.

Problem 1

Each chip in our expression array dataset contains M spiked-in control spots (these spots will be the same across all chips). We will index these different control spots with the variable i. We also denote the total number of chips by N, each indexed with the variable j.

We begin by assuming that the measured value x_{ij} at each control-spot is the combination of a "true" level of control m_i, with two additional sources of multiplicative error. The term r_j indicates a per-chip source of multiplicative error, while e_{ij} is per-spot multiplicative error.

$$x_{ij} = m_i \times r_j \times e_{ij} \tag{1}$$

In this problem, we will work in log-scale. Therefore, we will use $y_{ij} = \log x_{ij}$, $\mu_i = \log m_i$, $\rho_j = \log r_j$, and $\epsilon_{ij} = \log e_{ij}$. Equation (1) now becomes:

$$y_{ij} = \mu_i + \rho_j + \epsilon_{ij} \tag{2}$$

We assume that μ_i and ρ_j are fixed for each i and j, but that ϵ_{ij} is distributed normally, with zero mean and variance σ_i^2 that depends only on the identity of the control (and not on the chip). Therefore, our y_{ij} are normally distributed,

$$y_{ij} \sim \mathcal{N}(\mu_i + \rho_j, \sigma_i^2) \tag{3}$$

Finally, we make the assumption that the y_{ij} are independent, given the values of their parameters μ_i, ρ_j, and σ_i^2. With this assumption in mind, write down the value for the total log-likelihood, $\mathcal{L} = \log P(Y)$.

The log-likelihood $\mathcal{L}$ is a function of its parameters, μ_i, ρ_j, and σ_i^2. We could therefore talk about the "optimal" parameters, i.e., those parameters that maximize its value.

$$\langle \mu_i^*, \rho_j^*, (\sigma_i^*)^2 \rangle = \arg \max_{\mu_i, \rho_j, \sigma_i^2} \mathcal{L} \tag{4}$$

Choosing the values of these parameters that maximize $\mathcal{L}$ is known as the "maximum likelihood" setting of the parameters. But how to find this maximum? Write down the expressions for $\frac{\partial \mathcal{L}}{\partial \mu_i}$, $\frac{\partial \mathcal{L}}{\partial \rho_i}$, and $\frac{\partial \mathcal{L}}{\partial \sigma_i}$. Set them equal to 0, and use those equations to solve for the stationary values of each parameter. Write down those equations.

Write down the answers to the following questions:

- What do you notice about these equations? Are they "easily" soluble (do they depend only on values that are already known, or given in the data)?
- Can you suggest a method of using these equations to find optimal parameter settings? (You don't need to implement this or even prove its correctness, simply suggest a "reasonable" method for solving the stationary-value equations).