COM SCI C121 Week 10

Aidan Jan

June 4, 2024

How do we develop a model to analyze data from these experiments?

- "What targets (perturbations) affect the gene of interest (measurement)?"

$$y_{ng} = \beta_0 + \beta_g x_g$$

- where:
  - n_{ng} = reporter of protein expression
  - β_0 = mean expression under "normal" conditions
  - β_g = effect of perturbation
  - x_g = indicator if perturbation occurs in this observation
  - n = replicate
  - g = target (gene) being perturbed

We have unique data - let's think carefully

Some goals of our method:

1. Identify gene perturbations that change the reporter (gene of interest) distribution
   - multiple guides pere gene should show the same trend
2. Model the sampling distribution in its "native" state
3. Be well behaved in small sample sizes
4. Infer experimental specific parameters
   - Bin size, guide specific variance, etc.

Multiple guides target the *same* gene and shus should be correlated

Goal 1: Identify gene perturbations that change the reporter (gene of interest) distribution. Multiple guides should show similar effects.

- Suppose a gene we want to knock out is 20 base pairs long.
- When we knock it out, maybe the gene we knocked out doesn't matter much.
- However, maybe a gene that *actually* matters contains the same 20 base pairs (highly likely considering the entire genome is billions of base pairs long)
- This causes the data to show that knocking out that gene has a huge effect, even if it does nothing.

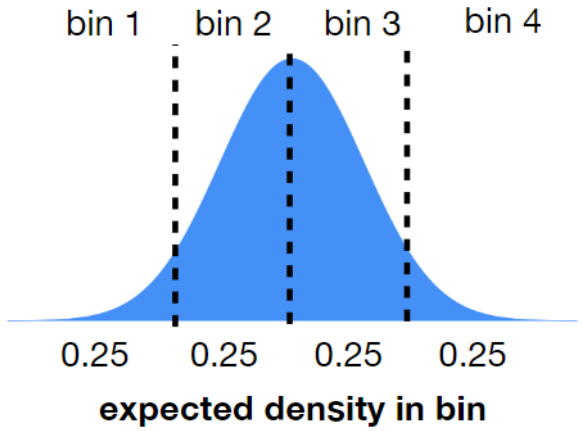
Approach: Guides have their own effect, but share a "parent" effect

$$\begin{aligned} \mu_T &\sim F(.) \\ \beta_g|\mu_T &\sim H(\mu_T) y_{ng} \qquad\qquad = \beta_0 + \beta_g x_g \end{aligned}$$

The equations give:

$$\begin{aligned} \mu_T|\sigma^2 &\sim N(0, \sigma^2) \\ \beta_g|\mu_T &\sim N(\mu_T, ?) \\ y_{ng}|\beta_g &\sim N(\beta_g x_g, 1) \end{aligned}$$

From these equations, we solve goal 2. We don't directly observe the reporter, we observe a noisy, quantized version.
Goal 2: Model the sampling distribution in its "native" state.



What if the expected counts per bin wasn't even? Then, the distribution is probably a Multinomial(C, p).

Dirichlet Multinomial

Our sampling distribution is an over dispersal Multinomial, aka *Dirichlet Multinomial*.

$$\text{DirMult}(c, \phi p)$$

$$B_{ng}|\phi, p \sim \text{DirMult}(c, \phi p)$$

We need to connect y_{ng} to our sampling distribution, so

$$B_{ng}|\phi, p \sim \text{DirMult}(c, \phi p(y_{ng}))$$

- B_{ng} is the data (observed bin counts)

In the previous equations:

- $\mu_T|\sigma^2 \sim N(0, \sigma^2)$ = Gene level effect
- $\beta_g|\mu_T \sim N(\mu_T, ?)$ = Guide level effect. (There's a ? because we don't know the exact guide level effect)
- $y_{ng}|\beta_g \sim N(\beta_g x_g, 1)$ = Unobserved reporter
- $B_{ng}|\phi, p \sim \text{DirMult}(c, \phi p(\beta_g))$ = Observed bin counts
- This is the perspective from the generator.

Goal 3: Behave well in small sample sizes (n = 3!?)

- One way to deal with this is to *shrink* parameters close to a *shared* value
- Additionally, we can enforce some sparsity

Spike-and-Slab Prior

The spike-and-slab prior enforces sparsity at the gene-level.
Does this *gene* have an effect? $\psi_T|\pi \sim \text{Bernoulli}(\pi)$
Yes, it does. $\mu_T|\psi_T = 1 \sim N(0, \sigma^2)$. But at the same time, no, it doesn't. $\mu_T|\psi_T = 0 \equiv 0$.

What is the ?, effect of guides?

$$\beta_g|\mu_T \sim N(\mu_g, ?)$$

$$B_{ng}|\phi, p \sim \text{DirMult}(c, \phi p(\beta_g))$$

Intuitively, we want the guides to be somewhat "similar"

- Guides should be similar
- If we had many guides per gene, we could have a different variance for each gene's guides, $\beta_g|\mu_T \sim N(\mu_g, \tau_T^2)$
  - The number of guides is usually 3-5.
- Enter shrinkage: $\beta_g|\mu_T \sim N(\mu_g, \tau^2)$