STA613/CBB540: Statistical methods in computational biology

Generalized Linear Models (1/29/13)

Lecturer: Barbara Engelhardt *Scribe: Yangxiaolu Cao*

When processing discrete data, two commonly used probability distributions are the binomial distribution and the Poisson distribution. The binomial distribution is used when an event only has two possible outcomes (success, failure); the Poisson distribution describes the count of the number of random events within a fixed interval of time or space with a known average rate. Generalized linear models are a generalization of the Gaussian linear model, in that the conditional distribution of the response variable is any distribution in the exponential family. Logistic regression is one GLM with a binomial distributed response variable. We will look at Poisson regression today.

1 Poisson Regression

Let $D = \{(x_1, y_1), \ldots, (x_n, y_n)\}$ be a set of paired data, where y_i is a scalar and x_i is a vector of length p. Let the parameter θ be a vector of length p. Then:

$$y_i|x_i, \theta \quad \sim \quad Poisson(x_i^T\theta)$$

Then

$$Pr(y_i|\lambda) = \frac{\lambda^{y_i}e^{-\lambda}}{y_i!}$$

where the mean $\mu = \lambda$ and the variance $\sigma^2 = \lambda$. Recall that the single-parameter exponential family is expressed as:

$$P(x|\eta) = h(x)\exp\{\eta^T T(x) - A(\eta)\}$$

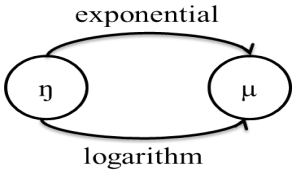
where η is the natural parameter, $T(x)$ is the sufficient statistics, $A(\eta)$: log partition function, and μ is the mean parameter.

We can write out the Poisson distribution in the exponential family form by applying the $\exp(\log(\cdot))$ function:

$$\begin{aligned} P(x|\eta) &= \exp\left\{\log\left(\frac{\lambda^x e^{-\lambda}}{x!}\right)\right\} \\ &= \exp\{x\log\lambda - \lambda - \log x!\} \\ &= \frac{1}{x!}\exp\{x\log\lambda - \lambda\} \end{aligned}$$

where $\eta = \log\lambda$, $T(x) = x$, $A(\eta) = \exp\{\eta\}$, and $\mu = \exp\{\eta\}$.

The relationship of θ and μ is $\mu = \exp\{\eta\}$ and $\eta = \log\mu$. Let us choose to set $\eta = \theta^T x$. Then, since we have set the natural parameter in this way, we can describe the *response function* as the canonical response function for Poisson regression, exp, which maps the natural parameter to the mean parameter, and the canonical *link function*, log, which maps the mean parameter to the canonical parameter.



The distribution can be written in terms of θ and x:

$$P(y_i|x_i,\theta) = \frac{1}{y_i!}\exp\{y_i\theta^T x_i - e^{\theta^T x_i}\}.$$

Further the mean of the distribution can be written as

$$\begin{aligned} \hat{y_i} &= E(y_i|x_i,\theta) \\ &= e^{\theta^T x} \\ &\Rightarrow \hat{y_i} \geq 0 \end{aligned}$$

If, as before, we have n observations from the joint distribution:

$$D = \{(x_1,y_1)(x_2,y_2)...(x_n,y_n)\} \sim P(y_i|x_i,\theta)$$

then we can rewrite the probability:

$$P(y_1y_2...y_n|x_1x_2...x_n,\theta) = \prod_{i=1}^{n}\frac{\exp\{y_i\theta^T x_i - e^{\theta^T x_i}\}}{y_i!}$$

2 Estimating parameter θ

In order to find the maximum likelihood estimate of θ, we use the log likelihood to calculate:

$$l(y|x,\theta) = \sum_{i=1}^{n}[y_i\theta^T x_i - \exp\{\theta^T x_i\} - \log(y_i!)]$$

As before, we can take the partial derivative of the log likelihood with respect to θ to maximize the log likelihood:

$$\begin{aligned} \frac{\partial l(y|x,\theta)}{\partial\theta} &= \sum_{i=1}^{n}[y_i x_i - x_i\exp\{\theta^T x_i\}] \\ &= \sum_{i=1}^{n} x_i(y_i - \exp\{\theta^T x_i\}) \\ &= \sum_{i=1}^{n} x_i(y_i - \hat{y_i}) \qquad \bigstar \end{aligned}$$

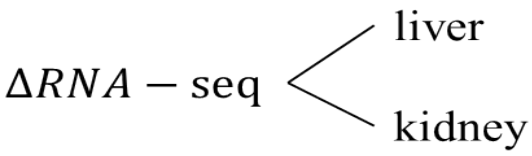
The form of this equation is identical to the same term for logistic regression. As in logistic regression, notice it is impossible to directly solve for θ:

$$\sum_{i=1}^{n} x_i(y_i - e^{\theta^T x}) = 0$$

Instead of a closed-form solution to this equation, we can use a gradient method to estimate θ. One efficient gradient method, described in the last lecture, is Iteratively reweighted least squares (IRLS).

3 Paper: "An assessment of technical reproducibility and comparison with gene expression arrays" [Marioni et al., 2008]

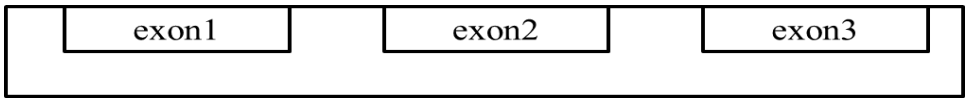
3.1 Example: RNA sequencing: differences in gene expression across cell types



From paper, recall from previous lecture: if there are 31,840 genes and, for a given test, 1487 are significant, and FDR=0.1% ,FPR=1%:

Q: How many False Positives based on FDR value? A = 1.4
Q: How many False Positives based on FPR value? A = 318

Genes often have multiple exons:



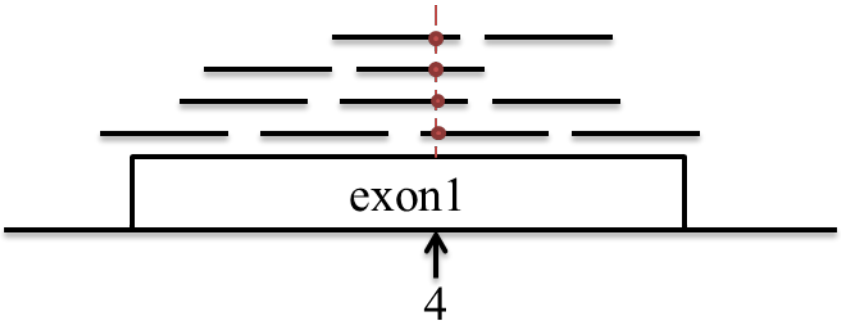
and different *isoforms*, which include different combinations of exons. Each isoform will include junction reads, or RNA-sequencing reads that span exons, for the exon pairs that are in that isoform, e.g.,

Choose i samples (e.g., liver, kidney), j genes and k lanes on each gene. Then

1. perform RNA-sequencing on this sample, generating a large set of sequencing reads, then

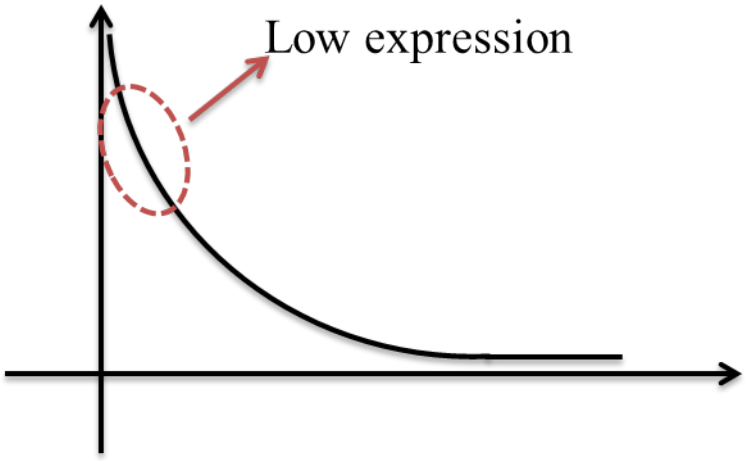
2. map the reads to the underlying (reference) genetic sequence; 32 base pair reads.

Then, the number of mapped reads to each exon will approximate transcription levels of each exon. These two steps both have a lot of bias. For example, if an exon has mutations, insertions, or deletions in the sequence relative to the reference sequence, it is less likely that reads from this exon in that individual will



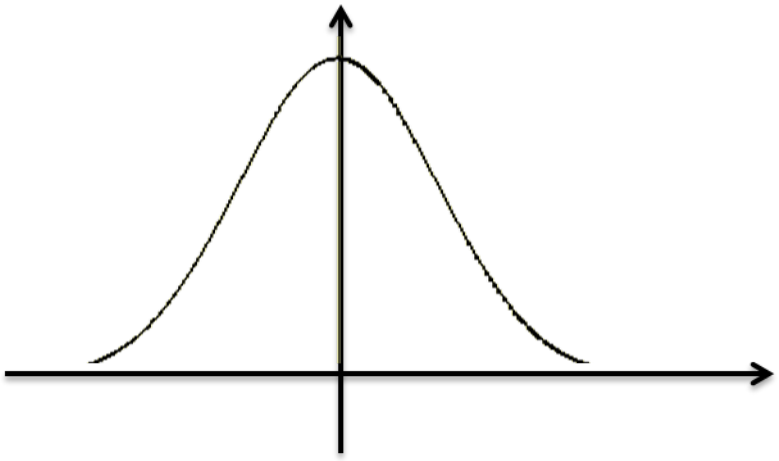
map onto the reference.

3.2 Data Processing



a

If we model the read counts using a linear Gaussian model, we may lose information about low expressed genes. For example, if we consider transforming the distribution of read counts to a standard normal, the genes in the bottom half of read counts will be mapped to the lower half of the Gaussian distribution:



Instead, we can use a Poisson regression model:

X_{ijk} : number of reads mapped to sequence
C_{ik} : total number of reads
λ_{ijk} : rate of transcription of lane k on gene j from sample i
Read counts are taken from particular lane on particular gene from particular sample. Here, they further constrain λ_{ijk} to be a rate parameter as follows: $\sum_{j=1}^{G} \lambda_{ijk} = 1$, and $0 \leq \lambda_{ijk} \leq 1$. Set

$$X_{ijk}|C_{ijk}, \lambda_{ijk} \sim Poisson(C_{ik}\lambda_{ijk})$$

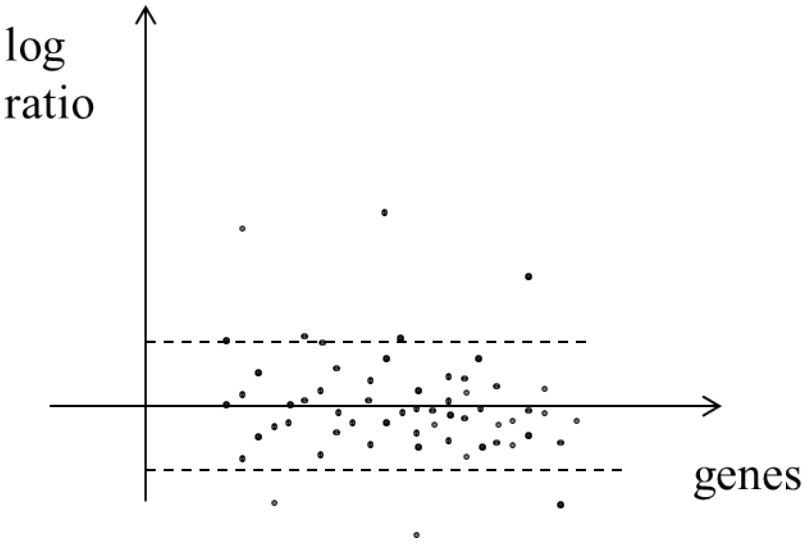
The mean of the distribution can be written as

$$E[X_{ijk}|C_{ijk}, \lambda_{ijk}] = C_{ik}\lambda_{ijk}$$

Note that this definition of the mean parameter is not the same as in the canonical Poisson regression model, since $\eta = \lambda = \mu = C_{ik}\lambda_{ijk}$.

Returning to the previous question, do we see a lane effect (is there differential expression of genes between two lanes in the same sample and the same cell type)? Consider

$$\log \frac{\lambda_{ik}}{\lambda_{ijk}}$$



To see whether there is lane effect, we have our classical hypothesis testing framework:

- H_0 : null hypothesis, there is no lane effect ($\lambda_{ijk} = \lambda_{ijk'}$)
- H_A : alternative hypothesis, there is a lane effect ($\lambda_{ijk} \neq \lambda_{ijk'}$).

In the paper, they only found a small number of genes to show differential expression across lanes, and so conclude that lane effects are minimal.

3.3 Likelihood Ratio Test

In the same framework, we can also consider the question about whether we see differential expression within genes across tissue types. Let us specify the null and alternative hypotheses more formally:

$$\begin{aligned} H_0 : \lambda_{ijk} &= \hat{\lambda_j} \\ H_A : \lambda_{ijk} &= \{\hat{\lambda_j^A}, \hat{\lambda_j^B}\} \end{aligned}$$

Here $A, B \in \{lanes, tissues\}$ and let A = liver, B = kidney. To simplify the equation, take the log of the probability:

$$\begin{aligned} \log P(D|H_0) &= \sum_{i=1}^{n} \log Pr(X_{ijk}|C_{ijk}, \hat{\lambda_j}) \\ \log P(D|H_A) &= \sum_{i=1}^{n} \log Pr(X_{ijk}|C_{ijk}, \hat{\lambda_j^A}, \hat{\lambda_j^B}) \end{aligned}$$

TS: test statistics

$$\begin{aligned} D_{TS} &= -2\log(\frac{P(D|H_0)}{P(D|H_A)}) \\ &= -2\log(P(D|H_0)) + 2\log(P(D|H_A)) \end{aligned}$$

$D_{TS} \sim \chi^2$, this test statistic is distributed according to a chi-squared distribution with degrees of freedom $\nu = 1$ (in this example, comparing two different cell types).

The significance threshold $t(x)$ to control the FDR at a given value was calculated using the method of Storey and Tibshirani (2003), through q-values (can be computed in R)