

CS-E5875 High-Throughput Bioinformatics

RNA-seq analysis: differential expression

Harri Lähdesmäki

Department of Computer Science
Aalto University

November 10, 2020

Contents

- ▶ Linear regression and generalized linear models: basics
- ▶ Differential gene expression analysis
- ▶ Transcript-level analysis

Linear regression¹

- ▶ Recall the multiple linear regression model

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ik} + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where

- ▶ y_i denotes the measured response for the i th sample/data point
- ▶ $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ denotes the regression coefficients
- ▶ $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ denotes the predictors for the i th sample/data point, and
- ▶ ϵ_i denotes the Gaussian observation error for the i th measurement, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$

¹See e.g. (Agresti, 2015) or (Murphy, 2012) or any book on (generalized) linear models

Linear regression¹

- ▶ Recall the multiple linear regression model

$$y_i = \beta_0 + \sum_{k=1}^p \beta_k x_{ik} + \epsilon_i = \mathbf{x}_i^T \boldsymbol{\beta} + \epsilon_i,$$

where

- ▶ y_i denotes the measured response for the i th sample/data point
 - ▶ $\boldsymbol{\beta} = (\beta_0, \beta_1, \dots, \beta_p)^T$ denotes the regression coefficients
 - ▶ $\mathbf{x}_i = (1, x_{i1}, \dots, x_{ip})^T$ denotes the predictors for the i th sample/data point, and
 - ▶ ϵ_i denotes the Gaussian observation error for the i th measurement, $\epsilon_i \sim \mathcal{N}(0, \sigma^2)$
- ▶ Assuming n measurements $\mathbf{y} = (y_1, \dots, y_n)^T$ and $X = (\mathbf{x}_1, \dots, \mathbf{x}_n)^T$, this can be written as

$$\mathbf{y} = X\boldsymbol{\beta} + \boldsymbol{\epsilon},$$

where X contains \mathbf{x}_i s as rows, $\boldsymbol{\epsilon} = (\epsilon_1, \dots, \epsilon_n)^T$ and $\boldsymbol{\epsilon} \sim \mathcal{N}(\mathbf{0}, \sigma^2 I_n)$

¹See e.g. (Agresti, 2015) or (Murphy, 2012) or any book on (generalized) linear models

Linear regression

- ▶ Parameters of the linear regression model are $\theta = (\beta, \sigma^2)$
- ▶ Equivalently, we can write the linear regression model with Gaussian noise as

$$\begin{aligned} p(\mathbf{y} \mid X, \theta) &= L(\theta \mid \mathbf{y}, X) \\ &= \mathcal{N}(\mathbf{y} \mid \boldsymbol{\mu}, \Sigma) \\ &= \mathcal{N}(\mathbf{y} \mid X\boldsymbol{\beta}, \sigma^2 I_n) \\ &= \prod_{i=1}^n \mathcal{N}(y_i \mid \mathbf{x}_i^T \boldsymbol{\beta}, \sigma^2) \\ &= \prod_{i=1}^n \mathcal{N}(y_i \mid \mathbb{E}[y_i], \sigma^2), \end{aligned}$$

where $\mu_i = \mathbb{E}[y_i] = \mathbf{x}_i^T \boldsymbol{\beta}$ denotes the expectation of random variable y_i and σ^2 specifies uncertainty around the expected value

Parameter estimation for linear model with Gaussian noise

- ▶ A common way to estimate parameters is to maximise the likelihood of the observed data w.r.t. model parameters, i.e., the maximum likelihood estimate (MLE)

$$\hat{\theta} = \arg \max_{\theta} p(\mathbf{y} \mid X, \theta)$$

Parameter estimation for linear model with Gaussian noise

- ▶ A common way to estimate parameters is to maximise the likelihood of the observed data w.r.t. model parameters, i.e., the maximum likelihood estimate (MLE)

$$\hat{\theta} = \arg \max_{\theta} p(\mathbf{y} \mid X, \theta)$$

- ▶ In this case it is useful to study the logarithm of the likelihood

$$\begin{aligned}\ell(\theta) &= \log p(\mathbf{y} \mid X, \theta) = \log \prod_{i=1}^n p(y_i | \mathbf{x}_i, \theta) = \sum_{i=1}^n \log p(y_i | \mathbf{x}_i, \theta) \\ &= \sum_{i=1}^n \log \left[\left(\frac{1}{2\pi\sigma^2} \right)^{\frac{1}{2}} \exp \left(-\frac{1}{2\sigma^2} (y_i - \mathbf{x}_i^T \beta)^2 \right) \right] \\ &= -\frac{n}{2} \log(2\pi\sigma^2) - \frac{1}{2\sigma^2} \sum_{i=1}^n (y_i - \mathbf{x}_i^T \beta)^2\end{aligned}$$

- ▶ Instead of maximizing $\ell(\theta)$ one can minimize $-\ell(\theta)$

Parameter estimation for linear model with Gaussian noise

- ▶ Minimum or maximum values of a (log) likelihood function w.r.t. parameters are obtained at parameter values where the gradient of the function, i.e. partial derivatives, are zero
- ▶ For some models, the minimum / maximum can be obtained in a closed form
- ▶ The linear regression model with additive Gaussian noise is one such model:

$$\begin{aligned}\hat{\beta} &= (X^T X)^{-1} X^T \mathbf{y} \\ \hat{\sigma}^2 &= \frac{1}{n} (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}) \\ &= \frac{1}{n} (\mathbf{y} - X\hat{\beta})^T (\mathbf{y} - X\hat{\beta}),\end{aligned}$$

assuming X has full rank and the inverse $(X^T X)^{-1}$ exists

Nonlinearity in the linear regression model

- ▶ To model non-linear function we can replace \mathbf{x} with some non-linear function $\phi(\mathbf{x})$
 - ▶ So-called basis function expansion
 - ▶ Model is still linear in parameters, thus called as linear regression
- ▶ For example, polynomial basis functions

$$\phi(\mathbf{x}) = (1, x, x^2, \dots, x^d)^T$$

- ▶ The above theory works for general basis functions as well

An illustration of the linear regression model with Gaussian noise

► Examples of linear and non-linear regression model fitting

- $\phi(x) = (1, x_1, x_2)$
- $\phi(x) = (1, x_1, x_2, x_1^2, x_2^2)$

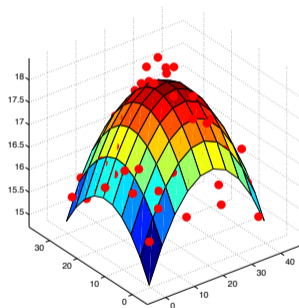
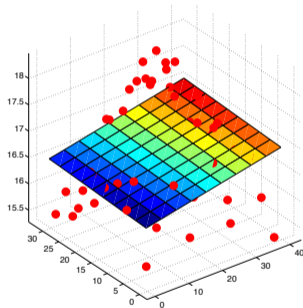


Figure: Figures from (Murphy, 2012)

Comparing two nested linear regression models

- ▶ Often one is interested in
 - ▶ Evaluating the model accuracy, or
 - ▶ Testing the significance of covariates/predictors of the model, either simultaneously or individually
- ▶ A natural measure of how well a model fits the data \mathbf{y} is the so-called residual sum of squares

$$\begin{aligned}\text{RSS} &= (\mathbf{y} - \hat{\mathbf{y}})^T (\mathbf{y} - \hat{\mathbf{y}}) \\ &= (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}})^T (\mathbf{y} - \mathbf{X}\hat{\boldsymbol{\beta}}) \\ &= \sum_{i=1}^n (y_i - \mathbf{x}_i^T \boldsymbol{\beta})^2\end{aligned}$$

- ▶ RSS quantifies the amount of signal in \mathbf{y} that a linear model cannot explain

Comparing two nested linear regression models

- ▶ Assume two nested multiple linear regression models

- ▶ Model 1: $y_i = \beta_0 + \sum_{k=1}^{p_1} \beta_k x_{ik} + \epsilon_i$

- ▶ Model 2: $y_i = \beta_0 + \sum_{k=1}^{p_1} \beta_k x_{ik} + \sum_{k=p_1+1}^{p_1+p_2} \beta_k x_{ik} + \epsilon_i$

Comparing two nested linear regression models

- ▶ Assume two nested multiple linear regression models
 - ▶ Model 1: $y_i = \beta_0 + \sum_{k=1}^{p_1} \beta_k x_{ik} + \epsilon_i$
 - ▶ Model 2: $y_i = \beta_0 + \sum_{k=1}^{p_1} \beta_k x_{ik} + \sum_{k=p_1+1}^{p_1+p_2} \beta_k x_{ik} + \epsilon_i$
- ▶ We can define a test statistic that compares the RSS values between two models as

$$F = \frac{\left(\frac{\text{RSS}_1 - \text{RSS}_2}{\text{df}_1} \right)}{\left(\frac{\text{RSS}_2}{\text{df}_2} \right)},$$

where $\text{df}_1 = (1 + p_1 + p_2) - (1 + p_1) = p_2$ and $\text{df}_2 = n - 1 - p_1 - p_2$

Comparing two nested linear regression models

- ▶ Assume two nested multiple linear regression models
 - ▶ Model 1: $y_i = \beta_0 + \sum_{k=1}^{p_1} \beta_k x_{ik} + \epsilon_i$
 - ▶ Model 2: $y_i = \beta_0 + \sum_{k=1}^{p_1} \beta_k x_{ik} + \sum_{k=p_1+1}^{p_1+p_2} \beta_k x_{ik} + \epsilon_i$
- ▶ We can define a test statistic that compares the RSS values between two models as

$$F = \frac{\left(\frac{\text{RSS}_1 - \text{RSS}_2}{\text{df}_1} \right)}{\left(\frac{\text{RSS}_2}{\text{df}_2} \right)},$$

where $\text{df}_1 = (1 + p_1 + p_2) - (1 + p_1) = p_2$ and $\text{df}_2 = n - 1 - p_1 - p_2$

- ▶ Under the null assumption that the p_2 additional covariates included in model 2 do not provide significantly better fit (i.e., $H_0 : \beta_{p_1+1} = \dots = \beta_{p_1+p_2} = 0$), the F test statistic has F distribution, with $(\text{df}_1, \text{df}_2)$ degrees of freedom
- Significance value from hypothesis testing

Likelihood ratio test

- ▶ Let $L(\hat{\theta}_1 \mid \mathbf{y}, X)$ and $L(\hat{\theta}_2 \mid \mathbf{y}, X)$ denote the maximum likelihoods for the two nested linear models, respectively
- ▶ The likelihood ratio measures how many times less likely the data are under one model (null hypothesis) than the other model (alternative hypothesis)

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\theta}_1 \mid \mathbf{y}, X)}{L(\hat{\theta}_2 \mid \mathbf{y}, X)}$$

- ▶ Intuition:
 - ▶ Values of $\Lambda(\mathbf{y})$ close to 1 indicate there is no difference between the null and alternative models
 - ▶ Small values (close 0) indicate that the alternative model can explain the data much better

Likelihood ratio test

- ▶ Let $L(\hat{\theta}_1 | \mathbf{y}, X)$ and $L(\hat{\theta}_2 | \mathbf{y}, X)$ denote the maximum likelihoods for the two nested linear models, respectively
- ▶ The likelihood ratio measures how many times less likely the data are under one model (null hypothesis) than the other model (alternative hypothesis)

$$\Lambda(\mathbf{y}) = \frac{L(\hat{\theta}_1 | \mathbf{y}, X)}{L(\hat{\theta}_2 | \mathbf{y}, X)}$$

- ▶ Intuition:
 - ▶ Values of $\Lambda(\mathbf{y})$ close to 1 indicate there is no difference between the null and alternative models
 - ▶ Small values (close 0) indicate that the alternative model can explain the data much better
- ▶ An asymptotic result for nested models: when $n \rightarrow \infty$, the test statistic $-2 \log \Lambda(\mathbf{y})$ is chi-squared distributed with degrees of freedom equal to the difference in the number of free parameters between the two models

The likelihood ratio test for the linear Gaussian model

- For the two nested linear regression models with Gaussian noise, the likelihood ratio test can be written as

$$\begin{aligned}\Lambda(\mathbf{y}) &= -2 \log \frac{\max_{\theta_1} L(\theta_1 \mid \mathbf{y}, X)}{\max_{\theta_2} L(\theta_2 \mid \mathbf{y}, X)} \\ &= -2 \log \frac{L(\hat{\theta}_1 \mid \mathbf{y}, X)}{L(\hat{\theta}_2 \mid \mathbf{y}, X)} \\ &= \dots = \left(1 + \frac{\text{RSS}_1 - \text{RSS}_2}{\text{RSS}_2} \right)^{-n/2} \\ &= \left(1 + \frac{p_2}{n - 1 - p_1 - p_2} F \right)^{-n/2}\end{aligned}$$

Generalized linear models

- ▶ Generalized linear models (GLM) are a generalization of linear regression models where the response/dependent variables can have an error distribution other than the normal distribution
- ▶ In standard GLMs the dependent variable is assumed to have a distribution in the exponential family, including e.g.
 - ▶ Normal, exponential, beta, gamma, Poisson, etc. distributions

Generalized linear models

- ▶ Recall that in the case of Gaussian likelihood, $\mathbb{E}[y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- ▶ In GLMs, the mean μ_i of the distribution of random variable y_i is assumed to depend on a linear model via an invertible link function g

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

- ▶ Thus:

$$\mathbb{E}[y_i] = \mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

- ▶ Note that in the case of Gaussian linear model, the link function $g(\cdot)$ is the identity function

Generalized linear models

- ▶ Recall that in the case of Gaussian likelihood, $\mathbb{E}[y_i] = \mu_i = \mathbf{x}_i^T \boldsymbol{\beta}$
- ▶ In GLMs, the mean μ_i of the distribution of random variable y_i is assumed to depend on a linear model via an invertible link function g

$$g(\mu_i) = \mathbf{x}_i^T \boldsymbol{\beta}$$

- ▶ Thus:

$$\mathbb{E}[y_i] = \mu_i = g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta})$$

- ▶ Note that in the case of Gaussian linear model, the link function $g(\cdot)$ is the identity function
- ▶ Variance of a GLM can follow the variance of the exponential family distribution or may be defined as a function $V(\cdot)$ of the predicted value

$$\text{Var}(y_i) \quad \text{or} \quad V(\mu_i, \phi) = V(g^{-1}(\mathbf{x}_i^T \boldsymbol{\beta}), \phi)$$

Generalized linear models

- ▶ Lets illustrate the GLM with the Poisson distribution for the response variables \mathbf{Y} (non-negative count data)
- ▶ Poisson rate parameter(s) λ must be positive, so logarithmic link function is appropriate

$$\log \lambda = X\beta \Leftrightarrow \lambda = \exp(X\beta)$$

- ▶ The variance of error distribution is defined by the Poisson distribution, i.e., $\text{Var}(Y_i) = V(\lambda_i) = \lambda_i = \exp(\mathbf{x}_i\beta)$

Generalized linear models

- ▶ Lets illustrate the GLM with the Poisson distribution for the response variables \mathbf{Y} (non-negative count data)
- ▶ Poisson rate parameter(s) $\boldsymbol{\lambda}$ must be positive, so logarithmic link function is appropriate

$$\log \boldsymbol{\lambda} = X\boldsymbol{\beta} \Leftrightarrow \boldsymbol{\lambda} = \exp(X\boldsymbol{\beta})$$

- ▶ The variance of error distribution is defined by the Poisson distribution, i.e., $\text{Var}(Y_i) = V(\lambda_i) = \lambda_i = \exp(\mathbf{x}_i\boldsymbol{\beta})$
- ▶ Likelihood of observed data $\mathbf{y} = (y_1, \dots, y_n)^T$ is then

$$L(\boldsymbol{\beta} \mid \mathbf{y}, X) = \prod_{i=1}^n \text{Poisson}(y_i \mid \lambda_i) = \prod_{i=1}^n \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!} = \prod_{i=1}^n \frac{\exp(\mathbf{x}_i\boldsymbol{\beta})^{y_i} \exp(-\exp(\mathbf{x}_i\boldsymbol{\beta}))}{y_i!}$$

Generalized linear models

- ▶ Lets illustrate the GLM with the Poisson distribution for the response variables \mathbf{Y} (non-negative count data)
- ▶ Poisson rate parameter(s) λ must be positive, so logarithmic link function is appropriate

$$\log \lambda = X\beta \Leftrightarrow \lambda = \exp(X\beta)$$

- ▶ The variance of error distribution is defined by the Poisson distribution, i.e., $\text{Var}(Y_i) = V(\lambda_i) = \lambda_i = \exp(\mathbf{x}_i\beta)$
- ▶ Likelihood of observed data $\mathbf{y} = (y_1, \dots, y_n)^T$ is then

$$L(\beta \mid \mathbf{y}, X) = \prod_{i=1}^n \text{Poisson}(y_i \mid \lambda_i) = \prod_{i=1}^n \frac{\lambda_i^{y_i} \exp(-\lambda_i)}{y_i!} = \prod_{i=1}^n \frac{\exp(\mathbf{x}_i\beta)^{y_i} \exp(-\exp(\mathbf{x}_i\beta))}{y_i!}$$

- ▶ GLMs are typically fitted using maximum likelihood (or Bayesian) approach
- ▶ Note that for GLMs no closed form solutions exist but numerical methods must be used

Hypothesis testing with GLMs

- ▶ For GLMs the null hypothesis is often stated by restricting the parameter vector

$$H_0 : \beta \in \Theta_0 \subset \mathbb{R}^{p+1}$$

- ▶ Consequently, the alternative hypothesis is defined via the complement of Θ_0 , i.e.,
 $\Theta_0^C = \mathbb{R}^{p+1} \setminus \Theta_0$

$$H_1 : \beta' \in \Theta_0^C$$

Hypothesis testing with GLMs

- ▶ For GLMs the null hypothesis is often stated by restricting the parameter vector

$$H_0 : \beta \in \Theta_0 \subset \mathbb{R}^{p+1}$$

- ▶ Consequently, the alternative hypothesis is defined via the complement of Θ_0 , i.e., $\Theta_0^C = \mathbb{R}^{p+1} \setminus \Theta_0$

$$H_1 : \beta' \in \Theta_0^C$$

- ▶ For example, if one is interested in testing a single predictor x_i , then
 - ▶ $H_0 : \beta_i = 0$ or equivalently $\beta \in \mathbb{R}^p$
 - ▶ $H_1 : \beta_i \neq 0$ or equivalently $\beta' \in \mathbb{R}^{p+1}$

Hypothesis testing with GLMs

- ▶ For GLMs the null hypothesis is often stated by restricting the parameter vector

$$H_0 : \beta \in \Theta_0 \subset \mathbb{R}^{p+1}$$

- ▶ Consequently, the alternative hypothesis is defined via the complement of Θ_0 , i.e., $\Theta_0^C = \mathbb{R}^{p+1} \setminus \Theta_0$

$$H_1 : \beta' \in \Theta_0^C$$

- ▶ For example, if one is interested in testing a single predictor x_i , then
 - ▶ $H_0 : \beta_i = 0$ or equivalently $\beta \in \mathbb{R}^p$
 - ▶ $H_1 : \beta_i \neq 0$ or equivalently $\beta' \in \mathbb{R}^{p+1}$
- ▶ An asymptotic result for nested models: when $n \rightarrow \infty$, the test statistic $-2 \log \Lambda(\mathbf{y})$ is chi-squared distributed with degrees of freedom equal to the difference in dimensionality of Θ_0 and Θ_0^C

Contents

- ▶ Linear regression and generalized linear models: basics
- ▶ Differential gene expression analysis
- ▶ Transcript-level analysis

Differential gene expression analysis

- ▶ Consider our hypothetical differential expression analysis using t -tests from lecture #1
- ▶ Two aspects
 - ▶ Expression difference: how large is the average expression difference between two groups?
 - ▶ Statistical significance: how sure are we that there is a true difference?
- ▶ The latter is a statistical question: hypothesis testing
- ▶ On the next slides we motivate the use of a negative binomial distribution by the following reasoning: multinomial \rightarrow binomial \rightarrow Poisson \rightarrow negative binomial

Multinomial distribution

- ▶ Sequence count data is discrete-valued, so it obviously has a non-Gaussian distribution
 - t -test based methods are not appropriate, or at least not optimal
- ▶ For a single sample, we can assume that read counts for genes (or transcripts) have a multinomial (sampling) distribution

Multinomial distribution

- ▶ Consider the following
 - ▶ A dice that has N different outcomes
 - ▶ The number of genes e.g. in the human genome is $\approx 20,000$
 - ▶ When a dice is rolled once, one of the outcomes will be chosen randomly with probability p_i , where $\sum_{i=1}^N p_i = 1$
 - ▶ “One roll” corresponds to picking a single RNA fragment from a very large pool of fragments for sequencing
 - ▶ Assume an experiment where dice is rolled N times (i.i.d.)
 - ▶ A sequencing run can produce e.g. 10M-1B sequencing reads
 - ▶ Denote the number of times each outcome is observed by $\mathbf{x} = (x_1, \dots, x_N)$, where $x_1 + \dots + x_N = n$ (the number of reads mapped to each gene)
 - ▶ Denote $\mathbf{p} = (p_1, \dots, p_N)$
 - ▶ The unknown abundances/proportions of different genes

Multinomial distribution

- ▶ Consider the following
 - ▶ A dice that has N different outcomes
 - ▶ The number of genes e.g. in the human genome is $\approx 20,000$
 - ▶ When a dice is rolled once, one of the outcomes will be chosen randomly with probability p_i , where $\sum_{i=1}^N p_i = 1$
 - ▶ “One roll” corresponds to picking a single RNA fragment from a very large pool of fragments for sequencing
 - ▶ Assume an experiment where dice is rolled N times (i.i.d.)
 - ▶ A sequencing run can produce e.g. 10M-1B sequencing reads
 - ▶ Denote the number of times each outcome is observed by $\mathbf{x} = (x_1, \dots, x_N)$, where $x_1 + \dots + x_N = n$ (the number of reads mapped to each gene)
 - ▶ Denote $\mathbf{p} = (p_1, \dots, p_N)$
 - ▶ The unknown abundances/proportions of different genes
- ▶ The probability mass function of the random variable $X = (X_1, \dots, X_N)$ that has the multinomial distribution

$$\begin{aligned}\text{Multinomial}(\mathbf{x}; n, \mathbf{p}) &= P(X_1 = x_1, \dots, X_N = x_N) \\ &= \begin{cases} \frac{n!}{x_1! \dots x_N!} p_1^{x_1} p_2^{x_2} \dots p_N^{x_N}, & \text{if } x_1 + \dots + x_N = n \\ 0, & \text{otherwise} \end{cases}\end{aligned}$$

Multinomial distribution

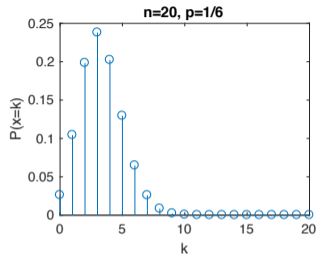
- ▶ Can be considered as sampling noise (or “technical” noise)
- ▶ The use of multinomial is somewhat challenging because we would need to model all genes at the same time

Binomial distribution

- ▶ Each of the components of a multinomial distribution separately (e.g. a gene) has a binomial distribution
 - ▶ For example, the probability that we obtain a sequencing read from gene i is $p = p_i$, and the probability that we obtain a sequencing read from any other gene is $1 - p = \sum_{j \neq i} p_j$
- ▶ Consider a binary-valued random variable that takes value 1 with probability p
- ▶ Take n independent random realizations of the binary-valued random variable
- ▶ Let X denote the number of success in n realizations
- ▶ The probability of getting exactly k successes in n trials is given by probability mass function of the binomial distribution

$$B(k; n, p) = P(X = k) = \binom{n}{k} p^k (1 - p)^{n-k}$$

Binomial distribution

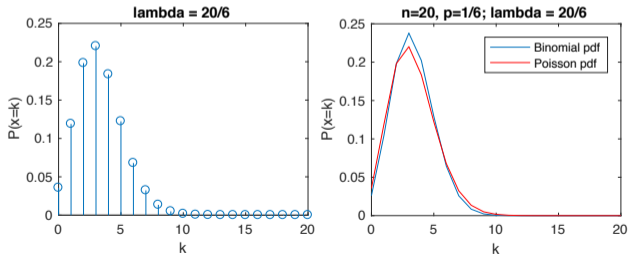


Poisson distribution

- ▶ Consider a discrete random variable X that can have values $0, 1, 2, \dots$
- ▶ The discrete random variable X has a Poisson distribution with rate parameter λ if

$$\text{Poisson}(k; \lambda) = P(X = k) = \frac{\lambda^k \exp(-\lambda)}{k!}$$

- ▶ For larger number of trials n (i.e., the number of sequencing reads in an experiment) with a small probability p , binomial can be approximated by Poisson distribution



Negative binomial distribution

- ▶ Read counts across biological replicates is observed to have a larger variance than what Poisson model suggests
 - ▶ So-called overdispersed noise
 - ▶ Biological variability/noise
- ▶ Negative binomial has been found to provide a good fit to sequencing count data

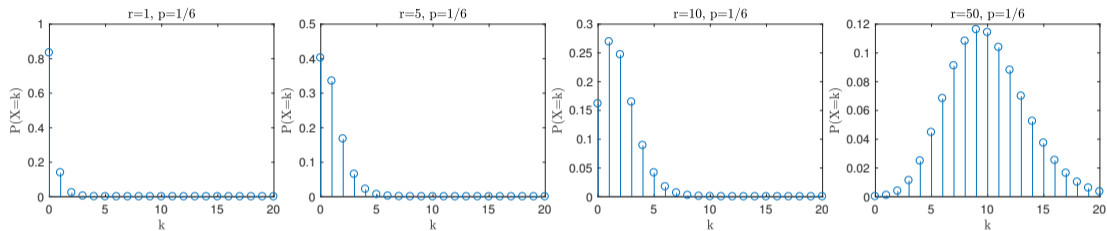
Negative binomial distribution

- ▶ Read counts across biological replicates is observed to have a larger variance than what Poisson model suggests
 - ▶ So-called overdispersed noise
 - ▶ Biological variability/noise
- ▶ Negative binomial has been found to provide a good fit to sequencing count data
- ▶ The negative binomial distribution is a discrete probability distribution of the number of successes (denoted X) in a sequence of i.i.d. Bernoulli trials (with probability p) before a specified (non-random) number of failures (denoted r) occurs
- ▶ Random variable X has the negative binomial distribution with probability mass function

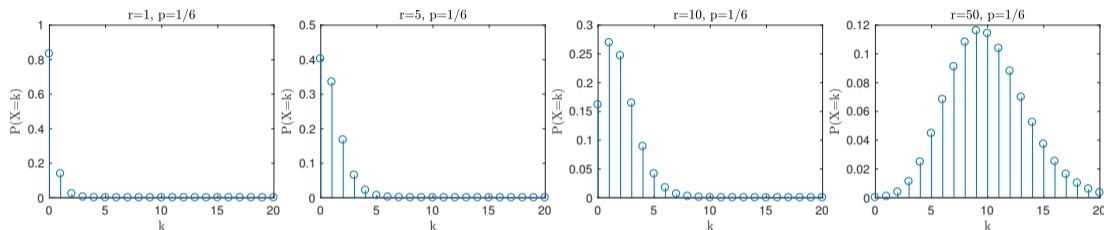
$$\text{NB}(k; r, p) = P(X = k) = \binom{r + k - 1}{k} p^k (1 - p)^r$$

- ▶ The negative binomial distribution has several alternative formulations: see e.g. https://en.wikipedia.org/wiki/Negative_binomial_distribution
- ▶ Be careful, especially when using in different programming languages!

Negative binomial distribution



Negative binomial distribution



- Negative binomial distribution occurs in many contexts
 - Negative binomial distribution can be derived as a continuous mixture of Poisson distributions where the mixing distribution is a gamma distribution

$$\text{NB}(k; r, p) = \int_0^{\infty} \text{Poisson}(k; \lambda) \text{Gamma}\left(\lambda; r, \frac{1-p}{p}\right) d\lambda$$

Compound distributions

- ▶ Assume a random variable X with a distribution F (and density p_f) with parameters θ
- ▶ Assume that the parameters θ of F have a mixing distribution G (density p_g)
 - ▶ Distribution F is compounded by G

$$p(x) = \int p_f(x|\theta)p_g(\theta)d\theta$$

- ▶ Recall the definition of the joint and marginal distributions

$$p(x, y) = p(x|y)p(y) \text{ and } p(x) = \int p(x, y)dy = \int p(x|y)p(y)dy$$

Compound distributions

- ▶ Typical usage:
 - ▶ Overdispersion modeling
 - ▶ Need to model a greater amount of variability than what would be expected by a given baseline model
 - ▶ Bayesian inference
 - ▶ Predictive distribution of future data $p(y^*|\theta)$ given the posterior distribution of model parameters θ conditioned on observed data y , $p(y^*|y) = \int p(y^*|\theta)p(\theta|y)d\theta$
- ▶ Commonly used compound distributions in bioinformatics
 - ▶ Gamma-Poisson, i.e., negative binomial
 - ▶ Beta-binomial
 - ▶ Dirichlet-multinomial

Gamma-Poisson compound distributions

$$\begin{aligned}f(k; r, p) &= \int_0^{\infty} f_{\text{Poisson}(\lambda)}(k) \cdot f_{\text{Gamma}\left(r, \frac{1-p}{p}\right)}(\lambda) \, d\lambda \\&= \int_0^{\infty} \frac{\lambda^k}{k!} e^{-\lambda} \cdot \lambda^{r-1} \frac{e^{-\lambda(1-p)/p}}{\left(\frac{p}{1-p}\right)^r \Gamma(r)} \, d\lambda \\&= \frac{(1-p)^r p^{-r}}{k! \Gamma(r)} \int_0^{\infty} \lambda^{r+k-1} e^{-\lambda/p} \, d\lambda \\&= \frac{(1-p)^r p^{-r}}{k! \Gamma(r)} p^{r+k} \Gamma(r+k) \\&= \frac{\Gamma(r+k)}{k! \Gamma(r)} p^k (1-p)^r.\end{aligned}$$

Copy-pasted from wikipedia: https://en.wikipedia.org/wiki/Negative_binomial_distribution

Negative binomial distribution

- The mean and variance of negative binomial distribution are

$$\mathbb{E}[X] = \mu = \frac{pr}{1-p} \quad \text{and} \quad \mathbb{V}[X] = \sigma^2 = \frac{pr}{(1-p)^2}$$

Negative binomial distribution

- ▶ The mean and variance of negative binomial distribution are

$$\mathbb{E}[X] = \mu = \frac{pr}{1-p} \quad \text{and} \quad \mathbb{V}[X] = \sigma^2 = \frac{pr}{(1-p)^2}$$

- ▶ For our application it is useful to reparameterized NB using the mean and variance

$$\text{NB}(\mu, \sigma^2) \doteq \text{NB}(r, p),$$

where

$$r = \frac{\mu^2}{\sigma^2 - \mu} \quad \text{and} \quad p = \frac{\sigma^2 - \mu}{\sigma^2}$$

Negative binomial distribution

- ▶ The mean and variance of negative binomial distribution are

$$\mathbb{E}[X] = \mu = \frac{pr}{1-p} \quad \text{and} \quad \mathbb{V}[X] = \sigma^2 = \frac{pr}{(1-p)^2}$$

- ▶ For our application it is useful to reparameterized NB using the mean and variance

$$\text{NB}(\mu, \sigma^2) \doteq \text{NB}(r, p),$$

where

$$r = \frac{\mu^2}{\sigma^2 - \mu} \quad \text{and} \quad p = \frac{\sigma^2 - \mu}{\sigma^2}$$

- ▶ Further, we will consider a parameterization

$$\text{NB}(\mu, \phi) \doteq \text{NB}(\mu, \sigma^2),$$

where ϕ defines the variance as $\sigma^2 = \mu + \phi\mu^2$

Differential gene expression analysis

- ▶ We will look at edgeR (McCarthy et al., 2012), a versatile and efficient modeling method for sequencing count data
- ▶ Assume that the number of aligned reads in sample j that are assigned to gene g can be modelled by negative binomial distribution

$$N_{gj} \sim \text{NB}(s_j \lambda_{gj}, \phi_g)$$

where

- ▶ s_j is the so-called library size: e.g. the total number of reads from sample j , or some other normalization quantity
- ▶ λ_{gj} is the proportion of RNA fragments that originate from gene g in sample j
 - ▶ Note that $\sum_g \lambda_{gj} = 1$
- ▶ ϕ_g is the dispersion for gene g that defines the over-dispersion and thus the variance in the negative binomial model

Differential gene expression analysis

- For the above definition of NB distribution the mean and variance for N_{gj} are

$$\mathbb{E}[N_{gj}] = \mu_{gj} = s_j \lambda_{gj} \quad (1)$$

$$\mathbb{V}[N_{gj}] = \mu_{gj} + \phi_g \mu_{gj}^2 = s_j \lambda_{gj} + \phi_g s_j^2 \lambda_{gj}^2 \quad (2)$$

- Recall that for the standard Poisson model $\mathbb{E}[N_{gj}] = \mu_{gj}$ and $\mathbb{V}[N_{gj}] = \mu_{gj}$

Differential gene expression analysis

- ▶ Often one is interested in comparing two populations A and B, i.e., $H_0 : \lambda_{gA} = \lambda_{gB}$
- ▶ edgeR implements a general linear model (GLM) with NB distribution that allows comparison of two population means as well as many other more complex experimental designs
- ▶ In GLM the mean $\mu_{gj} = s_j \lambda_{gj}$ of the NB is modeled with a log-linear model

$$\log \lambda_{gj} = \mathbf{x}_j^T \boldsymbol{\beta}_g$$

$$\log \mu_{gj} = \mathbf{x}_j^T \boldsymbol{\beta}_g + \log s_j$$

$$\log \mu_{gj} = \beta_0 + \sum_{k=1}^p x_{jk} \beta_{gk} + \log s_j,$$

- ▶ \mathbf{x}_j is a vector that contains all p covariates for sample j , and
 - ▶ $\boldsymbol{\beta}_g$ is a vector that contains the corresponding parameters for gene g
- ▶ The mean of the NB distribution is $\mu_{gj} = \exp(\mathbf{x}_j^T \boldsymbol{\beta}_g + \log s_j)$
- ▶ Recall that variance is defined as $\mu_{gj} + \phi \mu_{gj}^2$

Differential gene expression analysis

- ▶ Consider a simple example with 4 samples, 2 from group A and 2 from group B
- ▶ The linear model and the design matrix for the null hypothesis model (lets call it M_0) that assumes only one population/condition is (i.e., no difference between A and B)

$$\begin{pmatrix} \log \mu_{g1} \\ \log \mu_{g2} \\ \log \mu_{g3} \\ \log \mu_{g4} \end{pmatrix} = \begin{pmatrix} 1 \\ 1 \\ 1 \\ 1 \end{pmatrix} \begin{pmatrix} \beta_g \end{pmatrix} + \begin{pmatrix} \log s_1 \\ \log s_2 \\ \log s_3 \\ \log s_4 \end{pmatrix},$$

- ▶ The model for the alternative hypothesis with two conditions (M_1) can be written e.g.

$$\begin{pmatrix} \log \mu_{g1} \\ \log \mu_{g2} \\ \log \mu_{g3} \\ \log \mu_{g4} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ 0 & 1 \\ 0 & 1 \end{pmatrix} \begin{pmatrix} \beta_{gA} \\ \beta_{gB} \end{pmatrix} + \begin{pmatrix} \log s_1 \\ \log s_2 \\ \log s_3 \\ \log s_4 \end{pmatrix},$$

where samples 1 and 2 are from condition A and samples 3 and 4 are from condition B

Differential gene expression analysis

- ▶ Let's denote the observed read counts for gene g as $\mathbf{y}_g = (n_{g1}, \dots, n_{g4})^T$ (in the previous example we have 4 samples)
- ▶ In edgeR, statistical hypothesis testing for differential gene expression between conditions A and B can be implemented e.g. with the likelihood-ratio test

$$T = -2 \ln \frac{\ell(\hat{\beta}_g, \hat{\phi}_g | \mathbf{y}_g, M_0)}{\ell(\hat{\beta}_{gA}, \hat{\beta}_{gB}, \hat{\phi}_g | \mathbf{y}_g, M_1)}$$

- ▶ $\ell(\cdot)$ is the NB density function
- ▶ $\hat{\beta}_g$ denotes the maximum likelihood estimate of β_g given \mathbf{y}_g and M_0 (similarly for other parameters)
- ▶ The test statistic T is approximately chi-squared distributed with degrees of freedom equal to $\text{df}_{M_1} - \text{df}_{M_0}$, where df_M denotes the number of free parameters of model M
 - p -value
 - ▶ Remember multiple testing

Differential gene expression analysis

- ▶ In many applications the number of biological replicates is too small to allow accurate estimation of both λ_{gj} and ϕ_j
 - ▶ edgeR tool implements a moderated test where information between genes is shared that allows more accurate dispersion estimation
- ▶ The so-called adjusted profile likelihood (APL) for dispersion ϕ_g is

$$APL_g(\phi_g) = \ell(\phi_g | \mathbf{y}_g, \hat{\beta}_g) - \frac{1}{2} \log \det \mathcal{I}_g$$

- ▶ ϕ_g is free parameter
- ▶ $\hat{\beta}_g$ is the ML estimate of β_g that depends on ϕ_g
- ▶ \mathcal{I}_g is the Fisher information matrix

Differential gene expression analysis

- ▶ One possible assumption is that all genes have the same dispersion value $\phi_g = \phi$
- ▶ A shared dispersion can be estimated by maximizing the sum of the adjusted profile likelihoods

$$APL_S(\phi) = \sum_{g=1}^G APL_g(\phi)$$

- ▶ In essence, data across all genes is shared to estimate variance/dispersion
- ▶ edgeR tool provides also options for other dispersion estimates
 - ▶ Trended: group genes into bin that have similar mean read count
 - ▶ Gene-wise

Differential gene expression analysis

- ▶ An example from edgeR User Guide (Chen et al, 2017)
- ▶ Three patient with oral squamous cell carcinomas
 - ▶ Oral squamous cell carcinomas and matched normal tissue from each patient
 - ▶ RNA-seq experiments paired experimental design
- ▶ Goal: detect genes differentially expressed between tumour and normal tissue
- ▶ Samples: 8N, 8T, 33N, 33T, 51N, 51T
- ▶ Design matrix X is

	(Intercept)	Patient33	Patient51	TissueT
8N	1	0	0	0
8T	1	0	0	1
33N	1	1	0	0
33T	1	1	0	1
51N	1	0	1	0
51T	1	0	1	1

Figure from (Chen et al, 2017)

Differential gene expression analysis

- Variance dependence on the mean (biological coefficient of variation equals the square root of the dispersion)

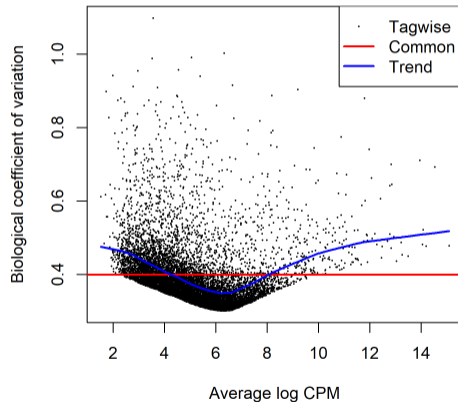


Figure from (Chen et al, 2017)

Differential gene expression analysis

- ▶ 1269 genes differentially expressed with FDR 5%
- ▶ Additionally, require at least 2-fold change (blue horizontal lines below)
- ▶ MA plot: a scatter plot where a dot corresponds to a gene g , x-axis shows mean gene expression $\frac{1}{2} \log X_{gA} X_{gB}$ and y-axis shows difference $\log \frac{X_{gA}}{X_{gB}}$

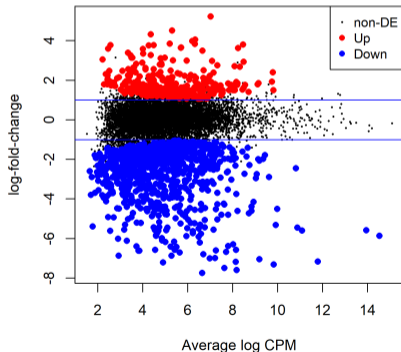


Figure from (Chen et al, 2017)

Contents

- ▶ Linear regression and generalized linear models: basics
- ▶ Differential gene expression analysis
- ▶ Transcript-level analysis

Transcript-level expression quantification

- ▶ Let us assume that each gene i is associated with J_i transcripts indexed by j , then

$$\begin{aligned}\theta_{ij} &= P(\text{sample a read from transcript } j \text{ associated with gene } i) \\ &= \frac{1}{Z} \mu_{ij} \ell_{ij},\end{aligned}$$

where

- ▶ μ_{ij} is the expression level of transcript j associated with gene i
 - ▶ ℓ_{ij} is the length of transcript j of gene i
 - ▶ Normalizing constant is $Z = \sum_{ij} \mu_{ij} \ell_{ij}$
- ▶ The true expression level of gene i is

$$\mu_i = \sum_{j=1}^{J_i} \mu_{ij}$$

Transcript-level expression quantification

- ▶ Lets denote the aligned RNA-seq reads as R_1, R_2, \dots, R_N
- ▶ Let us also make an unrealistic assumption that all reads are assigned **uniquely** to one of the transcripts
- ▶ Then the frequency estimator gives us

$$\hat{\theta}_{ij} = \frac{k_{ij}}{N},$$

where k_{ij} is the number of reads assigned uniquely to μ_{ij}

- ▶ Correspondingly, we can convert the estimates into expression values by normalizing by the transcript length

$$\hat{\mu}_{ij} \propto \sum_j \frac{\hat{\theta}_{ij}}{\ell_{ij}} = \sum_j \frac{k_{ij}}{\ell_{ij} N}$$

Transcript-level expression quantification

- Recall the union method for estimating the gene expression level

$$k_i = \sum_j k_{ij}$$

and the frequency estimator

$$\hat{\theta}_i = \frac{k_i}{\ell_i},$$

where ℓ_i is the length of the gene i

- Union method tends to underestimate the gene expression level because

$$\begin{aligned}\hat{\theta}_i &= \frac{\sum_j k_{ij}}{\ell_i} = \frac{k_{i1}}{\ell_i} + \dots + \frac{k_{iJ_i}}{\ell_i} \\ &\leq \frac{k_{i1}}{\ell_{i1}} + \dots + \frac{k_{iJ_i}}{\ell_{iJ_i}},\end{aligned}$$

where $\ell_i \geq \ell_{ij}$

Transcript-level expression quantification

- Consider a simple case of skipped exon

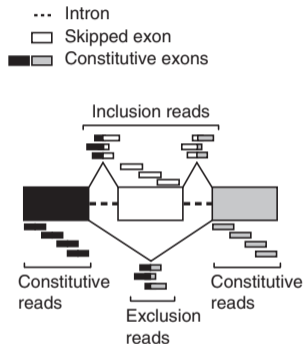


Figure from (Katz et al., 2010)

- We can use e.g. the reads in the skipped exon and the inclusion and exclusion reads together with the frequency estimator to estimate the relative expression of the two transcripts

Transcript-level expression quantification

- ▶ With paired end reads we can try to use all (non-uniquely) aligned reads assuming we can estimate insert length variability

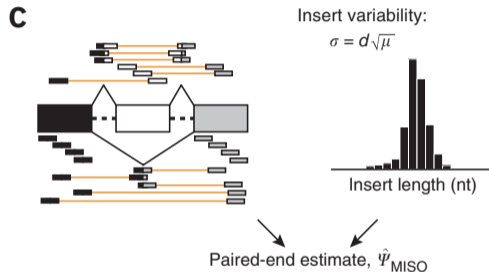


Figure from (Katz et al., 2010)

- ▶ Estimation can be done Markov chain Monte Carlo (MCMC) sampling (Katz et al., 2010)

References

- ▶ Agresti, Alan. Foundations of Linear and Generalized Linear Models, John Wiley & Sons, 2015
- ▶ Chen Y, et al., edgeR: differential expression analysis of digital gene expression data, User's Guide, 11 October 2017
- ▶ Murphy K, Machine learning: a probabilistic perspective, MIT Press, 2012
- ▶ Katz Y, et al., Analysis and design of rnA sequencing experiments for identifying isoform regulation, Nature Methods, 7(12):1009-15, 2010.