پردیس علوم
دانشکده ریاضی، آمار و علوم کامپیوتر

روش‌های آماری برای شناسایی ژن‌های متفاوت‌بیان‌شده

نگارنده: مرضیه خاکباز

استاد راهنما: دکتر سمانه افتخاری مهابادی

مقطع کارشناسی در رشته ریاضیات و کاربردها

بهمن ۱۴۰۱

چکیده

یک امر رایج در تجزیه و تحلیل داده‌های ریزآرایه، تعیین ژن‌هایی است که در دو نوع نمونه بافت یا نمونه‌های به دست آمده در دو شرایط آزمایشی بیان متفاوتی دارند. اخیراً چندین روش آماری برای دستیابی به این هدف پیشنهاد شده است، برای زمانی که نمونه‌های تکراری تحت هرکدام از شرایط وجود داشته باشند. با این حال، ممکن است مشخص نباشد که چگونه این روش‌ها با یکدیگر مقایسه می‌شوند. در اینجا هدف اصلی ما مقایسه‌ی سه روش آماری است؛ آزمون‌t، یک مدل رگرسیونی [۱] و یک مدل آمیخته [۲] با توجه ویژه به فرضیات هر مدل است.

هر سه روش مبتنی بر استفاده از tـآماره دو نمونه‌ای یا تغییرات جزئی آن هستند، اما در نحوه‌ی مرتبط کردن سطح معنی‌داری آماری با آماره‌ی مربوطه متفاوت هستند که منجر به تفاوت زیاد در سطح معنی‌داری و تعداد ژن‌های شناسایی شده می‌شود. در ادامه با استفاده از داده‌های سرطان خون [۳] این موارد را توضیح می‌دهیم.

[1] Thomas et al., Genome Res., 11, 1227–1236, 2001

[2] Pan et al., http://www.biostat.umn.edu/cgi-bin/rrs?print+2001, 2001a,b

[3] leukemia data of Golub et al. (Science, 285, 531–537, 1999)

فهرست مطالب

سپاسگزاری

از دکتر افتخاری عزیز که با حوصله و دقت فراوان مرا در انجام این پروژه یاری کردند کمال تشکر را دارم.

پیشگفتار

یکی از پیشرفت‌های هیجان‌انگیز در ژنگان‌شناسی [۴] ، استفاده از فناوری ریزآرایه برای نمایش هم‌زمان سطح بیان هزاران ژن می‌باشد. یک امر رایج، مقایسه سطح بیان ژن‌ها در نمونه‌های گرفته شده از دو بافت مختلف یا در دو نقطه زمانی یا شرایط مختلف است. اوایل، روش ساده‌ی تغییرات فولد [۵] مورد استفاده قرار می گرفت ، اما به دلیل عدم توجه آن به تغییرپذیری آماری، روشی غیرمطمئن شناخته شد، و از آن زمان به بعد روش های آماری پیچیده‌تری ارائه شدند [۶]. همچنین مشاهده شد که داده‌های مبتنی بر یک آرایه ممکن است قابل اعتماد نباشند و دارای نوفه‌های [۷] زیادی باشند. با پیشرفت فناوری، آزمایش‌های ریزآرایه ارزان‌تر شدند و استفاده از آرایه های چندگانه [۸] را امکان پذیر کردند. در این پروژه ما ژن‌های متفاوت بیان شده (با نمونه‌گیری تکراری) تحت هرکدام از شرایط را تشخیص می‌دهیم.

[4] genomics

[5] fold changes

[6] e.g.Chen et al., 1997; Efron et al., 2000; Ideker et al., 2000; Newton et al., 2001.

[7] noise

[8] multiple arrays

فصل ۱

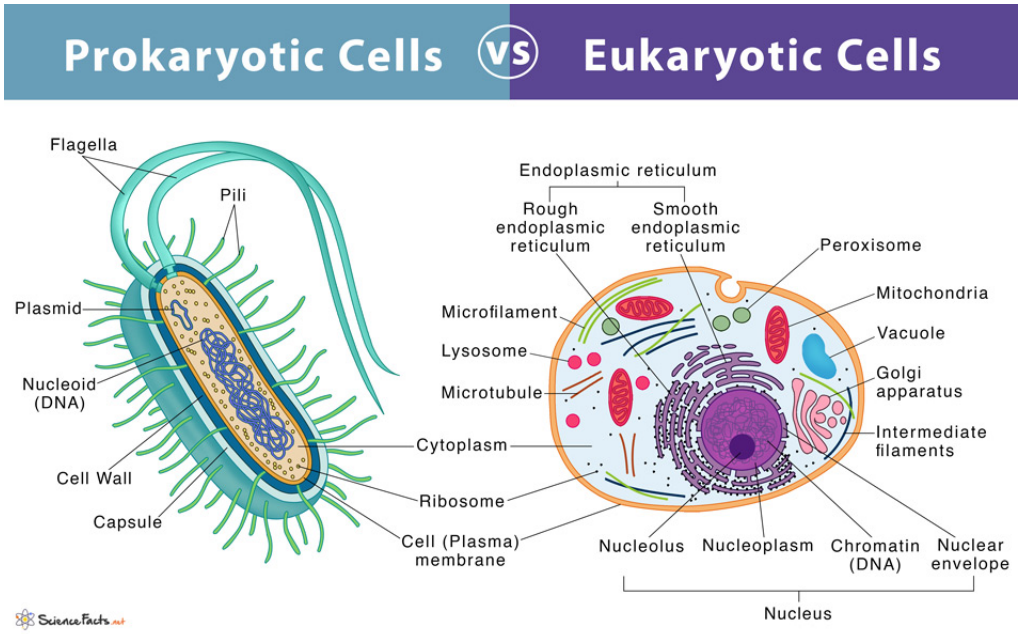
مفاهیم زیستی

۱.۱ سلول

سلول‌ها بنیادی‌ترین واحد سازنده بدن تمام موجودات زنده محسوب می‌شوند. برخی از موجودات تک‌سلولی هستند، به این معنا که یک سلول به خودی خود یک موجود زنده است. اما در گونه‌های بالاتر مانند حیوانات و گیاهان، یک موجود از هزاران میلیارد سلول تشکیل شده است.

سلول ها دو نوع دارند: سلول‌های پروکاریوتی و سلول‌های یوکاریوتی. همانطور که در شکل ۱.۱ مشاهده می‌کنید، سلول‌های یوکاریوتی هسته دارند و پروکاریوتی‌ها ندارد. موجودات زنده نیز بر این اساس، که آیا سلول‌هایشان هسته دارند یا نه، به دو دسته‌ی یوکاریتونی و پروکاریوتی تقسیم می‌شوند. پروکاریوت‌ها دارای قدمت بیشتری هستند که شامل باکتری‌ها و آرکی‌ها هستند. باقی موجودات پیشرفته‌تر یوکاریوتی هستند، مانند گیاهان و حیوانات.

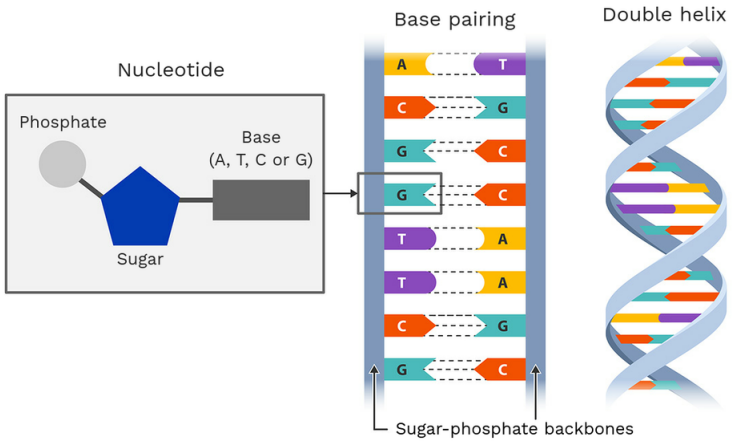
سلول‌های یوکاریوتی دارای ساختار پیچیده‌تری هستند و در هسته‌ی آن‌ها مواد ژنتیکی کلیدی، مانند DNA، به فرم کروموزوم یا کروموزوم‌ها قرار دارند.



شکل ۱.۱: سلول یوکاریوتی و پروکاریوتی

۲.۱ DNA

دئوکسی‌ریبونوکلئیک اسید یا به اختصار دی‌اِن‌ای ، گونه‌ای اسید نوکلئیک است که دارای دستورالعمل‌های ژنتیکی است که برای کارکرد و توسعهٔ زیستی جانداران و ویروس‌ها مورد استفاده قرار می‌گیرد. نقش اصلی مولکول دی‌ان‌ای ذخیره‌سازی طولانی مدت اطلاعات ژنتیکی و دستوری است. دی‌ان‌ای پلی‌مری است که مونومر آن نوکلئوتیدها هستند. یک نوکلئوتید شامل یک گروه فسفات و یک کربوهیدرات پنج‌کربنه (دئوکسی‌ریبوز) و یک باز آلی است. هر نوکلئوتید تنها یک باز آلی نیتروژن‌دار دارد که آن می‌تواند از گونه آدنین (A) ، گوانین (G) ، سیتوزین (C) یا تیمین (T) این بازها یا پریمیدین (تک‌حلقه‌ای) یا پورین (دو حلقه‌ای) می‌باشد. شکل ۲.۱ تصویری از ساختار DNA را نشان می‌دهد.



شکل ۲.۱: ساختار DNA

۳.۱ RNA

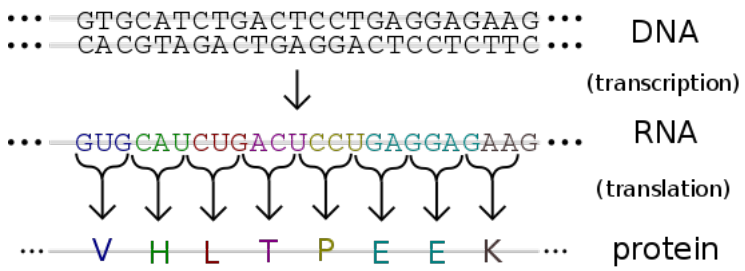
اسید ریبونوکلئیک یا RNA یک مولکول تک‌رشته‌ای است. و مانند DNA دارای چهار نوع باز است با این تفاوت که به جای تیمین، دارای یوراسیل (U) است.

۴.۱ رونویسی و ترجمه

همانطور که در شکل ۳.۱ مشاهده می‌کنید اطلاعات کدگذاری شده‌ی درون سلول، طی فرایندی به نوعی RNA که mRNA می‌نامیم، تبدیل می‌شود که این فرایند را رونویسی می‌گوییم. سپس اطلاعات درون mRNA به پروتئین تبدیل می‌شود که به آن ترجمه می‌گوییم.

۵.۱ ژن

ژن دنباله‌ای از نوکلوئوتیدها است که در برگیرنده اطلاعات لازم جهت تولید مولکول‌های RNA یا پروتئین‌های لازم برای سلول هست. هر ژن در بخشی از DNA سلول وجود دارد. به تعریفی



شکل ۳.۱: فرآیند رونویسی و ترجمه

دیگر ژن، بخشی از مولکول دی‌ان‌ای می‌باشد که روی یک رشته از آن قرار دارد طی فرایند رونویسی ژن‌ها به مولکول‌های RNA تبدیل می‌شوند که یا به شکل مستقیم در سلول استفاده می‌شوند یا دربرگیرنده اطلاعاتی جهت تولید پروتئین هستند و طی فرایند ترجمه، پروتئین مربوط به آن‌ها ساخته می‌شود. ژن‌ها تمامی صفات سلول را کنترل می‌کنند و عملکرد سلول‌ها به کمک ژن‌ها و پروتئین‌های ساخته‌شده از روی آن‌ها تعیین می‌شود. این ژن‌ها از پدر و مادر به ارث می‌رسند و ممکن است به مرور در اثر فرایند تقسیم سلولی یا در تعامل سلول با محیط بیرونی دچار تغییر شوند.

۶.۱ بیان ژن

بیان ژن فرایندی است که در آن اطلاعات درون ژن استفاده می‌شود تا یک محصول کاربردی از آن بدست آید. فرایند بیان ژن به وسیله تمام یوکاریوت‌ها و پروکاریوت‌ها (باکتری‌ها و...) انجام می‌شود. مراحل مختلفی را می‌توان برای فرایند بیان ژن در نظر گرفت که عموماً شامل رونویسی، ترجمه و تغییرات بعد از ترجمه یک پروتئین می‌باشد. تنظیم بیان ژن به سلول این امکان را می‌دهد تا بتواند ساختار و کاربرد خود را کنترل کند. از آنجا که تمام سلول‌های بدن انسان از یک سلول مشتق شده‌اند تفاوت‌ها و تمایزات بین سلول‌ها حاصل از بیان شدن یا نشدن قسمت‌هایی از ژن است. بیان ژن همچنان می‌تواند به عنوان یکی از زیر لایه‌های تکامل در نظر گرفته شود زیرا کنترل زمان بندی، مکان و مقدار ژن می‌تواند تأثیرات مهمی در عملکرد ژن‌ها درون سلول یا کل ارگانیسم

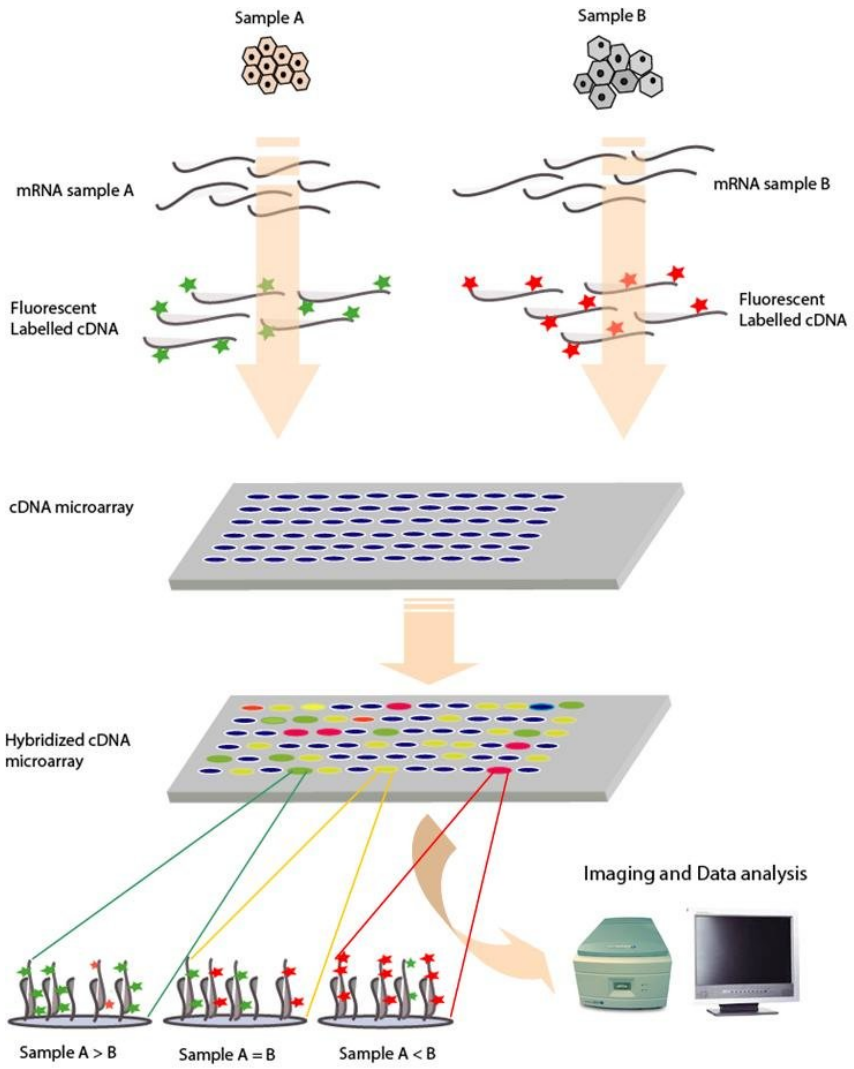
پرسلولی داشته باشد.

۷.۱ ریزآرایه

ریزآرایه‌های [1] دی‌ان‌ای، امکان غربال همزمان و سریع هزاران ژن را فراهم می‌سازد. ریزآرایهٔ دی‌ان‌ای (که به آن بیوچیپ نیز گفته می‌شود) متشکل از تعداد زیادی نقطه بر روی یک سطح جامد است (شکل ۴.۱). بر روی این نقاط رشته‌های دی‌ان‌ای سوار شده‌اند که به آنها پروب گفته می‌شود. دانشمندان به منظور اندازه‌گیری میزان بیان ژن‌ها از این روش استفاده می‌کنند. هر نقطه‌ای که بر روی صفحه قرار دارد حاوی ۱۲–۱۰ مول از یک توالی خاص دی‌ان‌ای است. این توالی می‌تواند بخش کوچکی از یک ژن یا بخش متصل شونده به یک cDNA یا cRNA باشد که تحت شرایط خاصی به توالی مکمل متصل می‌گردد. توالی‌های هدف معمولاً با موادی از جمله فلوئوروفور، نقره یا کمیلومینسنس [2] نشاندار شده‌اند و به این وسیله میزان اتصال رشته‌های نوکلئیک اسیدی به توالی‌های هدف تشخیص داده می‌شود. در مرحله بعد این داده‌ها توسط نرم‌افزارهای ویژه‌ای کمی سازی می‌گردند.

[1] microarrays

[2] chemiluminescence



شکل ۴.۱: فرآیند آنالیز داده‌های بیان ژن با ریزآرایه‌ها

فصل ۲

روش‌های تشخیص ژن‌های متفاوت بیان‌شده

۱.۲ داده‌ها

فرض‌کنید که Y_{jk} $(j = 1, ..., n; k = 1, ..., K_1, K_1 + 1, ..., K_1 + K_2)$ سطح بیان ژن j در آرایه‌ی k باشد. K_1 تا نمونه‌ی اول از بیماران با سرطان نوع اول ، و K_2 تا نمونه‌ی آخر از بیماران با سرطان نوع دوم است.

یک مدل آماری برای داده ها به صورت زیر است

$$Y_{jk} = a_j + b_j x_k + \epsilon_{jk} \qquad (۱.۲)$$

ϵ_{jk} خطای تصادفی با میانگین ۰ است، و x_k یک متغیر پیشگوی دو وجهی است که شرایط نمونه‌ی kام را نشان می‌دهد.

$$\begin{cases} x_k = 1 & 1 \leq k \leq K_1 \\ x_k = 0 & K_1 + 1 \leq k \leq K_1 + K_2 \end{cases}$$

در نتیجه میانگین سطح بیان ژن j ام برای دو نمونه به صورت زیر می‌باشد

$$\begin{cases} E(Y_{jk}|X) = a_j + b_j & 1 \leq k \leq K_1 \\ E(Y_{jk}|X) = a_j & K_1 + 1 \leq k \leq K_1 + K_2 \end{cases}$$

تعیین اینکه آیا ژنی بیان متفاوتی دارد یا خیر، معادل است با

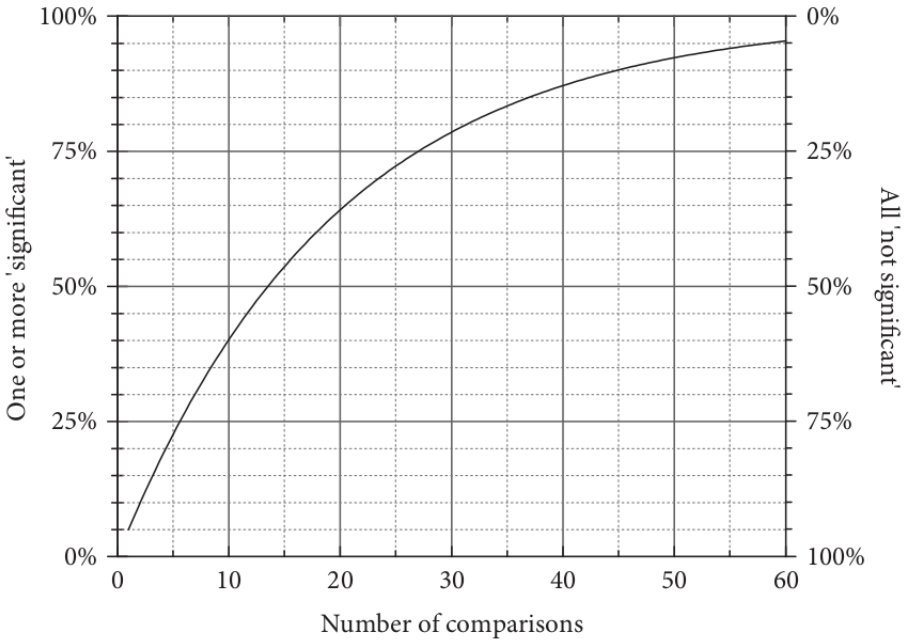
$$\begin{cases} H_0 : b_j = 0 \\ H_1 : b_j \neq 0 \end{cases}$$

انجام آزمون فرض شامل دو گام می‌باشد. گام اول ساخت آماره‌ی آزمون و گام بعدی تعیین سطح معناداری یا پی‌ـمقدارها مرتبط با آماره‌ی آزمون است.

پی‌ـمقدارها معمولا براساس توزیع آماره‌ی آزمون تحت فرض صفر محاسبه می‌شود.

۱.۱.۲ مساله‌ی مقایسه‌های چندگانه

زمانی که تعداد فرض‌های مورد آزمون در یک پژوهش افزایش می‌یابد، احتمال رخداد خطای نوع اول نیز بیشتر می‌شود (رد فرض H_0 وقتی که درست است). به طور مثال اگر $\alpha = 0.05$ و k تا فرض صفر داشته باشیم، شانس اینکه هیچ‌کدام معنی‌دار نباشند 0.95^k و شانس اینکه یکی (یا بیشتر) از مقایسه‌ها از لحاظ آماری معنی دار باشند $1 - 0.95^k$ است. شکل ۱.۲ این احتمال را برای تعداد مختلف مقایسه‌های چندگانه نشان می‌دهد.



شکل ۱.۲: به طور مثال اگر ۱۳ تا مقایسه‌ی مستقل داشته باشیم (همه‌ی فرض‌های صفر درست باشند) احتمال اینکه یکی (یا بیشتر) از پی‌ـمقدارها کمتر از 0.05 شود 50% است . یعنی یک شانس $50 : 50$ برای دستیابی به حداقل یک معنی‌داری کاذب وجود دارد.

۲.۱.۲ تصحیح بونفرونی

تصحیح بونفرونی [1] یک راه حل ساده برای حل مساله‌ی مقایسه‌های چندگانه با تقسیم α بر تعداد مقایسه‌ها می‌باشد. یعنی در یک پژوهش زمانی نتایج از لحاظ آماری معنی‌دار هستند که پی‌ـ‌مقدار کمتر از $\frac{\alpha}{k}$ (k تعداد مقایسه‌ها است) باشد.

در اینجا ما $\alpha = 0.01$ را برای سطح معناداری در سطح ژنوم را درنظر میگیریم و برای مقایسه‌های چندگانه از تصحیح بونفرونی و $\alpha^* = \frac{\alpha}{2n}$ (برای آزمون دوطرفه) استفاده می‌کنیم.

۲.۲ روش آزمون t

روش‌های گوناگونی برای بدست آوردن آماره‌ی آزمون برای دو نمونه وجود دارند که با توجه به سایز نمونه‌ها (K_1 و K_2) و برابری یا عدم برابری واریانس، یکی را انتخاب می‌کنیم . با توجه به اینکه اندازه‌های K_1 و K_2 اکثرا کوچک هستند و همچنین شواهد، عدم برابری واریانس را نشان می‌دهند، در نتیجه ما از آماره‌ی آزمون به روش ولچ [2] برای دو نمونه‌ی مستقل نرمال با واریانس‌های نابرابر، استفاده می‌کنیم.

میانگین نمونه‌ای و واریانس نمونه‌ای برای دو گروه به صورت زیر می‌باشد

$$\overline{Y}_{j(1)} = \frac{\sum_{k=1}^{K_1} Y_{jk}}{K_1}$$

$$\overline{Y}_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2} Y_{jk}}{K_2}$$

[1] Bonferroni Adjustment

[2] Welch

$$s^2_{j(1)} = \frac{\sum_{k=1}^{K_1}(Y_{jk} - \overline{Y}_{j(1)})^2}{K_1 - 1}$$

$$s^2_{j(2)} = \frac{\sum_{k=K_1+1}^{K_1+K_2}(Y_{jk} - \overline{Y}_{j(2)})^2}{K_2 - 1}$$

بنابراین

$$\overline{Y}_{j(1)} \sim N\left(\mu_1, \frac{s^2_{j(1)}}{K_1}\right)$$

و

$$\overline{Y}_{j(2)} \sim N\left(\mu_2, \frac{s^2_{j(2)}}{K_2}\right)$$

دو نمونه مستقل هستند. در نتیجه داریم

$$\overline{Y}_{j(1)} - \overline{Y}_{j(2)} \sim N\left(\mu_1 - \mu_2, \frac{s^2_{j(1)}}{K_1} + \frac{s^2_{j(2)}}{K_2}\right)$$

آماره‌ی آزمون به صورت زیر است

$$Z_j = \frac{\overline{Y}_{j(1)} - \overline{Y}_{j(2)}}{\sqrt{\frac{s^2_{j(1)}}{K_1} + \frac{s^2_{j(2)}}{K_2}}}$$

تحت فرض نرمال بودن دو نمونه، Z_j به صورت تقریبی دارای توزیع t با درجه‌آزادی زیر است

$$d_j = \frac{\frac{s^2_{j(1)}}{K_1} + \frac{s^2_{j(2)}}{K_2}}{(s^2_{j(1)}/K_1)^2/(K_1 - 1) + (s^2_{j(2)}/K_2)^2/(K_2 - 1)}$$

۳.۲ رویکرد مدل‌سازی رگرسیونی

توماس [۳] و همکاران یک روش رگرسیونی [۴] پیشنهاد دادند که مدل آن همان معادله‌ی ۱.۲ است و در آن (a_j, b_j) را به روش کمترین توان دوم موزون [۵] برآورد می‌کنند و سپس واریانس $\hat{b_j}$ را به کمک استوار [۶] یا برآوردگر فشرده‌ی واریانس [۷] برآورد می‌کنند. آماره‌ی آزمون به صورت زیر است

$$Z'_j = \frac{\hat{b_j}}{\sqrt{var(\hat{b_j})}}$$

توماس مشاهده کرد نتایج بدست آمده مبتنی بر استفاده از Z'_j بسیار نزدیک به Z_j است. توجه کنید که معادله‌ی ۱.۲ می‌تواند به عنوان یک مساله‌ی رگرسیونی نوشته شود. ماتریس قطری بلوکی XX^T [۸] به صورت زیر است

$$\begin{bmatrix} 1 \\ \vdots \\ 1 \\ 0 \\ \vdots \\ 0 \end{bmatrix} \times \begin{bmatrix} 1 & \dots & 1 & 0 & \dots & 0 \end{bmatrix} = \begin{bmatrix} 1 & \dots & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 1 & \dots & 1 & \dots & 0 \\ \vdots & \ddots & \vdots & \ddots & \vdots \\ 0 & \dots & 0 & \dots & 0 \end{bmatrix}$$

از این رو برای ژن‌های مختلف، $(\hat{a_j}, \hat{b_j})$های بدست‌آمده با روش کمترین توان دوم از هم

[3] Thomas et al.
[4] regression model
[5] weighted least square method
[6] robust
[7] sandwich variance estimator
[8] block-diagonal matrix

مستقل هستند.

اگر برآوردگر $\hat{b_j}$ بنویسم

$$\hat{b_j} = \frac{S_{xy}}{S_{xx}} = \frac{\sum y_{jk}(x_k - \overline{x})}{S_{xx}} = \frac{\sum y_{jk}x_k - \overline{x}\sum y_{jk}}{\sum_{k=1}^{K_1}(x_k - \overline{x})^2 + \sum_{k=K_1+1}^{K_1+K_2}(x_k - \overline{x})^2}$$

و میانگین پیشگوها برابر است با

$$\overline{x} = \frac{\sum_{i=1}^{K_1+K_2} x_i}{N} = \frac{K_1}{K_1 + K_2}$$

با جایگذاری $x_k = \{1, 0\}$ ها و $\overline{x}$ عبارت زیر بدست می‌آید

$$\hat{b_j} = \overline{Y}_{j(1)} - \overline{Y}_{j(2)}$$

مک‌کولا [9] و درام [10] نشان دادند که برآورد واریانس استوار $\hat{b_j}$ برابر عبارت زیر است

$$var(\hat{b_j}) = \frac{s^2_{j(1)}}{K_1}\frac{K_1 - 1}{K_1} + \frac{s^2_{j(2)}}{K_2}\frac{K_2 - 1}{K_2}$$

واضح است که اگر K_1 و K_2 به بی‌نهایت میل کنند، Z_j و Z'_j با هم معادلند. گرچه برای K_1 و K_2های کوچک ،به علت نااریب بودن برآوردگر واریانس ، Z_j ترجیح داده می‌شود.
علاوه بر این، استفاده از توزیع نرمال استاندارد برای محاسبه‌ی پی‌ـمقدار برای Z'_j براساس فرض بزرگی K_1 و K_2 است، که در بسیاری از آزمایش‌های ریزآرایه برقرار نیست. در نتیجه استفاده از این فرض عملکرد خوبی ندارد.

[9] McCullagh
[10] Drum

۴.۲ رویکرد مدل‌سازی آمیخته

مشکل آزمون t و روش رگرسیونی این است که فرض قوی روی توزیع صفر آماره‌ی آزمون می‌گذارند. در مقابل افرون [11] و همکاران، توشر[12] و همکاران ، پن[13] و همکاران روشی [14] مستقیم برای برآورد مستقیم توزیع صفر (f_0) پیشنهاد دادند. در این روش ما از داده‌های تکراری استفاده می‌کنیم و نیاز داریم که جفت اعداد K_1 و K_2 زوج باشند.

اگر فرض صفر برقرار باشد، توزیع آماره‌ی آزمون برابر با توزیع صفر است. در اینجا فرض این است که ژن‌های متناظر دارای سطح‌بیان متفاوت نیستند. برای برآورد توزیع صفر Z_j می‌توانیم از آزمون جایگشتی استفاده کنیم.

۱.۴.۲ برآورد توزیع صفر

در این روش ما آماره‌ی z_j را برای هر ژن می‌سازیم و از آن برای برآورد توزیع Z_j تحت فرض صفر (f_0) استفاده می‌کنیم.
ابتدا آماره‌ی صفر z_j را می‌سازیم

$$z_j = \frac{Y_{j(1)}p_j/K_1 - Y_{j(2)}q_j/K_2}{\sqrt{\frac{s^2_{j(1)}}{K_1} + \frac{s^2_{j(2)}}{K_2}}} \qquad (۲.۲)$$

که در آن

$$Y_{j(1)} = (Y_{j1}, \ldots, Y_{j,K_1})$$

[11] Efron et al. (2000)
[12] Tusher et al. (2000)
[13] Pan et al. (2000)
[14] mixture modeling approach

$$Y_{j(2)} = (Y_{j,K_1+1}, \ldots, Y_{j,K_1+K_2})$$

و p_j یک بردار ستونی از جایگشت تصادفی $K_1/2$ تا 1 و -1 است و q_j یک بردار ستونی از جایگشت تصادفی $K_2/2$ تا 1 و -1 است. از z_j استفاده می‌کنیم تا توزیع f_0 را به روش ناپارامتری برآورد کنیم. z_j از همان داده‌های آماره‌ی Z_j ساخته شده است.
در آماره‌ی z_j نصف داده‌های هر دسته از نصف دیگر کم می‌شوند. به طورمثال

$$z_j = \frac{Y_{j1} - Y_{j2}, \ldots, Y_{j,K_1-1} - Y_{j,K_1}}{K_1 s_{j(1)}} + \frac{Y_{j,K_1+1} - Y_{j,K_1+2}, \ldots, Y_{j,K_1+K_2-1} - Y_{j,K_1+K_2}}{K_2 s_{j(2)}}$$

با قراردادن $x_k = \{1, 0\}$ در Y_{jk} ها داریم

$$z_j = \frac{(a_j + b_j + \epsilon_{j1}) - (a_j + bj + \epsilon_{j2}), \ldots, +(a_j + b_j + \epsilon_{j,K_1-1}) - (a_j + bj + \epsilon_{j,K_1})}{K_1 s_{j(1)}}$$

$$+ \frac{(a_j + \epsilon_{j,K_1+1}) - (a_j + \epsilon_{j,K_1+2}), \ldots, +(a_j + \epsilon_{j,K_1+K_2-1}) - (a_j + \epsilon_{j,K_1+K_2})}{K_2 s_{j(2)}}$$

بعد از انجام محاسبات داریم

$$z_j = \frac{\epsilon_{j1} - \epsilon_{j2}, \ldots, +\epsilon_{j,K_1-1} - \epsilon_{j,K_1}}{K_1 s_{j(1)}} + \frac{\epsilon_{j,K_1+1} - \epsilon_{j,K_1+2}, \ldots, +\epsilon_{j,K_1+K_2-1} - \epsilon_{j,K_1+K_2}}{K_2 s_{j(2)}}$$

فرض‌کنید که توزیع Z_jها f و توزیع z_jها f_0 است. با استفاده از داده‌های z_j و Z_j ، می‌توانیم f_0 و f را به‌صورت مستقیم برآورد کنیم. نکته‌ی کلیدی این است که برای ژن‌هایی که سطح بیان متفاوتی ندارند، توزیع Z_j ها هم f_0 است. اگر فرض‌کنیم که توزیع Z_jها برای ژن‌هایی با سطح بیان تغییریافته f_1 باشد، می‌توانیم f را به صورت ترکیبی از f_0 و f_1 بنویسیم

$$f = p_0 f_0 + p_1 f_1$$

که p_1 برابرست با نسبت ژن‌های با سطح بیان تغییریافته و $p_0 = 1 - p_1$.

در اینجا ما از یک رویکرد فراوانی‌گرای [15] ناپارامتری برای برآورد f_0 و f_1 ، به صورت مستقیم، استفاده می‌کنیم. باوجود اندازه‌ی نمونه‌ی بزرگ n ، انتخاب یک روش پارامتری غیرضروری به نظر می‌رسد. در اینجا ما ترجیح می‌دهیم برای برآورد توابع چگالی f_0 و f_1 از مدل‌های متناهی آمیخته [16] استفاده‌کنیم که یک ابزار انعطاف‌پذیر و قدرتمند برای مدل‌سازی پدیده‌های مختلف تصادفی ارائه می‌دهند.

برای داده‌های پیوسته، مانند داده‌های بیان ژن، استفاده از مؤلفه‌های نرمال در یک توزیع آمیخته، طبیعی است. توجه داشته باشید که یک مدل آمیخته‌ی نرمال اساساً یک برآوردگر ناپارامتری چگالی است. یک مدل آمیخته‌ی نرمال، برآورد پایدارتری برای احتمالات دم ارائه می‌دهد. احتمالات دم نقش مهمی در معنی‌داری آزمون آماری دارند. همچنین تعیین ناحیه‌ی رد برای آزمون نسبت درستنمایی را تسهیل می‌بخشند.

۲.۴.۲ مدل آمیخته‌ی نرمال

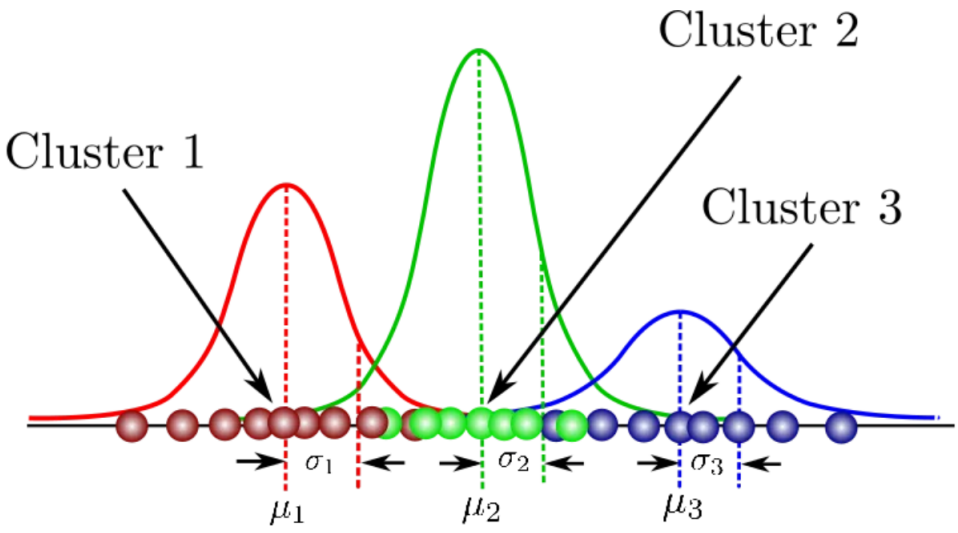
یک آمیخته‌ی نرمال، یک تابع متشکل از $g \in \{1, \ldots, G\}$ توزیع نرمال است، که G تعداد توزیع‌های مجموعه داده‌ها است. شکل ۲.۲ یک مثال از مدل آمیخته را نشان‌می‌دهد.

هر توزیع نرمال در توزیع آمیخته متشکل از پارامترهای زیر است

- یک میانگین μ
- یک کوواریانس Σ
- یک احتمال آمیخته‌ی π که میزان بزرگی تابع نرمال را مشخص می‌کند

[15] frequentist

[16] finite mixture models



شکل ۲.۲: مدل آمیخته‌ی نرمال وقتی که $G = 3$

ضرایب آمیخته [۱۷] احتمالاتی هستند به طوریکه

$$\Sigma_{g=1}^{G} \pi_g = 1$$

تابع چگالی نرمال در حالت کلی به‌صورت زیر است

$$\mathcal{N}(x|\mu, \Sigma) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} exp\left(-\frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu)\right)$$

برای پیداکردن مقدار بهینه‌ی پارامترها از بیشینه درستنمایی استفاده می‌کنیم که برای آن از تابع چگالی لگاریتم می‌گیریم

[17] mixing coefficients

$$\ln \mathcal{N}(x|\mu, \Sigma) = -\frac{D}{2} \ln 2\pi - \frac{1}{2} \ln \Sigma - \frac{1}{2}(x-\mu)^T \Sigma^{-1}(x-\mu) \qquad (۳.۲)$$

حال می‌توانیم با مشتق‌گیری و برابر صفر قراردادن معادله‌ی بالا، مقادیر بهینه‌ی پارامترها را بدست بیاوریم. ولی چون ما تنها با یک توزیع روبه‌رو نیستیم و چند تا توزیع وجود دارد، این راه بسیار پیچیده خواهد شد، در نتیجه ما احتیاج به استفاده از روش‌هایی برای ساده‌تر کردن مساله داریم.

عبارت زیر میزان احتمال اینکه داده‌ی x_n از توزیع g ام بیاید را بیان می‌کند

$$p(m_{ng} = 1|\mathbf{x_g})$$

که در آن m متغیر پنهان [۱۸] است که تنها دو مقدار صفر و یک را می‌گیرد. اگر x از توزیع g آمده باشد مقدار m یک، و در غیر اینصورت صفر است.

به همین ترتیب می‌توانیم قرار دهیم

$$\pi_g = p(m_g = 1)$$

بدین معنی است که احتمال اینکه یک داده‌ی مشاهده شده از توزیع g ام بیاید برابر با ضریب آمیخته معادل است. منطقی بنظر می‌رسد، زیرا هر چه توزیع نرمال بزرگتر شود، ما انتظار داریم که این احتمال هم بزرگتر شود. حال $\mathbf{M}$ را مجموعه‌ی تمام متغیرهای پنهان m درنظر بگیرید.

$$\mathbf{M} = \{\mathbf{m_1}, \ldots, \mathbf{m_g}\}$$

[18] latent variable

می‌دانیم که m ها مستقل از هم رخ می‌دهند و تنها زمانی مقدار یک را می‌گیرند که نقطه‌ی داده شده از خوشه‌ی gام بیاید. از این رو داریم

$$p(\mathbf{M}) = p(m_1 = 1)^{m_1} p(m_2 = 1)^{m_2} \dots p(m_g = 1)^{m_g} = \prod_{m=1}^{G} \pi_g^{m_g}$$

حال احتمال مشاهده‌ی داده ی n ام به شرط آنکه از توزیع g ام بیاید به‌صورت زیر است

$$p(\mathbf{x_n}|\mathbf{m}) = \prod_{g=1}^{G} \mathcal{N}(\mathbf{x_n}|\mu_g, \mathbf{\Sigma_g})^{m_g}$$

بااستفاده از قانون بیز [19] داریم

$$p(\mathbf{x_n}, \mathbf{m}) = p(\mathbf{x_n}|\mathbf{m})p(\mathbf{m})$$

با توجه به اینکه ما نیاز به $p(\mathbf{x_n})$ خواهیم داشت، نه $p(\mathbf{x_n}, \mathbf{m})$، از حاشیه‌ای سازی [20] برای از بین‌بردن آن استفاده می‌کنیم.

$$p(\mathbf{x_n}) = \sum_{g=1}^{G} p(\mathbf{x_n}|\mathbf{m})p(\mathbf{m}) = \sum_{g=1}^{G} \pi_g \mathcal{N}(\mathbf{x_n}|\mu, \Sigma)$$

این معادله‌ای است که نرمال آمیخته را نشان می‌دهد که به وضوح می‌توان دید که به پارامترهایی که قبلا ذکر کردیم وابسته است. برای تعیین مقدار بهینه‌ی پارامترها نیاز به تعیین بیشینه درستنمایی مدل داریم که از طریق احتمال توأم همه‌ی مشاهدات x_n آن را بدست می‌آوریم

[19] Bayes rule

[20] Marginalization

$$p(\mathbf{X}) = \prod_{n=1}^{N} p(\mathbf{x_n}) = \prod_{n=1}^{N} \sum_{g=1}^{G} \pi_g \mathcal{N}(\mathbf{x_n}|\mu_g, \Sigma_g)$$

سپس از طرفین لگاریتم می‌گیریم

$$\ln p(\mathbf{X}) = \sum_{n=1}^{N} \ln \sum_{g=1}^{G} \pi_g \mathcal{N}(\mathbf{x_n}|\mu_g, \Sigma_g) \qquad (۴.۲)$$

بنظر می‌رسد عبارت بالا برای بدست آوردن پارامترها مناسب باشد، ولی لگاریتمی که دومین مجموع را شامل می‌شود محاسبه را سخت می‌کند که در ادامه با استفاده از روشی تکراری [۲۱] پارامترها را برآورد می‌کنیم. اما قبل از آن ما باید احتمال m به شرط $\mathbf{x}$ را بدست آوریم.

با استفاده از قانون بیز داریم

$$p(m_g = 1|\mathbf{x_n}) = \frac{p(\mathbf{x_n}|m_g = 1)p(m_g = 1)}{\sum_{j=1}^{G} p(\mathbf{x_n}|m_j = 1)p(m_j = 1)} \qquad (۵.۲)$$

در قسمت‌های قبل عبارات زیر را بدست آوردیم

$$p(m_g = 1) = \pi_g, \quad p(m_g = 1|\mathbf{x_n}) = \mathcal{N}(\mathbf{x_n}|\mu_g, \Sigma_g)$$

با جایگذاری این عبارات در ۵.۲ داریم

$$p(m_g = 1|\mathbf{x_n}) = \frac{\pi_g \mathcal{N}(\mathbf{x_n}|\mu_g, \Sigma_g)}{\sum_{g=1}^{G} \pi_g \mathcal{N}(\mathbf{x_n}|\mu_j, \Sigma_j)} = \gamma(m_{ng}) \qquad (۶.۲)$$

[21] iterative method

۳.۴.۲ الگوریتم امید ریاضی- بیشینه‌سازی

در قسمت قبل دیدیم که بدست آوردن مقدار بهینه‌ی پارامترها از معادله‌ی ۴.۲ بسیار سخت می‌باشد. ما برای دستیابی به این هدف از یک روش تکراری استفاده می‌کنیم. این روش را امید ریاضی- بیشینه سازی [۲۲] می‌گویند که کاربرد گسترده‌ای در مسائل بهینه‌سازی دارد.

پارامترهای مدل ما مقادیر زیر هستند

$$\theta = \{\pi, \mu, \Sigma\}$$

در قدم اول θ را مقدار دهی اولیه می‌کنیم. برای مثال، می‌توانیم نتایج بدست آمده از الگوریتم K -میانگین را به عنوان نقاط شروع استفاده کنیم.

در قدم بعدی باید مقدار تابع زیر را حساب کنیم

$$\mathcal{Q}(\theta^*, \theta) = \mathbb{E}[\ln p(\mathbf{X}, \mathbf{M}|\theta^*) = \sum_{M} p(\mathbf{M}|\mathbf{X}, \theta) \ln p(\mathbf{X}, \mathbf{M}|\theta^*) \qquad (۷.۲)$$

در عبارت بالا $p(\mathbf{M}|\mathbf{X}, \theta)$ همان $\gamma(m_{ng})$ است که قبلا بدست آوردیم. حال می‌خواهیم مقدار $p(\mathbf{X}, \mathbf{M}|\theta^*)$ را بدست آوریم که درستنمایی کامل [۲۳] مدل است که شامل $\mathbf{X}$ و $\mathbf{M}$ است. آن را به صورت زیر می‌نویسیم

$$p(\mathbf{X}, \mathbf{M}|\theta^*) = \prod_{n=1}^{N} \prod_{g=1}^{G} \pi^{m_{ng}} \mathcal{N}(\mathbf{x_n}|\mu_g, \Sigma_g)^{m_{ng}}$$

این عبارت احتمال تؤام داده‌ها و متغیرهای پنهان است و یک توسیع برای مشتقات آغازین [۲۴]

[22] Expectation-Maximization Algorithm

[23] complete likelihood

[24] initial derivations

$p(\mathbf{X})$ است. لگاریتم این عبارت به صورت زیر است

$$\ln p(\mathbf{X}, \mathbf{M}|\theta^*) = \sum_{n=1}^{N} \sum_{g=1}^{G} m_{ng}[\ln \pi_g + \ln \mathcal{N}(\mathbf{x_n}|\mu_g, \Sigma_g)] \qquad (۸.۲)$$

حال با وجود معادله‌ی بالا می‌توان خیلی راحت‌تر مقادیر بهینه‌ی پارامترها را بدست آورد و مانند معادله‌ی ۴.۲ سخت نیست. جمع‌های بالا هنگامی محاسبه می‌شوند که مقدار m یک باشد. در نتیجه می‌توان عبارت بالا را ساده‌تر کرد.

با جایگذاری ۶.۲ و ۸.۲ در ۷.۲ داریم

$$\mathcal{Q}(\theta^*, \theta) = \sum_{M} \gamma(m_{ng})[\ln \pi_g + \ln \mathcal{N}(\mathbf{x_n}|\mu_g, \Sigma_g)] \qquad (۹.۲)$$

در قدم سوم از $\mathcal{Q}$ نسبت به پارامترها مشتق می‌گیریم و مساوی صفر قرار می‌دهیم. مقادیر بهینه‌ی پارامترها برابرند با

$$\pi_g^* = \frac{\sum_{n=1}^{N} \gamma(m_{ng})}{N}$$

$$\mu^* = \frac{\sum_{n=1}^{N} \gamma(m_{ng})\mathbf{x_n}}{\sum_{n=1}^{N} \gamma(m_{ng})}$$

$$\Sigma_g^* = \frac{\sum_{n=1}^{N} \gamma(m_{ng})(\mathbf{x_n} - \mu_g)(\mathbf{x_n} - \mu_g)^T}{\sum_{n=1}^{N} \gamma(m_{ng})}$$

۴.۴.۲ معیار اطلاع بیزی‌ـشوارتز

معیار اطلاع بیزی‌ـشوارتز [۲۵] یک معیار برای انتخاب مدل از میان مجموعه‌ی محدودی از مدل‌ها است. این معیار تا حدی مبتنی بر تابع درستنمایی می‌باشد.
زمانیکه که مدل را برازش می‌دهیم، با افزودن پارامترها، درستنمایی افزایش می‌یابد و ممکن است موجب بیش‌برازشی [۲۶] شود. معیار بیزی‌ـشوارتز این مشکل را به کمک جمله‌ی تاوان [۲۷] برای تعداد پارامترها در مدل حل می‌کند.

بیان ریاضی معیار به‌صورت زیر است

$$BIC = v_k \ln(N) - 2 \ln \hat{L}$$

که در عبارت بالا

- $\hat{L} = f_0(x; \theta_k)$ یا همان مقدار بیشینه شده‌ی تابع درستنمایی مدل است.
- $v_k = 3G - 1$ تعداد پارامترهای $\hat{L}$ است.

روش کار بدین صورت است که این معیار مجموعه‌ای از مدل‌ها با مقادیر مختلف g را برازش می‌دهد و سپس g متناظر با اولین کمینه محلی [۲۸] معیار را انتخاب می‌کند.

[25] Bayesian information criterion
[26] overfitting
[27] penalty term
[28] local minimum

۵.۴.۲ آزمون نسبت درستنمایی

آزمون نسبت درست‌نمایی یکی از روش‌های آزمون فرض آماری است که بین درستی یک فرضیه و متمم آن تصمیم می‌گیرد. به بیان دقیق‌تر، برای آزمودن فرضیه $H_0 : \theta \in \Theta_0$ در برابر متمم آن.

$$\lambda(\mathbf{x}) = \frac{\sup_{\theta\in\Theta_0} \mathcal{L}(\theta|\mathbf{x})}{\sup_{\theta\in\Theta_0^c} \mathcal{L}(\theta|\mathbf{x})}$$

که در آن $\mathcal{L}(\theta|\mathbf{x})$ درستنمایی داده است. یک آزمون نسبت درست‌نمایی، آزمونی است که فرض H_0 را در صورت $\lambda(\mathbf{x}) \leq c$ رد می‌کند.

در آمار، آزمون نسبت احتمال، خوب بودن تناسب دو مدل آماری رقیب را بر اساس نسبت احتمالات آنها ارزیابی می‌کند، به خصوص یکی با حداکثرسازی در کل فضای پارامتر و دیگری پس از اعمال برخی محدودیت‌ها پیدا شده است. اگر محدودیت (یعنی فرضیه صفر) توسط داده‌های مشاهده شده پشتیبانی شود، دو احتمال نباید بیش از خطای نمونه‌گیری متفاوت باشد. بنابراین آزمون نسبت احتمال آزمایش می‌کند که آیا این نسبت به طور قابل توجهی با یک تفاوت دارد یا به طور معادل آیا لگاریتم طبیعی آن از صفر تفاوت معناداری دارد.

حال در مسأله‌ی مورد نظر می‌خواهیم برای هر Z داده شده آزمون کنیم که آیا توزیع آن با توزیع فرض صفر یکسان است یا خیر. برای هر Z داده شده می‌توانیم آماره‌ی آزمون نسبت درستنمایی زیر را بسازیم

$$LR(Z) = f_0(Z)/f(Z)$$

مقادیر کوچک برای $LR(Z)$ ، یعنی $LR(Z) < c$ ، شواهدی برای رد H_0 فراهم می‌کند. سطح معنی‌داری در سطح ژنوم به‌صورت زیر است، همچنین نقطه ی برش [۲۹] c طوری تعیین

[29] cut-off point

می‌شود که خطای نوع اول برابر عبارت زیر شود

$$\frac{\alpha}{n} = \int_{LR(Z)<c} f_0(z)dz \qquad (۱۰.۲)$$

فصل ۳

پیاده‌سازی

۱.۳ داده

برای اجرای روش‌ها ما از مجموعه داده‌های سرطان خون که شامل ۲۷ نمونه از لوسمی حاد لنفوئیدی [۱] (ALL) و ۱۱ نمونه از لوسمی حاد میلوئیدی [۲] (AML) می‌باشد، استفاده می‌کنیم. هدف پیدا کردن ژن‌های متفاوت بیان شده در این دو بیماری است. باتوجه به اینکه در مدل آمیخته اندازه‌ی نمونه‌ها باید عددی زوج باشد، ما به صورت تصادفی $K_1 = 26$ ALL و $K_2 = 10$ AML نمونه را انتخاب می‌کنیم. $n = 7129$ ژن در هر نمونه وجود دارد. همانگونه که پیش‌تر ذکر شد، $\alpha = 0.01$ را در نظر می‌گیریم و بعد از اعمال تصحیح بونفرونی برای مقایسه‌های چندگانه داریم

$$\alpha^* = \frac{0.01}{7129 \times 2} = 7.014 \times 10^{-7}$$

پیش‌پردازش داده‌ها برای هر نمونه به صورت زیر انجام گرفته است (اختلاف از میانه تقسیم

[1] Acute lymphoblastic leukemia

[2] Acute Myeloid Leukemia

بر اختلاف چندک اول و سوم)

$$x' = \frac{x - median(x)}{Q3 - Q1}$$

۲.۳ برازش مدل‌های آمیخته

مدل آمیخته را برای f و f_0 با تعداد مؤلفه‌های 1 تا 3 برازش دادیم که نتایج آن در جدول ۱.۳ است. مبتنی بر معیار بیزی‌ـشوارتز مدل دو مؤلفه‌ای بهترین مدل است. مدل‌های برازش داده شده به صورت زیر هستند

g	1	2	3
f_0	21656.7	21585.4	21598.2
f	28986.1	28833.3	28857.1

جدول ۱.۳: مقادیر معیار بیزی‌ـشوارتز برای مدل آمیخته

$$f_0(z) = 0.479\mathcal{N}(z; -0.746, 0.697) + 0.521\mathcal{N}(z; 0.739, 0.641)$$

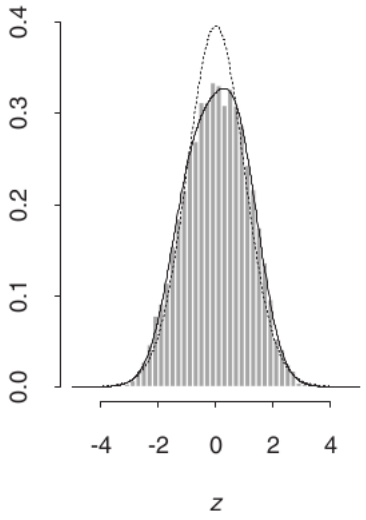
$$f(z) = 0.518\mathcal{N}(z; -0.318, 1.803) + 0.482\mathcal{N}(z; 0.781, 4.501)$$

در شکل‌های ۱.۳ و ۲.۳ بافت‌نگار [3] و مدل‌های برازش داده شده را مشاهده می‌کنید. در شکل۱.۳ می‌توان دید دم‌های تابع چگالی توزیع t سنگین‌تر از f_0 برآورد شده است. و در شکل ۲.۳ هر دو تابع f و f_0 قابل مشاهده هستند و می‌توان دید که f در مقایسه با f_0 دارای دم‌های سنگین‌تری است.

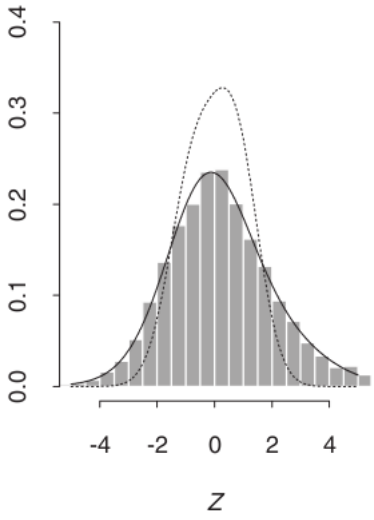
تابع نسبت درستنمایی در شکل ۳.۳ ترسیم شده است. با استفاده از روش دوبخشی [4] نقطه‌ی

[3] histogram

[4] Bisection Method



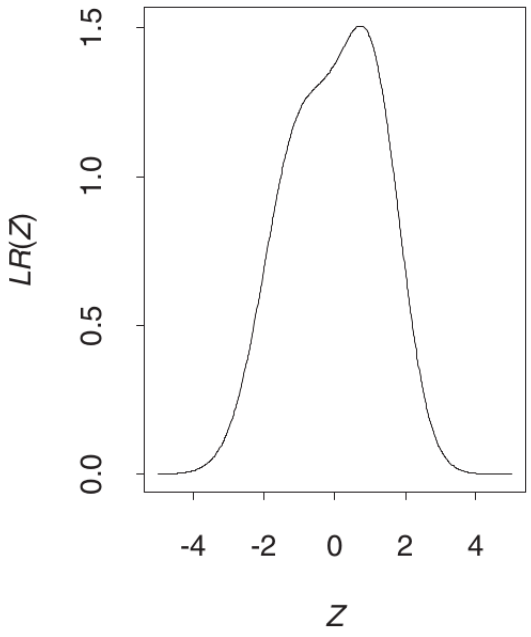
شکل ۱.۳: بافت‌نگار و مدل برازش داده شده‌ی آمیخته برای z_j (خط توپر) و توزیع t با ۳۴ درجه آزادی (نقطه‌چین)



شکل ۲.۳: بافت‌نگار و مدل برازش داده شده‌ی آمیخته برای Z_j (خط توپر) و f_0 برازش داده شده (نقطه‌چین)

$c = 0.0003437$ را بدست می‌آوریم و ناحیه‌ی رد برای فرض صفر به صورت زیر است

$$\{Z : Z < -4.8877 \ \ or \ \ Z > 4.4019\}$$



شکل ۳.۳: تابع نسبت درستنمایی

۳.۳ تعداد ژن‌ها با بیان متفاوت

آماره‌های Z_j و Z'_j به سادگی محاسبه می‌شوند. اما ناحیه‌ی رد فرض صفر سه روش متفاوت هستند.

در روش آزمون t چون درجه‌ی آزادی با تغییر j عوض می‌شود، ناحیه‌ی رد

$$\{Z : |Z| > t(\alpha^*, d_j)\}$$

می‌باشد که $t(\alpha^*, d_j)$، α^*ـصدک بالایی برای یک توزیع t با درجه آزادی d_j است.

باتوجه به توزیع n دامنه‌ی d_j از ۹ تا ۳۴ با میانگین ۱۷.۸ و میانه‌ی ۱۹.۶ محاسبه می‌شود. هرچقدر d_j کوچکتر باشد، مقدار $t(\alpha^*, d_j)$ بزرگتر است. برای محاسبه‌ی ناحیه‌ی رد محافظه‌کارانه عمل می‌کنیم و $d_j = 34$ را انتخاب می‌کنیم که در نهایت داریم

$$\{Z : |Z| > 5.8369\}$$

در مدل رگرسیونی، با توجه به اینکه فرض کردیم توزیع صفر نرمال استاندارد است، ناحیه‌ی رد α^*ـصدک بالایی برای توزیع نرمال استاندارد است.

و همانطور که در قسمت قبل گفتیم، ناحیه‌ی رد برای مدل آمیخته به صورت زیر است

$$\{Z : Z < -4.8877 \ \ or \ \ Z > 4.4019\}$$

با مقایسه‌ی ناحیه‌های رد سه مدل و f_0 برآورد شده و توزیع t در شکل ۱.۳ متوجه می‌شویم که روش آزمون t خیلی محافظه‌کارانه است. این روش ۲۰ ژن را که سطح بیان‌شان به طرز معنی‌داری تغییر کرده است را تشخیص می‌دهد، در حالی که روش رگرسیونی ۱۵۷ و روش مدل آمیخته ۱۸۷ ژن را پیدا کردند.

فصل ۴

نتیجه‌گیری

در گردایه‌ی پیش‌رو به بررسی کامل سه روش آزمون t، مدل رگرسیونی و یک مدل آمیخته پرداختیم. اکنون به ارائه یک جمع‌بندی درباره این سه روش می‌پردازیم.

آماره‌ی Z'_j که در روش رگرسیونی استفاده می‌شود، مشابه آماره‌ی آزمون t در روش مدل آمیخته است. از این رو سه روش، در قسمت آماره، معمولا نتایج مشابهی دارند. اما در قسمت تعیین سطح معنی‌داری و ناحیه‌ی رد با هم متفاوت هستند. در دو روش اول ما فرضیات پارامتری قوی روی توزیع آماره‌ها داریم. در روش t فرض نرمال بودن خطاهای تصادفی و توزیع t آماره‌ی Z_j، و در روش رگرسیونی فرض قوی نرمال بودن توزیع Z'_j. این فرضیات پارامتری در نمونه‌های کوچک ممکن است نقض شود، در نتیجه این دو روش مجانباً معتبر [۱] هستند (نمونه با اندازه‌ی بزرگ). در مقابل، توزیع صفر در روش مدل آمیخته به‌صورت مستقیم برآورد می‌شود. این روش از وجود نمونه‌های چندگانه [۲] برای ساخت z_j استفاده می‌کند، در نتیجه تعداد ژن‌های بیشتر، آن را برای برآورد f_0 به‌صورت ناپارامتری بهبود می‌بخشند. توجه‌کنید که f_0 توزیع خطاهای تصادفی است نه سطح بیان ژن. البته در روش مدل آمیخته هم فرضیاتی وجود دارد فرض در این مورد که خطاهای

[1] asymptotically valid

[2] multiple samples

تصادفی دارای توزیع متقارن هستند، و بعد از یک استانداردسازی مناسب (در اینجا تقسیم بر واریانس نمونه‌ای) خطاهای تصادفی همه‌ی ژن‌ها، یک توزیع مشترک دارند. البته این فرضیات ضعیف و منطقی هستند، یا به طور کلی ضعیف‌تر از فرض نرمال بودن در روش t هستند.

توجه داشته باشید که فرض نرمال بودن در روش t تنها برای K_1 و K_2 کوچک وجود دارد. اگر این دو مقدار به بی‌نهایت میل کنند، d_j هم به بی‌نهایت میل می‌کند ، در نتیجه توزیع صفر مورد نظر نرمال استاندارد می‌شود. از این رو اگر اندازه‌ی K_1 و K_2 بزرگتر از ۳۰ شود، می‌توان توزیع صفر را همان نرمال استاندارد در نظرگرفت. همچنین فرض نرمال تقریبی توزیع صفر Z'_j در مدل رگرسیونی هم نیاز به K_1 و K_2 بزرگ دارد. در نتیجه اگر این دو مقدار بزرگ باشند، این دو روش ناپارامتری خواهند بود.

واژه‌نامه انگلیسی به فارسی

رویکرد ..... approach

مجانباً ..... asymptotically

دوبخشی ..... bisection

قطری‌ـ بلوکی ..... block-diagonal

سلول ..... cell

ضریب ..... coefficient

برش ..... cut-off

برآوردگر ..... estimator

فراوانی‌گرا ..... frequentist

ژنگان‌شناسی ..... genomics

بافت‌نگار ..... histogram

تکراری ..... iterative

پنهان ..... latent

درستنمایی ..... likelihood

محلی ..... local

حاشیه‌ای‌سازی ..... marginalization

ریزآرایه ..... microarray

آمیخته ..... mixture

چندگانه ..... multiple

نوفه ..... noise

بیش‌برازشی ..... overfitting

تاوان ..... penalty

رگرسیون ..... regression

استوار ..... robust

نمونه ..... sample

جمله ..... term

معتبر ..... valid

واریانس ..... variance

توان دوم موزون ..... weighted least square

کتاب‌نامه

[1] Bradley Efron, Robert Tibshirani, John D Storey and Virginia Tusher (2001) Empirical Bayes Analysis of a Microarray Experiment,*Journal of the American Statistical Association*, 96:456, 1151-1160, DOI: 10.1198/016214501753382129

[2] Chen Y, Dougherty ER, Bittner ML. Ratio-based decisions and the quantitative analysis of cDNA microarray images. *J Biomed Opt.* 1997 Oct;2(4):364-74. doi: 10.1117/12.281504. PMID: 23014960.

[3] Efron,B., Tibshirani,R., Goss,V. and Chu,G. (2000) Microarrays and their use in a comparative experiment. Manuscript (available at http://www-stat.stanford.edu/ tibs/research.html).

[4] Ideker T, Thorsson V, Siegel AF, Hood LE. Testing for differentially-expressed genes by maximum-likelihood analysis of microarray data. *J Comput Biol.* 2000;7(6):805-17. doi: 10.1089/10665270050514945. PMID: 11382363.

[5] Newton MA, Kendziorski CM, Richmond CS, Blattner FR, Tsui KW. On differential variability of expression ratios: improving statistical inference about gene expression changes from microarray data.*J Comput Biol*. 2001; 8(1):37-52. doi: 10.1089/106652701300099074. PMID: 11339905.

[6] Pan W. A comparative review of statistical methods for discovering differentially expressed genes in replicated microarray experiments. *Bioinformatics*. 2002 Apr;18(4):546-54. doi: 10.1093/bioinformatics/18.4.546. PMID: 12016052.

[7] Pan W, Lin J, Le CT. A mixture model approach to detecting differentially expressed genes with microarray data. *Funct Integr Genomics*. 2003 Jul;3(3):117-24. doi: 10.1007/s10142-003-0085-7. Epub 2003 Jul 1. PMID: 12844246.

[8] Pan W, Lin J, Le CT. How many replicates of arrays are required to detect gene expression changes in microarray experiments? A mixture model approach. *Genome Biol*. 2002;3(5):research0022. doi: 10.1186/gb-2002-3-5-research0022. Epub 2002 Apr 22. PMID: 12049663; PMCID: PMC115224.

[9] Thomas JG, Olson JM, Tapscott SJ, Zhao LP. An efficient and robust statistical modeling approach to discover differentially expressed genes using genomic expression profiles. *Genome Res*. 2001

Jul;11(7):1227-36. doi: 10.1101/gr.165101. PMID: 11435405; PMCID: PMC311075.

Abstract

A common task in analyzing microarray data is to determine which genes are differentially expressed across two kinds of tissue samples or samples obtained under two experimental conditions. Recently several statistical methods have been proposed to accomplish this goal when there are replicated samples under each condition. However, it may not be clear how these methods compare with each other. Our main goal here is to compare three methods, the t-test, a regression modeling approach [3] and a mixture model approach [4] with particular attention to their different model- ing assumptions.

It is pointed out that all three methods are based on using the two-sample t-statistic or its minor variation, but they differ in how to associate a statistical significance level to the corresponding statistic, leading to the possibly large difference in the resulting significance levels and the numbers of genes detected. Using the leukemia data [5], we illustrate these points.

[3]Thomas et al., Genome Res., 11, 1227–1236, 2001

[4]Pan et al., http://www.biostat.umn.edu/cgi-bin/rrs?print+2001, 2001a,b

[5]Golub et al. (Science, 285, 531–537, 1999)

College of Science

School of Mathematics, Statistics, and Computer Science

statistical methods for discovering differentially expressed genes

Marziye Khakbaz

Supervisor:

Dr. Samaneh Eftekhari Mahabadi

A thesis submitted in partial fulfillment of the requirements for the degree of B.Sc. in Applied Mathematics

2023