和沈老师讨论的进一步结果

数 33 赵丰

January 17, 2017

1 仿真结果

采用前面所述的方法进行仿真，只考虑单位点的情形，每个位点采四个数据，01 序列由之前所述的 Markov 链生成，将贝叶斯方法与传统的 t 检验进行对比，根据仿真实验的参数，贝叶斯方法最小错误率为 $2P(Z < -1) \approx 0.32$ 对长为 1000 的链，由于其出现 01 的概率相等，所以用贝叶斯方法的错误率实际上只有理论值的一半，为 0.159 左右。传统的 t 检验由于只是控制第一类错误的概率，没有考虑到 $\theta \neq 0$ 时 $\bar{x}$ 具体的分布，因此没有充分利用 θ 的先验的非 0 即 1 的信息，在样本值为 1000 的情况下，通过调整显著性水平 α 的值，总的错误率只能降低到 27% 左右。相比较而言，贝叶斯统计更有优势。

2 参数估计

虽然 01 序列的各个位点有很强的相关性，但一旦 01 序列给定，不同位点的观测值彼此独立，如前节假定 $X \sim N(\mu_1, \sigma_1^2)|\theta = 0), X \sim N(\mu_2, \sigma_2^2)|\theta = 1)$. 通过对 RNA 序列的统计，可以得到单双链的比例 p, 即 $\theta \sim B(p)$。由于 RNA 序列一般较长，即使单个位点采集的样本点很少，但总共的信息很多，我们可以利用全局的信息对 $\mu_1, \mu_2, \sigma_1, \sigma_2$ 用矩估计的方法做出估计。方法如下：将所有位点的观测值 $X_1, X_2, ..X_n$ 分别求 1 到 4 阶矩，X_i 之间彼此独立，由于 n 很大，由大数定律可得:

$$\frac{\sum_{i=1}^{n} X_i^j}{n} \approx (1-p)E(X^j|\theta = 0) + pE(X^j|\theta = 1), j = 1, 2, 3, 4 \qquad (1)$$

由正态分布的密度函数可分别算出其前 4 阶原点矩，由此得如下关于 $\mu_1, \mu_2, \sigma_1, \sigma_2$ 的 4 元方程：

$$\begin{aligned}
&\frac{\sum_{i=1}^{n} X_i}{n} \approx (1-p)\mu_1 + p\mu_2 \\
&\frac{\sum_{i=2}^{n} X_i^2}{n} \approx (1-p)(\mu_1^2 + \sigma_1^2) + p(\mu_2^2 + \sigma_2^2) \\
&\frac{\sum_{i=2}^{n} X_i^3}{n} \approx (1-p)(\mu_1^3 + 3\sigma_1^2\mu_1) + p(\mu_2^3 + 3\sigma_2^2\mu_2) \\
&\frac{\sum_{i=2}^{n} X_i^4}{n} \approx (1-p)(\mu_1^4 + 6\sigma_1^2\mu_1^2 + 3\sigma_1^4) + p(\mu_2^4 + 6\sigma_2^2\mu_2^2 + 3\sigma_2^4)
\end{aligned} \tag{2}$$

根据原始数据可以求出上述方程组左边的值，由此解出待估计参数。

3 参数估计数值实验

使用李盼提供的数据（只用 RT) 读取某实验条件下实验组 2 组，对照组两组，每组数据长度均为 1870。先将对照组数据取平均值，再用实验组除以取平均值后的对照组，R 代码如下：

```
x1<-scan('cy_D1.rt',what=numeric(0),n=1e6)
x2<-scan('cy_D2.rt',what=numeric(0),n=1e6)
x3<-scan('cy_N1.rt',what=numeric(0),n=1e6)
x4<-scan('cy_N2.rt',what=numeric(0),n=1e6)
x_case_1<-2*x3/(x1+x2);
x_case_2<-2*x4/(x1+x2);
#cor(x1,x2)=0.93
#cor(x3,x4)=0.99
```

进一步计算两组 case 的统计信息得下表：从上表可以看出两组数据相

参数	x_case_1	x_case_2
最小值	0.083	0.108
最大值	5.882	5.457
一阶原点矩	1.729	1.807
二阶原点矩	4.106	4.377
三阶原点矩	11.963	12.681
四阶原点矩	40.513	41.610

差不大，可以用将两组数据合并用来估计四个参数。有 icshape 实验值按 0.5 的阈值二值化得 01 序列，计算出单链比例为 0.53，近似取 p=0.5.

代入已知数据，求解上述非线性方程组得：

$$\mu_1 = 1.165, \mu_2 = 2.371, \sigma_1^2 = 0.269, \sigma_2^2 = 1.236$$

下面利用单位点的两组数据根据前面的贝叶斯统计模型做判决，R 代码如下：

```
mu_1=2.3711
mu_2=1.1649
sigma_1=1.2362
sigma_2=0.2686
n=2
my_y=c()
for(i in 1:length(x_case_1)){
xbar=(x_case_1[i]+x_case_2[i])/2;
p1=log(sigma_1)+n*(xbar-mu_1)^2/sigma_1^2;
p2=log(sigma_2)+n*(xbar-mu_2)^2/sigma_2^2;
if(p1>p2){
my_y=c(my_y,1);
}
else{my_y=c(my_y,0);}#0 is single-chain
}
```

4 结果比较

使用上述贝叶斯统计的方法获得 my_y01 序列，其与标准结构的差别率为 51.9%, 而使用 icshape 的方法获得 y 序列，其与标准结构的差别率为 36%. 比较 Bayes 方法和 icshape 的方法结果，差别率为 51%. 但是比较 x_i_case 和结果的相关性，icshape 的方法只有-0.19(RT 值越大，更倾向 $\theta = 0$, 即 y[i]=0 所以是负相关) 而 Bayes method 达到-0.34，可见 Bayes method 更优。主要问题是 4 组数据变 2 组用直接相除法可能不太理想。如果在原来实验组-α 乘以对照组的基础上用 Bayes method 效果可能比较好。

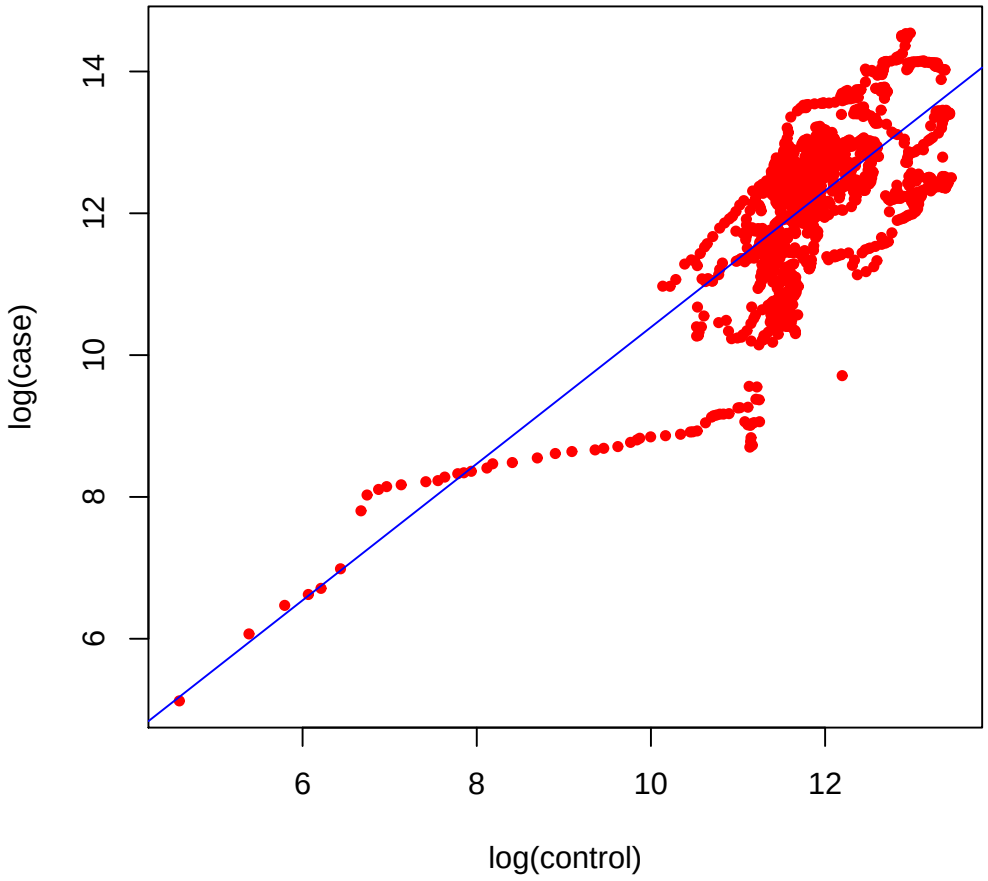
如果尝试用 $\log(case) = A * \log(control) + B + signal$ 对模型进行拟合：则 R 代码如下：

```
control<-(x1+x2)/2;
plot(x=log(control),y=log(x3),pch=20,col='red',
ylab='log(case)')
abline(lm(log(x3)~log(control)),col='blue')
my_coff=lm(log(x3)~log(control))
```

```
intercept=my_coff$coefficients[1]
slope=my_coff$coefficients[2]
signal_value_1=log(x3)-(intercept+slope*log(control))
signal_value_2=log(x3)-(intercept+slope*log(control))
```

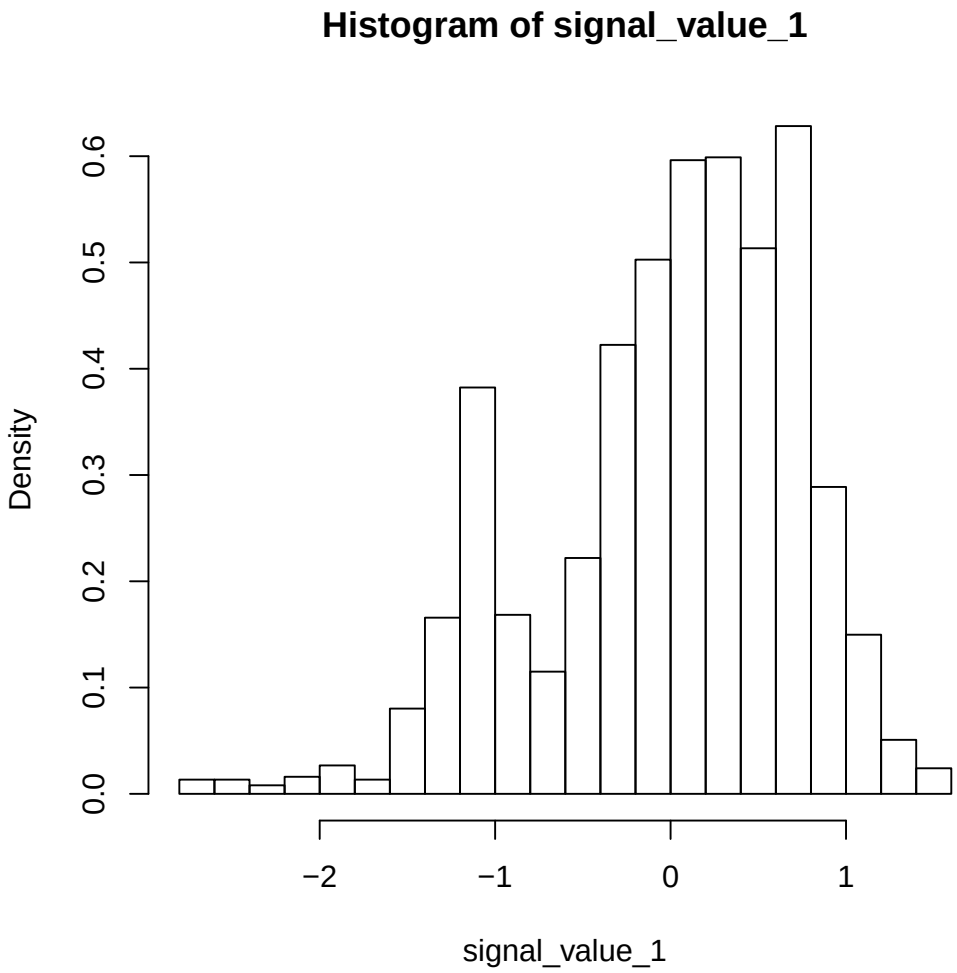
上面做出的 $\log(case) \sim \log(control)$ 如下所示

Figure 1: 线性回归拟合



对 $signal_value_1$ 做直方图如下：

Figure 2: $signal_value_1$ 的直方图



由上图可以看出，$signal_value_1$ 有两个峰，可以近似认为是两组正态样本混叠在一起。使用上述得到 signal 的方法用贝叶斯统计估计原 01 序列，得 my_y,icshape 给出的 01 序列与 signal_value_1 的相关性只有-0.23, 而用用贝叶斯统计的方法结果与 signal_value_1 的相关性可达-0.45. 两种方法差别率为 43%。

5 参考文献

References

[1] 生物信息学