D^{3}M: Detection of differential distributions of methylation patterns

Yusuke Matsui, Masahiro Mizuta, Satoru Miyano and Teppei Shimamura

1 Simulation model

Here, we describe the details of the simulation models used in section 3. We generate the data using two types of distribution considering cases 1-8 in Table 1. The control and case groups are represented by normal and normal mixture distributions, respectively. In each case, there are 300 samples; 160 and 140 for case and control groups, respectively. We start with the normal mixture distributions and then we generate the normal distributions. According to section 3, we place distribution functions of control and case groups at site s_i as $F_i(x)$ and $G_i(y)$ respectively, and we put the mean and standard deviation of $F_i(x)$ ($G_i(y)$) as $\mu_{i,1}$ and $\sigma_{i,1}$ ($\mu_{i,2}$ and $\sigma_{i,2}$).

The mixture distributions in the case groups are as follows;

- The number of mixture components: two
- Mixture proportions: 90% and 10%

The formula for the moments of two component mixture distributions is;

$$\mu_i = 0.9\mu_{i,1} + 0.1\mu_{i,2} \tag{1}$$

and that for variance is

$$\sigma_i^2 = 0.9(\mu_i^{(1)} - \mu_i)^2 + 0.1(\mu_i^{(2)} - \mu_i)^2 + (\sigma_i^{(1)})^2 + (\sigma_i^{(2)})^2 \tag{2}$$

where $\mu_i^{(1)}$ (respectively $\mu_i^{(2)}$) and $\sigma_i^{(1)}$ (respectively $\sigma_i^{(2)}$) represent the mean and standard deviation of the 1^{st} (2^{nd}) component of the mixture distribution, respectively. In the simulation, we fix $\mu_i = 5$ and obtain $\mu_i^{(2)}$ and $\sigma_i^{(2)}$ from (1) and (2) after generating $\mu_i^{(1)}$ and $\sigma_i^{(1)}$.

Next, we generate the distributions of control group. Here, we regard the significant difference between means and between variances as $|\mu_{i,1} - \mu_{i,2}| \geq$

1.96 and $|\sigma_{i,1}-\sigma_{i,2}| \geq 1.96$, respectively. In cases 1-4 of Table 2 where $\mu_{i,1} \neq \mu_{i,2}$, we set the mean by $\mu_{i,2} = \mu_{i,1} - 1.96$, which satisfies the inequalities of the significant difference. In addition to this, we add the noise to $\mu_{i,2}$ (respectively $\sigma_{i,2}$) using a truncate normal distribution, whose parameters are lower bound a_i, upper bound b_i, mean $\mu_i^{(\mathrm{TN})}$, and variance $(\sigma_i^{(\mathrm{TN})})^2$, using R package `truncnorm`. We set $\mu_i^{(\mathrm{TN})} = \mu_{i,2}$ and $\sigma_i^{(\mathrm{TN})} = 0.1$. In particular, in cases 3, 4, 7 and 8 where $\sigma_1 \neq \sigma_2$, since the variance needs to be positive, we set $a_i = 0$ and $b_i = \sigma_{i,1} - 1.96$ if $\sigma_{i,1} - 1.96 > 0$, and otherwise $b_i = 0.1$. We generate 5,000 datasets ($i = 1, 2, \ldots, 5000$) in each case and the resulting summary statistics are shown in Table 1. Finally, we normalized the each dataset to be included in [0,1].

Table 1: Summary statistics of generated 5,000 data in each case

	$\mathrm{E}[\mu_1 - \mu_2]$	$\sqrt{\mathrm{Var}[\mu_1 - \mu_2]}$	$\mathrm{E}[\sigma_1 - \sigma_2]$	$\sqrt{\mathrm{Var}[\sigma_1 - \sigma_2]}$
case1	0.000	0.006	0.001	0.004
case2	0.293	0.135	0.080	0.098
case3	-0.001	0.035	2.926	0.060
case4	3.696	0.117	-0.001	0.033
case5	0.293	0.012	14.962	0.010
case6	5.301	0.130	0.077	0.095
case7	3.687	0.118	2.926	0.062
case8	5.295	0.012	14.962	0.010