change the normal distribution

2020-01-20

Previously, we looked at coefficients of the lme model: $z = (\beta + \Gamma(\alpha' x) + b) \sim MVN(\beta + \Gamma(\alpha' x), D)$, and used these normal distributions to calculate the purity.

We may also treat the outcome Y as normal distributions and fit $Y = X(\beta + \Gamma(\alpha' x)) + Zb \sim MVN(X(\beta + \Gamma(\alpha' x)), ZDZ')$ to calculate the purity.

However, this method has some problem, since in our example, the covariance matrix ZDZ' is non-inversable.

The true D matrix is

	(Intercept)	tt
(Intercept)	8.868634	2.766242
tt	2.766242	1.015725

The Z matrix is

```
##      [,1] [,2]
## [1,]    1    1
## [2,]    1    2
## [3,]    1    3
## [4,]    1    4
## [5,]    1    5
## [6,]    1    6
## [7,]    1    7
```

Then

```
z %*% d1 %*% t(z)
```

```
##          [,1]     [,2]     [,3]     [,4]     [,5]     [,6]     [,7]
## [1,] 15.41684 19.19881 22.98078 26.76274 30.54471 34.32668 38.10864
## [2,] 19.19881 23.99650 28.79419 33.59188 38.38958 43.18727 47.98496
## [3,] 22.98078 28.79419 34.60761 40.42102 46.23444 52.04786 57.86127
## [4,] 26.76274 33.59188 40.42102 47.25017 54.07931 60.90845 67.73759
## [5,] 30.54471 38.38958 46.23444 54.07931 61.92417 69.76904 77.61391
## [6,] 34.32668 43.18727 52.04786 60.90845 69.76904 78.62963 87.49022
## [7,] 38.10864 47.98496 57.86127 67.73759 77.61391 87.49022 97.36654
```

```
eigen(z %*% d1 %*% t(z))$values
```

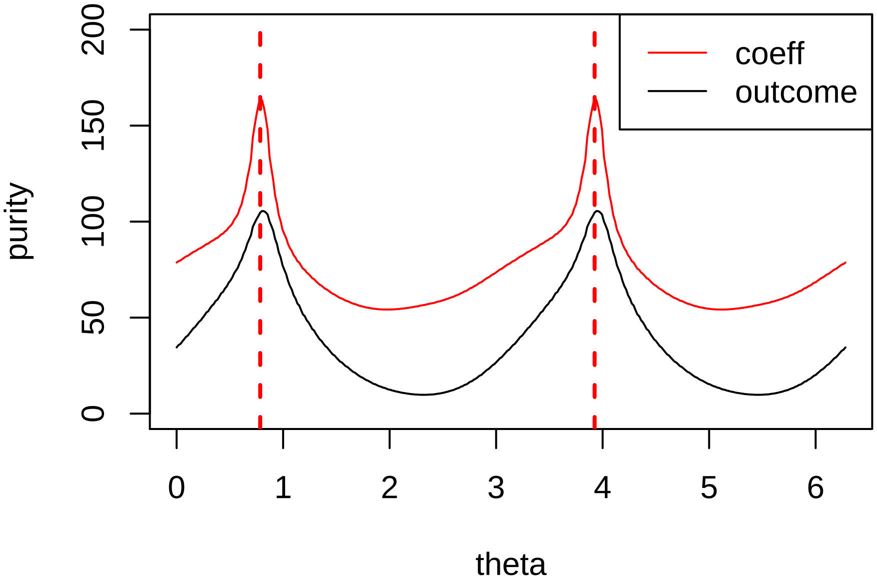
```
## [1]  3.584500e+02  7.414585e-01  1.554740e-14  7.479788e-16 -3.154540e-15
## [6] -5.836469e-15 -6.748817e-15
```

it is not inversable. I then calculate the following simulations by using its generalized inverse.

Simulation with one dataset

simulate one data set, check whether the two method can return the same estimated α.

set the true $\alpha = [\cos(\frac{\pi}{4}), \sin(\frac{\pi}{4})]$



The true θ value is $\frac{\pi}{4} \approx 0.785$. The estimated θ that maximizes the purity by fiting the coefficients as normal distributions:

```
## [1] 0.7853982
```

The estimated θ that maximizes the purity by fiting the outcome as normal distributions:

```
## [1] 0.8028514
```

Simulation with 1000 repetitions

Estimate the coefficient as $z = (\beta + \Gamma(\alpha' x) + b) \sim MVN(\beta + \Gamma(\alpha' x), D)$ and then calculate the purity

theta	purity	purity_sd	coverage	cossim	cos_sd	theta_est	theta_sd	theta_cov
0	1.505	0.533	0.910	0.965	0.075	0.800	0.253	0.955
60	261.166	51.682	0.925	0.999	0.001	0.785	0.032	0.965
120	201.892	41.082	0.930	0.999	0.001	0.785	0.034	0.945
180	21.870	3.551	0.920	0.997	0.004	0.787	0.072	0.935

The columns contain:

- truekl: the true purity
- Purity: the mean estimated purity
- sdp: the standard deviation of the estimated purity
- coverage: how many times the true purity is contained in the estimated confidence interval
- theta_est: the estimated θ, where $\alpha = [\cos(\theta), \sin(\theta)]$
- cossim: the cosine similarity of the estimated $\hat{\alpha}$ and α

Estimate the outcomes as $Y = X(\beta + \Gamma(\alpha' x)) + Zb \sim MVN(X(\beta + \Gamma(\alpha' x)), ZDZ')$ and then calculate the purity

theta	purity	purity_sd	coverage	cossim	cos_sd	theta_est	theta_sd	theta_cov
0	0.226	0.414	0.925	0.974	0.038	0.793	0.231	0.957

theta	purity	purity_sd	coverage	cossim	cos_sd	theta_est	theta_sd	theta_cov
60	13.675	2.761	0.023	0.988	0.051	0.787	0.144	0.965
120	7.474	1.728	0.068	0.946	0.146	0.810	0.285	0.935
180	4.124	1.120	0.130	0.991	0.012	0.790	0.136	0.958

The columns contain:

- truekl: the true purity
- Purity: the mean estimated purity
- sdp: the standard deviation of the estimated purity
- coverage: how many times the true purity is contained in the estimated confidence interval
- theta_est: the estimated θ, where $\alpha = [\cos(\theta), \sin(\theta)]$
- cossim: the cosine similarity of the estimated $\hat{\alpha}$ and α

By using this method, the standard deviation of estimated puirty get much smaller. But the cosine similarity get worse than using our pervious method.