semi_simulation

```
library(MASS)
library(Iso)
```

```
## Iso 0.0-17
```

```
library(mvtnorm)
```

Model overview

Basic model

Recently Wang and Zhou (2018) proposed a parametric transformation model for biomarker data, of the form

$$f(\mathbf{t}|\mathbf{x},\mathbf{z}) = \sum_{d=0}^{L-1} \pi_d(\mathbf{z}) \prod_{k=1}^{K} J(t_k,\lambda_k) g(H_k(t_k,\lambda_k|\mathbf{x}),$$

where $g(\cdot|\mathbf{x})$ is a known parametric density, $H(\cdot,\cdot)$ is a given transformation, $J(t_k,\lambda_k)$ is the transformation Jacobian, and $\pi_d(\mathbf{z})$ is the mixture proportion, which is specified as a logit form. In their model, the joint density is independent given the covariate $\mathbf{x}$

Our extension

Here our model is more general, the joint density is not independent, and the monotone transformations $H_k(\cdot)$'s are also unknown to be estimated The observed data is $D_n = \{(\mathbf{y}_i,\mathbf{x}_i)\} : i = 1,...,n\}$ from n subjects. $\mathbf{y}_i = (y_{i1},...,y_{ik})^T$ is the observation of the i-th subject with k biomarkers, $\mathbf{x}_i = (x_{i1},...,x_{iq})$ is the corresponding covariates, we assume the $(\mathbf{y}_i,\mathbf{x}_i)$'s are iid.

The goal is to classify each subject to one of the two groups, normal and disabled. For this, we first need to specify a model, then estimate the model parameters and finally, classify the subjects.

The normal model is easy to use but not robust to model assumption. For robustness we propose the following semi-parametric transformation model for the observed data. First, for each margin, we specify the transformation

$$g_j(y_{ij}) = \mathbf{x}_i^T\beta_j + \epsilon_{ij} \qquad (i = 1,2,...,n; j = 1,...,k) (1)$$

where $g_j(\cdot) \in \mathcal{G}$, $\mathcal{G}$ is the collection of bounded monotone non-decreasing functions, and ϵ_{ij} is the noise. We assume $\epsilon_i := (\epsilon_{i1},...,\epsilon_{ik})^T \sim N(\mathbf{0},\Omega)$, $\Omega = (\omega_{ij})_{k\times k}$. Denote $\mathbf{g}(\mathbf{y}_i) = \big(g_1(y_{i1}),...,g_k(y_{ik})\big)^T$ and $\mathbf{x_i} = (x_{i1},...,x_{iq})^T$, $\mathbf{beta}_j = (\beta_{j1},...,\beta_{jq})^T$ and $\mathbf{B}^T = (\beta_1,...,\beta_k)$. Let $\phi(\cdot|\Omega)$ be the density function of the $N(\mathbf{0},\Omega)$ distribution. Next, since the group status of each subject is unknown, the joint model is specified as a mixture

$$f(\mathbf{y}_i|\mathbf{x}_i) = \sum_{d=0}^{1} \pi_d(\mathbf{x}_i)\phi(\mathbf{g}(\mathbf{y}_i) - \mathbf{x}_i^T\mathbf{B}|\Omega) (2)$$

where

$$\pi_0(\mathbf{x}_i|\eta) = 1 - \pi_1(\mathbf{x_i}|\eta) = \frac{\exp(\mathbf{x}_i^T\eta)}{1+\exp(\mathbf{x}_i^T\eta)}$$

is the mixing proportions. $\pi_d(\mathbf{x}_i|\eta)$ is the probability subject i belongs to group d.

Simulation settings:

- Assumption1: joint density is not independent, and $H_k(\cdot) = I_k(\cdot)$, $k = 2$ (2 biomarkers) with $q = 2$ (2 covariates).

We need to estimate $\eta = \{\eta_1, \eta_2, ..., \eta_q\}$, $\theta = \{\eta_{jd},\ \omega_{jr} : j, r = 1, ..., k; d = 0, 1\}$ and $\mathbf{g}(\cdot) = (g_1(\cdot), ..., g_k(\cdot))$. Note that since the ω_{jr}'s are symmetric, we only need to estimate ω_{jr} for $j = 1, ..., k; r \geq j$. Let δ_i be the latent group indicator of the i-th subject, $\delta_i = 0$ if this subject belongs to group 0, and $\delta_i = 1$ otherwise. Let $D_{n,c} = \{(\mathbf{y}_i, x_i, \delta_i) : i = 1, ..., n\}$ be the "complete" data. The log-likelihood of the complete data is

$$\ell(\eta, \theta, \mathbf{g}|D_{n,c}) = \sum_{i=1}^{n}\sum_{d=0}^{1}\Big[I_d(C_i)\Big(\log \pi_d(\mathbf{x}_i|\eta) - \frac{1}{2}(\mathbf{g}(\mathbf{y}_i) - \mathbf{x}_i^T\mathbf{B})^T\Omega^{-1}(\mathbf{g}(\mathbf{y}_i) - \mathbf{x}_i^T\mathbf{B})\Big)$$

$$-\frac{n}{2}\log|\Omega|. \quad (3)$$

We estimate $(\eta, \theta, \mathbf{g})$ by

$$(\hat{\eta}, \hat{\theta}, \hat{\mathbf{g}}) = \arg\max_{(\eta,\theta,\mathbf{g})\in(\mathbf{B},\mathbf{\Theta},\mathcal{G}^k)} \ell(\eta, \theta, \mathbf{g}|D_{n,c}). \quad (4)$$

- Simulation function

```
no_gold_standard_simu<- function(eta.r, theta.r, b.0.r, b.1.r, o.r, D){
  iter = 1
  res = c()
  res[iter] = 1000
  while(res[iter]>0.01){
  # print(iter)
  par_old = c(eta.r, theta.r)
  # print(iter)
  # if(iter == 1){
    pi1.r = exp(as.matrix(D$X)%*%eta.r)/(1+exp(as.matrix(D$X)%*%eta.r))
    pi0.r = 1-pi1.r

    p0.r = dmvnorm(g.r-as.matrix(D$X)%*%b.0.r,mean = c(0,0), sigma = o.r)
    p1.r = dmvnorm(g.r-as.matrix(D$X)%*%b.1.r,mean = c(0,0), sigma = o.r)
    P.r = (pi1.r*p1.r)/(pi1.r*p1.r+pi0.r*p0.r)
    # res_eta = mean(abs(D$d_prob - P.r))
  # }


  # label.new = ifelse(P.r>1-P.r,1,0)
  # mean.0 = apply((g.r-as.matrix(D$X))[label.new==0,],2,mean)
  # mean.1 = apply((g.r-as.matrix(D$X))[label.new==1,],2,mean)
  #
  # P.r.cat = as.factor(ifelse(P.r>0.5,1,0))
    eta.r<- glm(P.r~.-1, data = D$X,family = binomial)$coef

  # sprintf("Estimated eta: %s", eta.r)
  # sprintf("True Estimated eta: %s", eta)

  mytheta = function(par){
      o = matrix(0,2,2)
      o[1,1]<-par[1]
      o[2,2]<-par[2]
      o[1,2] <- par[3]*prod(sqrt(diag(o)))
      o[2,1] <- o[1,2]
```

```
    beta1 <- par[4:5]
    beta2 <- par[6:7]
    b0 <- matrix(c(beta1,beta2),2)

    beta1 <- par[8:9]
    beta2 <- par[10:11]
    b1 <- matrix(c(beta1,beta2),2)

    o.det = prod(par[1],par[2])-o[1,2]^2
    o.det = det(o)
    o.inv = matrix(0,2,2)
    o.inv[1,1]<-par[2]/o.det
    o.inv[2,2]<-par[1]/o.det
    o.inv[1,2] <- -o[1,2]/o.det
    o.inv[2,1] <- -o[2,1]/o.det
    # o.inv = solve(o)
    # sum(c(log(P.r))*(-0.5)*c(apply(as.matrix(g.r)-as.matrix(D$X)%*%b1,1,function(x) t(as.matrix(x))
    #     c(1-log(P.r))*(-0.5)*c(pply(as.matrix(g.r)-as.matrix(D$X)%*%b0,1,function(x) t(as.matrix(x)
    #     -N/2*log(o.det)
    sum(
      P.r*-0.5*c(apply(as.matrix(g.r)-as.matrix(D$X)%*%b1,1,function(x) t(as.matrix(x))%*%o.inv%*%as.m
        ,
        (1-P.r)*-0.5*c(apply(as.matrix(g.r)-as.matrix(D$X)%*%b0,1,function(x) t(as.matrix(x))%*%o.inv
       ,-N/2*log(o.det))
}
result = try(optim(theta.r,mytheta, method = 'BFGS',
                   control= list(fnscale=-1)))


if(inherits(result, "try-error")==TRUE){
  res = res
#   result2.1 = try(pava(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.0
#                   +0.5*(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.
#                   -g.r[,2])*o.inv[1,2]/o.inv[1,1]))
#
#
#   result2.2 = try(pava(diag(as.matrix(D$X)%*%as.matrix(apply(1-P.r,1, function(x) (1-x)*as.matrix(b
# +0.5*(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.0.r[,1],nrow=2))+a
#       -g.r[,1])*o.inv[1,2]/o.inv[2,2]))
#
#   if(inherits(result2.1, "try-error")==TRUE | inherits(result2.2, "try-error")==TRUE){
#     res = res
#   }else{
#     g.r.1 <- matrix(0, nrow = 1000, ncol = 2)
#     g.r.1[,1] <- pava(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.0.
#                   +0.5*(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.
#                   -g.r[,2])*o.inv[1,2]/o.inv[1,1])
#
#
#     g.r.1[,2] <- pava(diag(as.matrix(D$X)%*%as.matrix(apply(1-P.r,1, function(x) (1-x)*as.matrix(b.
# +0.5*(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.0.r[,1],nrow=2))+a
#       -g.r[,1])*o.inv[1,2]/o.inv[2,2])
#   g.r <- g.r.1
```

```
#   }
}else{
  theta.opt = optim(theta.r,mytheta, method = 'BFGS',
                    control= list(fnscale=-1))
  theta.r <- theta.opt$par
theta.conv <- theta.opt$convergence
# print(ifelse(theta.conv==0,'success','fail'))
# sprintf("Estimated theta: %s, %s and %s ", theta.r[1],theta.r[2], theta.r[3])
# sprintf("True theta: %s, %s and %s", O[1,1],O[2,2], rho)
#
# sprintf("Estimated beta0: %s, %s, %s,%s",
#         theta.r[4],theta.r[5],theta.r[6],theta.r[7])
#
# sprintf("True beta0: %s, %s, %s,%s",
#        B.0[1],B.0[2],B.0[3],B.0[4])
#
# sprintf("Estimated beta1: %s, %s, %s,%s ",
#        theta.r[8],theta.r[9],theta.r[10],theta.r[11])
#
# sprintf("True beta1: %s, %s, %s,%s",
#        B.1[1],B.1[2],B.1[3],B.1[4])

b.0.r = matrix(theta.r[4:7],2,2)
b.1.r = matrix(theta.r[8:11],2,2)

o.r = matrix(0,2,2)
o.r[1,1]<-theta.r[1]
o.r[2,2]<-theta.r[2]
o.r[1,2] <- theta.r[3]*prod(sqrt(diag(o.r)))
o.r[2,1] <- o.r[1,2]
o.det = prod(diag(o.r))-o.r[1,2]^2
o.inv = matrix(0,2,2)
o.inv[1,1]<-o.r[2,2]/o.det
o.inv[2,2]<-o.r[1,1]/o.det
o.inv[1,2] <- -o.r[2,1]/o.det
o.inv[2,1] <- -o.r[2,1]/o.det

# g.r.1 <- matrix(0, nrow = 1000, ncol = 2)
# g.r.1[,1] <- pava(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.0.r[,1
#                   +0.5*(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.
#                   -g.r[,2])*o.inv[1,2]/o.inv[1,1])
#
# g.r.1[,2] <-  pava(diag(as.matrix(D$X)%*%as.matrix(apply(1-P.r,1, function(x) (1-x)*as.matrix(b.0.r
# +0.5*(diag(as.matrix(D$X)%*%as.matrix(apply(P.r,1, function(x) (1-x)*as.matrix(b.0.r[,1],nrow=2))+a
#       -g.r[,1])*o.inv[1,2]/o.inv[2,2])
#
# res2 = mean(abs(g.r-g.r.1))
# g.r <- g.r.1
# res[iter] = mean(abs(g.r-D$Y))
 par_new = c(eta.r, theta.r)

 # par_new = c(res_eta, theta.r)
res = append(res, max(abs(par_old-par_new)), after = length(res))
```

```
  # res = append(res,max(res2,abs(par_old-par_new)), after = length(res))
  # g.r <- g.r.1
  # print(res[iter])
  iter = iter+1
  }
}



# mean(as.matrix(D$X[P.r<0.5,])%*%B.0 - D$Y[P.r<0.5,])


print(paste("Converge in %s iteration:", iter))
print("Converge covariance matrix: ")
print(o.r)

print("Original covariance matrix: ")
print(O)
return(list(eta.r = eta.r, theta.r = theta.r))
}
```

- Basic Settings: 2 biomarkers, 2 covairates and 2 classes of diseases(0-no disease, 1-disease)

```
set.seed(123)

k <-2; #number of biomarker
m <-2; #covariates dimension
d <-2; #number of class(0-no disease 1-disease)
N<- 1000 #number of observation

## B
beta1 <- c(0.1,0.3)
beta2 <- c(0.2,0.4)
B.0 <- matrix(c(beta1,beta2),2)

beta1 <- c(-0.1,0.2)
beta2 <- c(-0.2,0.3)
B.1 <- matrix(c(beta1,beta2),2)

# B.1 = B.0

## Delta
# delta <- sample(c(0,1),N, replace = T)


## X
x1 = rnorm(N,0,1)
x2 = rnorm(N,0,2)
X = data.frame(x1 = x1, x2 = x2)

## g.y
eta<- c(1,2)
d_prob = exp(as.matrix(X)%*%eta)/(1+exp(as.matrix(X)%*%eta))
d_cat = rbinom(N,1,d_prob)
```

```
d_prob.0 = 1-d_prob[d_cat==0]
d_prob.1 = d_prob[d_cat==1]


X.0 = X[d_cat==0,]
X.1 = X[d_cat==1,]


## D
D <- list()
D$X <- rbind(X.0,X.1)
d_cat <- c(rep(0,nrow(X.0)),rep(1,nrow(X.1)))
D$d_cat <-  d_cat
D$d_prob <- c(d_prob.0, d_prob.1)
```

- Simulation1: ###Initial var1 = var2 = 0.3, and correlation = 0, without transformation

```
### True value of covariance matrix
var <- c(0.3,0.3)
O = abs(diag(var,2))
rho = 0
O[1,2] <- rho*sqrt(var[1])*sqrt(var[2])
O[2,1] <- O[1,2]

## Control covariance matrix
rho.r <- rho
var1 <- var[1]
var2 <- var[2]

## g.y
eps <- mvrnorm(n = N, c(0,0), O)
eps.0 = eps[d_cat==0]
eps.1 = eps[d_cat==1]
#
Y.0 = as.matrix(X.0)%*%B.0 + eps.0
Y.1 = as.matrix(X.1)%*%B.1 + eps.1
D$Y <- rbind(Y.0,Y.1)



### Initial values:
eta.r <- rnorm(2)
theta.r = runif(3,0,1)
o.r = abs(diag(theta.r[c(1,2)],2))
o.r[1,2] <- theta.r[3]*sqrt(theta.r[1])*sqrt(theta.r[2])
o.r[2,1] <- o.r[1,2]

# b.r <- matrix(rnorm(4),nrow = 2)
b.0.r <- matrix(rnorm(4),nrow = 2)
b.1.r <- matrix(rnorm(4),nrow = 2)
theta.r = c(theta.r, c(b.0.r), c(b.1.r))
g.r <- D$Y # g.r stays constant
```

```
simu1_result = no_gold_standard_simu(eta.r, theta.r, b.0.r, b.1.r, o.r, D)
```

```
## [1] "Converge in %s iteration: 40"
## [1] "Converge covariance matrix: "
##             [,1]        [,2]
## [1,]  0.29675290 -0.01982971
## [2,] -0.01982971  0.29664699
## [1] "Original covariance matrix: "
##      [,1] [,2]
## [1,]  0.3  0.0
## [2,]  0.0  0.3
```

- Simulation2: ###Initial var1 = 0.3, var2 = 0.4, and correlation = 0.2, without transformation

```
var <- c(0.3,0.4)
O = abs(diag(var,2))
rho = 0.2
O[1,2] <- rho*sqrt(var[1])*sqrt(var[2])
O[2,1] <- O[1,2]

## Control covariance matrix
rho.r <- rho
var1 <- var[1]
var2 <- var[2]

## g.y
eps <- mvrnorm(n = N, c(0,0), O)
eps.0 = eps[d_cat==0]
eps.1 = eps[d_cat==1]
#
Y.0 = as.matrix(X.0)%*%B.0 + eps.0
Y.1 = as.matrix(X.1)%*%B.1 + eps.1
D$Y <- rbind(Y.0,Y.1)



### Initial values:
eta.r <- rnorm(2)
theta.r = runif(3,0,1)
o.r = abs(diag(theta.r[c(1,2)],2))
o.r[1,2] <- theta.r[3]*sqrt(theta.r[1])*sqrt(theta.r[2])
o.r[2,1] <- o.r[1,2]

# b.r <- matrix(rnorm(4),nrow = 2)
b.0.r <- matrix(rnorm(4),nrow = 2)
b.1.r <- matrix(rnorm(4),nrow = 2)
theta.r = c(theta.r, c(b.0.r), c(b.1.r))
g.r <- D$Y # g.r stays constant

simu2_result = no_gold_standard_simu(eta.r, theta.r, b.0.r, b.1.r, o.r, D)
```

```
## [1] "Converge in %s iteration: 31"
## [1] "Converge covariance matrix: "
##             [,1]       [,2]
## [1,] 0.28138079 0.05491051
```

```
## [2,] 0.05491051 0.36318869
## [1] "Original covariance matrix: "
##            [,1]       [,2]
## [1,] 0.30000000 0.06928203
## [2,] 0.06928203 0.40000000
```