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

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

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

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

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

- Simulation1