Semiparametric analysis in multivariate mixture models: A biased sampling approach

Denis Leung
School of Economics and Social Sciences,
Singapore Management University
e-mail: denisleung@smu.edu.sg

Jing Qin
Biostatistics Research Branch,
National Institute of Allergy and Infectious Diseases, USA
e-mail: jingqin@niaid.nih.gov

Biased sampling

1. Convenient and economic sampling of data (James J. Heckman, 2000 Nobel Prize in Economics)

2. Cancer screening studies - patients who turn up for screening are different from those who don't - Response biased sampling.

3. Comparison of diagnostic tools - an effective tool will pick up "cases" sooner and at an ealier stage - proportion of cases proportional to time since detection - Length-biased

4. Prevalent sampling in which subject who has reached a particular disease stage are followed - Truncation sampling (Case-Control study)

5. General missing data problems

Length biased sampling (Cox 1966, Vardi 1982, 1985)

$$X \sim dF(X), \qquad \textbf{but } x_1, ... x_n \sim \textbf{ iid } G(x) = x\frac{dF(x)}{\int x dF(x)} \qquad \Rightarrow g(x) = w(x) f(x)$$

F **can be estimated by**

$$\hat{F}(x) = \frac{A(x)}{A(\infty)}, \qquad A(x) = \frac{1}{n}\sum_{i=1}^{n} \frac{1}{x_i} I(x_i \leq x)$$

Applications:

1. Cancer screening

2. Wild animal study

What happens if an additional (training) sample from F **is available**

$$y_1, ... y_m \sim \textbf{ iid } F(y)$$

How can we estimate F **by combining data from** X **and** Y**? Vardi (1982)**

Case-control study

$$P(D=1|x) = \frac{\exp(\alpha+\beta x)}{1+\exp(\alpha+\beta x)}$$

$$x_1, ..., x_{n_0} \sim \textbf{ iid } \ f(x|D=0)$$

$$x'_1, ..., x'_{n_1} \sim \textbf{ iid } \ f(x|D=1)$$

1. n_0 **and** n_1 **are fixed**

2. β **is the "odds-ratio". Anderson (1972) and Prentice and Pyke (1979) found that the prospective likelihood is valid even for the retrospective sampling problem for** β**.**

3. **Bayes Theorem gives**

$$\frac{f(x|D=1)}{f(x|D=0)} = \exp(\alpha^* + \beta x) \qquad \rightarrow \qquad f(x|D=1) = \overbrace{\exp(\alpha^* + \beta x)}^{w(x)} f(x|D=0)$$

$$\alpha^* = \alpha + \log\{P(D=0)/P(D=1)\}$$

4. **Biased sampling problem with weight function** $w(x) = \exp(\alpha^* + x\beta)$ **(Vardi, 1982, 1985; Gill, Vardi and Wellner, 1988; Gilbert, Lele and Vardi, 1999)**

Missing data problems

1. **Assume a regression model**

$$f(y|x) = f(y|x,\beta)$$

2. **Observed data:** $\{(\delta = 1, y_i, x_i)\}_{i=1}^{n_1}$ **and** $\{(\delta = 0, ?, x_i)\}_{i=n_1}^{n}$ **where** $\delta = 1$ **if** Y **is observed and** 0 **otherwise**

3. **Non-ignorable missing: Missing probability**

$$P(\delta = 1|y,x) = P(\delta = 1|y) = \frac{\exp(\alpha_0 + \alpha_1 y)}{1 + \exp(\alpha_0 + \alpha_1 y)} = h(y)$$

4. **The complete observations** $\{(\delta = 1, y_i, x_i)\}_{i=1}^{n_1}$ **have density**

$$f(y|\delta = 1, x) = \frac{h(y)f(y|x,\beta)}{\int h(y)f(y|x,\beta)dy} = w(y)f(y|x,\beta) \neq f(y|x,\beta)$$

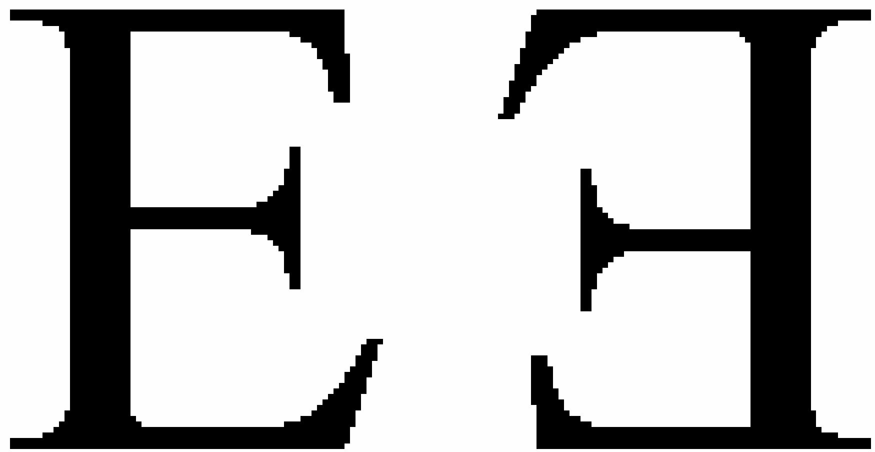
5. **Missing at random (MAR): Missing probability**

$$P(\delta = 1|y,x) = P(\delta = 1|x) = \frac{\exp(\alpha_0 + \alpha_1 x)}{1 + \exp(\alpha_0 + \alpha_1 x)} = h(x)$$

6. **The complete observations** $\{(\delta = 1, y_i, x_i)\}_{i=1}^{n_1}$ **have density**

$$f(y|\delta = 1, x) = \frac{h(x)f(y|x,\beta)}{\int h(x)f(y|x,\beta)dx} = \frac{h(x)f(y|x,\beta)}{h(x)\int f(y|x,\beta)dx} = f(y|x,\beta)$$

Reaction Time Experiment



Reaction Time Experiment (2)

Reaction time (RT) task problem (Cruz-Medina et al., 2004)

1. $N = 196$ **9 years old children**

2. **Each child given** $m = 6$ **experiments, well separated in time**

3. **Each child's RT's, in millisecond are recorded**

4. **Groups of children represent different cognitive developmental stages**

5. **Cruz-Medina et al. (2004) found two groups**

6. **Interest is in the proportion of children in the two groups**

Multivariate mixture model

1. **Multivariate data $\{(x_{i1}, x_{i2}, x_{i3})\}_{i=1}^{n}$ with density**

$$h(x_1, x_2, x_3) = \lambda f(x_1, x_2, x_3) + (1 - \lambda) g(x_1, x_2, x_3)$$

2. **$X = (X_1, X_2, X_3)$ represents the different "test" results on an observation and λ is the mixture probability**

3. **Main interest is in the mixture proportion λ**

4. **Often no "gold standard" or training samples**

5. **Parametric vs nonparametric analysis**

6. **Nonparametric models unidentifiable if the dimension of X is below 3 (Hall and Zhou, 2003)**

7. **Conditional iid model (Cruz-Medina, Hettmansperger and Thomas, 2004)**

$$(x_1, x_2, x_3) \sim h(x_1, x_2, x_3) = \lambda f(x_1) f(x_2) f(x_3) + (1 - \lambda) g(x_1) g(x_2) g(x_3)$$

   i.e. given the group membership, components have independent and identical distributions

Semiparametric approach

1. **No need for identically distributed components**

2. **No need to discretize the data**

3. **For the i-th component, assume the density ratio model (Anderson, 1979)**

$$\frac{g_i(t)}{f_i(t)} = \exp(\alpha_i + \beta_i t + \gamma_i t^2) \quad \Rightarrow \quad g_i(t) = \overbrace{\exp(\alpha_i + \beta_i t + \gamma_i t^2)}^{w(t)} f_i(t)$$

4. **Applications in malaria study (Qin and Leung, 2005), quantitative traits analysis (Zou, Fine and Yandell, 2003), melanoma study (Qin et al., 2002)**

5. **With the density ratio model, the joint density is**

$$h(x_{i1}, x_{i2}, x_{i3}) = [\lambda + (1-\lambda)\exp\{\sum_{j=1}^{3} \alpha_j + \sum_{j=1}^{3} \beta_j x_{ij} + \sum_{j=1}^{3} \gamma_j x_{ij}^2\}] \prod_{j=1}^{3} f_j(x_{ij})$$

Semiparametric approach (2)

1. The likelihood is

$$L = \prod_{i=1}^{n} [\lambda + (1-\lambda)\exp\{\sum_{j=1}^{3} \alpha_j + \sum_{j=1}^{3} \beta_j x_{ij} + \sum_{j=1}^{3} \gamma_j x_{ij}^2\}] \prod_{j=1}^{3} dF_j(x_{ij})$$

$F_j(x_{ij}) = p_{ij}$ **jumps at** $x_{ij}, i = 1, 2, ..., n; j = 1, 2, 3$

2. The log-likelihood is

$$l = \sum_{i=1}^{n} \log[\lambda + (1-\lambda)\exp\{\sum_{j=1}^{3} \alpha_j + \sum_{j=1}^{3} \beta_j x_{ij} + \sum_{j=1}^{3} \gamma_j x_{ij}^2\}] + \sum_{j=1}^{3}\sum_{i=1}^{n} \log p_{ij}$$

3. For fixed $(\lambda, \alpha, \beta, \gamma)$**, we maximize the** p_{ij}**'s subject to the constraints**

$$\sum_{i=1}^{n} p_{ij} = 1, \quad p_{ij} \geq 0, \quad \sum_{i=1}^{n} p_{ij}\{\exp(\alpha_j + \beta_j x_{ij} + \gamma_i x_{ij}^2) - 1\} = 0$$

Semiparametric approach (3)

1. **We can use a Lagrange multiplier argument, which leads to**

$$p_{ij} = \frac{1}{n}\frac{1}{1+\eta_j\{\exp(\alpha_j+\beta_j x_{ij}+\gamma_j x_{ij}^2)-1\}},$$

where η_j **is a Lagrange multiplier determined by the equation**

$$\sum_{i=1}^{n}\frac{\exp(\alpha_j+\beta_j x_{ij}+\gamma_j x_{ij}^2)-1}{1+\eta_j\{\exp(\alpha_j+\beta_j x_{ij}+\gamma_j x_{ij}^2)-1\}}=0.$$

2. **Substituting the** p_{ij}**'s back into the log-likelihood, we have a semiparametric log-likelihood**

$$\begin{aligned} l(\lambda,\alpha,\beta,\gamma) &= \sum_{i=1}^{n}\log[\lambda+(1-\lambda)\exp\{\sum_{j=1}^{3}\alpha_j+\sum_{j=1}^{3}\beta_j x_{ij}+\sum_{j=1}^{3}\gamma_j x_{ij}^2\}] \\ &- \sum_{j=1}^{3}\sum_{i=1}^{n}\log\{1+\eta_j[\exp(\alpha_j+\beta_j x_{ij}+\gamma_j x_{ij}^2)-1]\} \end{aligned}$$

The underlying parameters can be estimated by maximizing ℓ **with respect to** $(\lambda,\alpha,\beta,\gamma)$**.**

Semiparametric approach (4)

1. The joint cumulative distribution can be estimated by

$$\hat{H}(t_1, t_2, t_3) = \hat{\lambda}\hat{F}(t_1, t_2, t_3) + (1 - \hat{\lambda})\hat{G}(t_1, t_2, t_3)$$

where

$$\hat{F}(t_1, t_2, t_3) = \sum_{j=1}^{n} I(x_{1j} \le t_1, x_{2j} \le t_2, x_{3j} \le t_3)\hat{p}_{1j}\hat{p}_{2j}\hat{p}_{3j}$$

$$\begin{aligned}\hat{G}(t_1, t_2, t_3) \;=\; & \sum_{j=1}^{n} I(x_{1j} \le t_1, x_{2j} \le t_2, x_{3j} \le t_3)\hat{p}_{1j}\hat{p}_{2j}\hat{p}_{3j} \\ & \exp\{\sum_{i=1}^{3} \hat{\alpha}_i + \sum_{i=1}^{3} \hat{\beta}_i x_{ij} + \sum_{i=1}^{3} \hat{\gamma}_i x_{ij}^2\}\end{aligned}$$

2. The marginal distributions F_i and G_i can be estimated by

$$\hat{F}_i(t) = \sum_{j=1}^{n} \hat{p}_{ij} I(x_{ij} \le t), \qquad \hat{G}_i(t) = \sum_{j=1}^{n} \hat{p}_{ij} \exp(\alpha_i + \beta_i x_{ij} + \gamma_i x_{ij}^2) I(x_{ij} \le t)$$

Semiparametric approach (5)

1. **Difficult to maximize the semi-parametric likelihood**

$$\begin{aligned} l(\lambda,\alpha,\beta,\gamma) &= \sum_{i=1}^{n} \log[\lambda + (1-\lambda)\exp\{\sum_{j=1}^{3}\alpha_j + \sum_{j=1}^{3}\beta_j x_{ij} + \sum_{j=1}^{3}\gamma_j x_{ij}^2\}] \\ &- \sum_{j=1}^{3}\sum_{i=1}^{n} \log\{1+\eta_j[\exp(\alpha_j+\beta_j x_{ij}+\gamma_j x_{ij}^2)-1]\} \end{aligned}$$

2. **Use a semi-parametric EM algorithm**

3. **If** $(d_i, x_{i1}, x_{i2}, x_{i3})$ **are observed,**

$$\ell_F(\xi) = \sum_{i=1}^{n}[d_i \log\lambda + (1-d_i)\log(1-\lambda)] + \sum_{i=1}^{n}\sum_{j=1}^{3}[\log p_{ij} + (1-d_i)(\alpha_j+\beta_j x_{ij}+\gamma_j x_{ij}^2)],$$

where $\xi = (\lambda, \theta_1 = (\alpha_1,\beta_1,\gamma_1), \theta_2 = (\alpha_2,\beta_2,\gamma_2), \theta_3 = (\alpha_3,\beta_3,\gamma_3), \eta = (\eta_1,\eta_2,\eta_3))$

4. **Given the current estimate** $\xi^k = (\lambda^k, \theta_1^k, \theta_2^k, \theta_3^k)$ **and conditional on the observed data gives**

$$E(\ell_F(\xi)|\xi^k) = \sum_{i=1}^{n}[w_i^k \log\lambda^k + (1-w_i^k)\log(1-\lambda^k)] + \sum_{j=1}^{3}\sum_{i=1}^{n}[\log p_{ij} + (1-w_i^k)(\alpha_j^k+\beta_j^k x_{ij}+\gamma_j^k x_{ij}^2)],$$

where

$$w_i^k = \frac{\lambda^k}{\lambda^k + (1-\lambda^k)\exp\{\Sigma_{j=1}^3(\alpha_j^k+\beta_j^k x_{ij}+\gamma_j^k x_{ij}^2)\}}.$$

Semiparametric approach (6)

1. Imposing the constraints

$$\sum_{i=1}^{n} p_{ij} = 1, \quad p_{ij} \geq 0, \quad \sum_{i=1}^{n} p_{ij}\{\exp(\alpha_j + \beta_j x_{ij} + \gamma_j x_{ij}^2) - 1\} = 0,$$

gives a profilied "Expected" log-likelihood

$$\begin{aligned} E(\ell_F(\xi)|\xi^k) &= \sum_{i=1}^{n}[w_i^k \log\lambda + (1-w_i^k)\log(1-\lambda)] + \sum_{j=1}^{3}\sum_{i=1}^{n}(1-w_i^k)(\alpha_j + \beta_j x_{ij} + \gamma_j x_{ij}^2) \\ &- \sum_{j=1}^{3}\sum_{i=1}^{n}\log[1 + \eta_j\{\exp(\alpha_j + \beta_j x_{ij} + \gamma_j x_{ij}^2) - 1\}]. \end{aligned}$$

For given w_i^k, maximizing ℓ_P with respect to $(\lambda, \alpha_j, \beta_j, \gamma_j)$ gives ξ^{k+1}

2. Update

$$w_i^{k+1} = \frac{\lambda^{k+1}}{\lambda^{k+1} + (1-\lambda^{k+1})\exp\{\Sigma_{j=1}^{3}(\alpha_j^{k+1} + \beta_j^{k+1} x_{ij} + \gamma_j^{k+1} x_{ij}^2)\}}.$$

3. Iterate Steps 1 and 2 until convergence

Reaction time (RT) task problem

1. For illustration, we use the first three test results (components) from Cruz-Medina et al's (2004) data

2. The first and second components are significantly different ($p = 0.000226$) $\Rightarrow$ identically distribution assumption not valid

3. Monotonic transform of the original data

$$Y_{ij} = \{\log(X_{ij}) - a\}/b, i = 1, 2, ..., 197; j = 1, 2, 3$$

   where a, b are mean and sd of $\log(X_{ij}), i = 1, 2, ..., 197; j = 1, 2, 3$

4. We applied the density ratio model to the transformed data

5. $\hat{\lambda} = 0.568$ (95% CI 0.420 to 0.705)

Figure 1: Histograms of observed components 1, 2 and 3

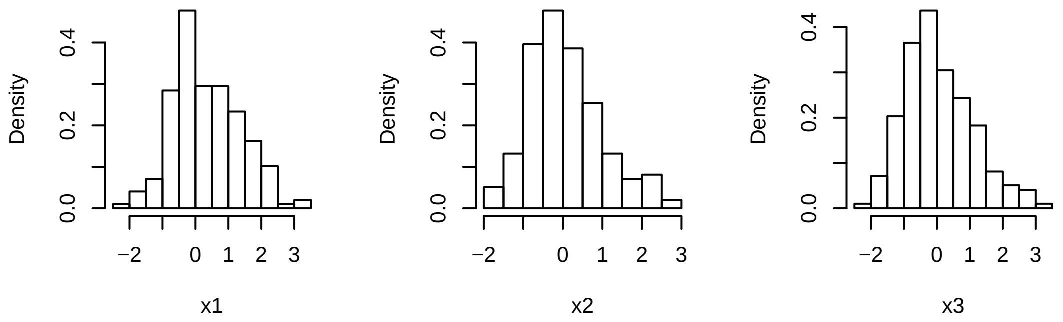


Figure 2: Semiparametric estimation of $F_i, G_i, i = 1, 2, 3$

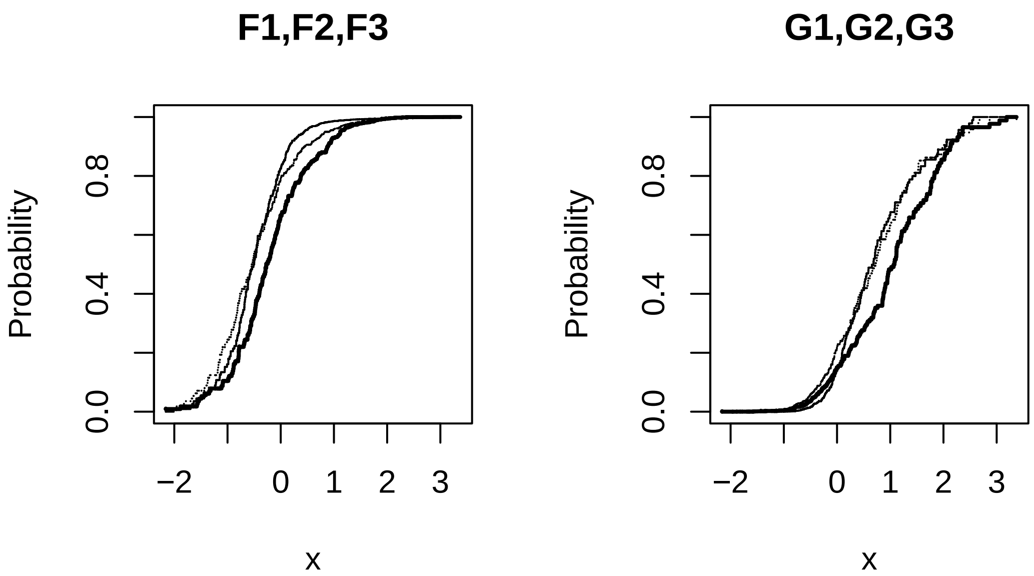
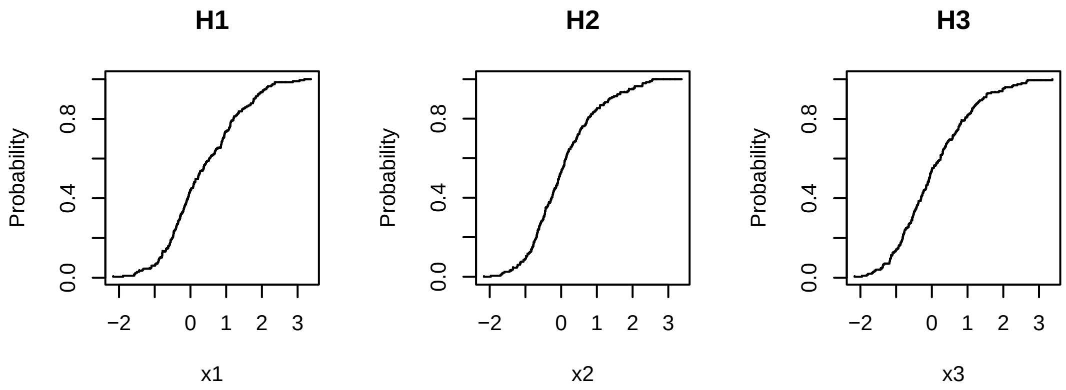


Figure 3: Empirical and semiparametric estimation of $H_i, i = 1, 2, 3$



Simulation Results

1. Trivariate normal mixture model:

$$(x_1, x_2, x_3) \sim \lambda N(0,1)N(0,1)N(0,1) + (1-\lambda)N(\mu_1, \sigma_1^2)N(\mu_2, \sigma_2^2)N(\mu_3, \sigma_3^2)$$

2. Small separation: $(\mu_1, \mu_2, \mu_3) = (1, 1.5, 2.5)$**,** $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1, 1, 1)$

Large separation: $(\mu_1, \mu_2, \mu_3) = (2, 2.5, 3)$**,** $(\sigma_1^2, \sigma_2^2, \sigma_3^2) = (1.5, 2, 1)$

3. 250 **simulations each with** $n = 500$**,** $n = 1000$

4. Comparison to MLE

Mean (variance) of the parametric and semiparametric estimates based on 250 simulations, sample size=1000.

		$(X,Y,Z) \sim \lambda N(0,1)N(0,1)N(0,1) + (1-\lambda)N(1,1)N(1.5,1)N(2.5,1)$		
True	Estimate	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.8$
-	$\hat{\lambda}$	0.30082 (0.00047)	0.50009 (0.00049)	0.80015 (0.00043)
-	$\hat{\lambda}_F$	0.30322 (0.00025)	0.50453 (0.00027)	0.80245 (0.00025)
-0.5	$\hat{\alpha}_1$	-0.50562 (0.00527)	-0.51543 (0.00472)	-0.52636 (0.01183)
-	$\hat{\alpha}_{F1}$	-0.49851 (0.00382)	-0.50667 (0.00564)	-0.54619 (0.02442)
1	$\hat{\beta}_1$	1.01779 (0.01557)	1.03180 (0.01529)	1.06606 (0.04352)
-	$\hat{\beta}_{F1}$	0.98398 (0.01317)	0.99389 (0.01774)	1.06254 (0.06971)
0	$\hat{\gamma}_1$	0.00127 (0.00662)	-0.00078 (0.00563)	-0.01970 (0.01034)
-	$\hat{\gamma}_{F1}$	-0.00232 (0.00844)	-0.00141 (0.00793)	-0.02772 (0.01729)
-1.125	$\hat{\alpha}_2$	-1.16680 (0.01393)	-1.16635 (0.01338)	-1.21708 (0.05504)
-	$\hat{\alpha}_{F2}$	-1.13784 (0.01847)	-1.14315 (0.02075)	-1.19579 (0.09433)
1.5	$\hat{\beta}_2$	1.55555 (0.04556)	1.54390 (0.04697)	1.59298 (0.12831)
-	$\hat{\beta}_{F2}$	1.49737 (0.02717)	1.49921 (0.03325)	1.55584 (0.12707)
0	$\hat{\gamma}_2$	0.00674 (0.01685)	0.00323 (0.01216)	-0.00901 (0.01926)
-	$\hat{\gamma}_{F2}$	-0.00481 (0.00776)	-0.01202 (0.00542)	-0.01840 (0.01337)
-3.125	$\hat{\alpha}_3$	-3.27271 (0.32721)	-3.23815 (0.40743)	-3.41140 (1.28678)
-	$\hat{\alpha}_{F3}$	-3.12622 (0.04783)	-3.14263 (0.05807)	-3.31905 (0.30482)
2.5	$\hat{\beta}_3$	2.63104 (0.68137)	2.60930 (0.78394)	2.69273 (1.54765)
-	$\hat{\beta}_{F3}$	2.46494 (0.08358)	2.47943 (0.11337)	2.74958 (0.54392)
0	$\hat{\gamma}_3$	0.00236 (0.11023)	-0.01118 (0.10637)	-0.01988 (0.13486)
-	$\hat{\gamma}_{F3}$	-0.01394 (0.00905)	-0.01474 (0.00755)	-0.05948 (0.01997)

Mean (variance) of the parametric and semiparametric estimates based on 250 simulations, sample size=1000.

		$(X,Y,Z) \sim \lambda N(0,1)N(0,1)N(0,1) + (1-\lambda)N(2.0,1.5)N(2.5,2)N(3.0,1)$		
True	Estimate	$\lambda = 0.3$	$\lambda = 0.5$	$\lambda = 0.8$
-	$\hat{\lambda}$	0.30074 (0.00026)	0.50144 (0.00028)	0.79971 (0.00020)
-	$\hat{\lambda}_F$	0.30281 (0.00021)	0.50442 (0.00023)	0.80029 (0.00016)
-1.33333	$\hat{\alpha}_1$	-1.53862 (0.01113)	-1.56951 (0.01192)	-1.58445 (0.03663)
-	$\hat{\alpha}_{F1}$	-0.97706 (0.01649)	-1.01759 (0.04451)	-1.02214 (0.08872)
1.33333	$\hat{\beta}_1$	1.35424 (0.02243)	1.36952 (0.02886)	1.39207 (0.07640)
-	$\hat{\beta}_{F1}$	0.96076 (0.01731)	0.99896 (0.03973)	1.00075 (0.07700)
0.16667	$\hat{\gamma}_1$	0.16474 (0.00768)	0.15600 (0.01174)	0.15923 (0.02107)
-	$\hat{\gamma}_{F1}$	0.24323 (0.00345)	0.24662 (0.00417)	0.25286 (0.00898)
-1.5625	$\hat{\alpha}_2$	-1.95385 (0.01900)	-1.98385 (0.06357)	-2.00770 (0.10458)
-	$\hat{\alpha}_{F2}$	-0.86195 (0.01213)	-0.93308 (0.01827)	-1.00566 (0.03812)
1.25	$\hat{\beta}_2$	1.28867 (0.02664)	1.28982 (0.04029)	1.32560 (0.09575)
-	$\hat{\beta}_{F2}$	0.67836 (0.00925)	0.73100 (0.01407)	0.78783 (0.02331)
0.25	$\hat{\gamma}_2$	0.25576 (0.00098)	0.24927 (0.00878)	0.24005 (0.01269)
-	$\hat{\gamma}_{F2}$	0.35506 (0.00697)	0.34111 (0.00388)	0.34259 (0.00305)
-4.5	$\hat{\alpha}_3$	-4.61913 (0.42892)	-4.69476 (0.57040)	-4.65478 (0.88890)
-	$\hat{\alpha}_{F3}$	-4.46534 (0.06361)	-4.48544 (0.07980)	-4.56179 (0.26862)
3	$\hat{\beta}_3$	3.13069 (0.82084)	3.15496 (0.93593)	3.06798 (1.14193)
-	$\hat{\beta}_{F3}$	2.94575 (0.09337)	2.93681 (0.10696)	3.06447 (0.41936)
0	$\hat{\gamma}_3$	-0.01308 (0.10958)	-0.01514 (0.09888)	0.01300 (0.10101)
-	$\hat{\gamma}_{F3}$	0.00073 (0.01073)	-0.01450 (0.00569)	-0.02310 (0.01414)