Online Supplement: Discussion on "PROBABILISTIC INDEX MODELS" by Olivier Thas, Jan De Neve, Lieven Clement and Jean-Pierre Ottoy (JRSS B 2012)

Emanuel Parzen and Subhadeep Mukhopadhyay (Deep)
Department of Statistics, Texas A&M University
Feb 15, 2012

1 (X, Y) Modeling, Population and Sample Probability

Modern Applied Statistics (unifying many cultures of Statistical Science and Statistical Learning methods,) is the aim of current research by Parzen and Deep. Eventual goal is high dimensional data modeling (big and small data science). Here our primary goal is to identify and estimate models for (X, Y), where X and Y are continuous or discrete variables.

X continuous, Y continuous	Regression (linear and non-linear).
X binary, Y continuous	Two sample.
X discrete, Y continuous	Multi-sample (analysis of variance).
X continuous, Y binary	Logistic regression.
X continuous, Y discrete	Multiple logistic regression.
X binary , Y binary	2×2 Contingency table.
X discrete, Y discrete	r by c Contingency table.

This outline does not use separate notation for population probability and sample probability. Fundamental to our approach is to define population concepts with estimators defined analogously from sample distribution functions.

To describe marginal probability law of a random variable X use

Distribution function	$F(x;X) = F_X(x)$.
Probability density function	$f(x;X) = f_X(x)$.
Probability mass function	$p(x;X) = p_X(x)$.
Quantile function	$Q(u;X) = F^{-1}(u;X), 0 < u < 1$.
Mid-distribution function	$F^{\text{mid}}(x;X) = \Pr[X < x] + .5\Pr[X = x]$.
Rank Transform	$F^{\text{mid}}(X;X) = (\text{Rank}(X) - .5)/n$ (sample of size n from F).

2 Comparison Density, Comparison Distribution, CMPI$(Y|X)$

To compare (measure differences between) distributions F and G of a variable define comparison density $d(u;G,F)$, $0 < u < 1$ and comparison distribution $D(u;G,F) = \int_0^u d(t;G,F)\,\mathrm{d}t$. When G, F continuous satisfying $f_G(x) = 0$ implies $f_F(x) = 0$, we have,

$$d(u;G,F) = \frac{f_F\big(Q_G(u)\big)}{f_G\big(Q_G(u)\big)}, \text{ and } D(u;G,F) = F\big(Q_G(u)\big). \tag{2.1}$$

When F, G discrete distributions define

$$d(u;G,F) = \frac{p_F\big(Q_G(u)\big)}{p_G\big(Q_G(u)\big)}, \text{ and } D(u;G,F) = F\big(Q_G(u)\big), \tag{2.2}$$

at u G-exact, $u = G(x)$ for some x.

Estimating $d(u) = d(u;G,F)$ provides *Weighted model* for unknown density $f_F(x) = f_G(x)d\big(G(x)\big)$.

PP-plot: When G, F discrete with jumps at $x_1 < \ldots < x_r$, graph of $D(u;G,F)$ is called PP plot; it joins linearly $(0,0), \big(G(x_i), F(x_i)\big), i = 1, \ldots, r$.

CMPI: Area under the PP graph (Fig. 3) is a statistic, denoted CMPI, called Comparison Mid-Probability Index,

$$\text{CMPI}(G,F) \,=\, \sum p_G(x_i)F_X^{\text{mid}}(x_i) \,=\, \mathbb{E}_G[F^{\text{mid}}(X;X)]. \tag{2.3}$$

Verify $\text{CMPI}(F,F) = .5$. Study dependence of (X,Y) by $\text{CMPI}(Y|X) = \mathbb{E}\big[F^{\text{mid}}(Y;Y) \mid X\big]$, $\text{CMPI}(X|Y) = \mathbb{E}\big[F^{\text{mid}}(X;X) \mid Y\big]$, conditional quantile $Q(u;F^{\text{mid}}(Y)|X)$, which we estimate by new nonparametric methods.

3 Copula Density of (X, Y), Bayes Theorem for Conditional Comparison Densities

A central role in the extension of multivariate analysis to possibly non-normal data (X, Y) is estimation of copula density function $\mathrm{cop}(u, v; X, Y)$ which is traditionally defined as the ratio of the joint probability density to the product of marginal densities. To extend the concept to both continuous and discrete variables we propose the following definition with two alternative formulas involving conditional comparison density functions, whose equality is an application of Bayes Theorem for random variables:

$$\text{Copula Density}: \quad \mathrm{cop}(u, v; X, Y) \;=\; d\left(v; Y, Y | X = Q_X(u)\right) \;=\; d\left(u; X, X | Y = Q_Y(v)\right). \quad (3.1)$$

Copula distribution function $\mathrm{Cop}(u, v; X, Y)$ is defined as the double integral over $0 < s < u$, $0 < t < v$ of $\mathrm{cop}(s, t; X, Y)$. One can show that for F-exact u, G-exact v

$$\mathrm{Cop}(u, v; X, Y) \;=\; F_{X,Y}\big(Q_X(u), Q_Y(v)\big). \quad (3.2)$$

Rank Transform Interpretation: We interpret $\mathrm{cop}(u, v; X, Y)$ as joint probability density of rank transformed data $\big(F_X^{\mathrm{mid}}(X), F_Y^{\mathrm{mid}}(Y)\big)$. We estimate it by maximum entropy density estimation rather than by two dimensional kernel density estimation.

Verify: $\mathrm{CMPI}(Y | X = Q_X(u)) \;=\; \int_0^1 v \, \mathrm{cop}(u, v; X, Y) \, \mathrm{d}v$.

4 Example Biostatistical Data, Score Functions, LP Comoments

We model by choosing sufficient statistics (score functions) before estimating parameters of probability models. To form estimators of above concepts we first construct for each variable X_k orthonormal score functions which are polynomial functions of $F^{\mathrm{mid}}(X_k)$. We change notation for observations to (X_1, X_2, X_3). As example we consider childhood respiratory disease study discussed in Thas et al (2012); data (from Bernard Rosner, Fundamentals of Biostatistics) studies FEV (lung function measure, volume of exhaled air) as a function of age (from 3 to 19) and smoking history (yes or no). We observe sample of size $n = 654$ of (X_1=FEV, X_2=age, X_3=0 or 1 for smoking history).

Step 0. Marginal Quantiles. For each variable X_k, compute (sample) quantile $Q(u; X_k), 0 < u < 1$. Compute unique values x_i in sample; $u_i = F(x_i; X_k); u_i^{\text{mid}} = F^{\text{mid}}(x_i; X_k); u_0 = 0$. For $u_{i-1} < u \leq u_i, Q(u; X_k) = x_i$.

Mid-Quantile, Median, Quartiles: Define $Q^{\text{mid}}(u; X_k)$ to join linearly (u_i^{mid}, x_i). Define median $Q2$, quartiles $Q1, Q3$ by $Q2 = Q^{\text{mid}}(.5; X_k)$, $Q1 = Q^{\text{mid}}(.25; X_k)$, $Q3 = Q^{\text{mid}}(.75; X_k)$. Measures of location and scale are mid-quartile $MQ = .5(Q1+Q3)$, quartile deviation $SQ = 2(Q3 - Q1)$. Skewness, tail index measured by $\text{QIQ}(u) = (Q^{\text{mid}}(u) - MQ)/SQ$, values at $.5, .05, .95$. Usual Tukey definition of X being outlier equivalent to $QI(X) = (X - MQ)/SQ$ satisfying $|QI(X)| > 1$. Five Number summary of X: $MQ, SQ, \text{QIQ}(.5), \text{QIQ}(.05), \text{QIQ}(.95)$.

Step 0. Example. Quantiles of FEV, Age and Smoke (Fig. 1).

Step 1. Score Functions. $S_{j,k}(X_k)$. Approximately (useful for variables with many X_k values)

$$S_{j,k}(X_k) = \text{Len}_j \left[F_{X_k}^{\text{mid}}\big(Q_{X_k}(u)\big) \right], \tag{4.1}$$

where $\text{Len}_j(u), 0 < u < 1, j > 0$, are orthonormal Legendre polynomials shifted to unit interval. Exact construction starts with score function $S_{1,k}(u; X_k) = S_{1,k}(Q(u; X_k))$ defined as

$$S_{1;k}(X_k) = (F^{\text{mid}}(X_k; X_k) - .5)/\sigma_{\text{mid}}, \;\; \sigma_{\text{mid}}^2 = \text{Var}[F^{\text{mid}}(X_k; X_k)] = (1/12)\left(1 - \mathbb{E}\big[|p(X_k; X_k)|^2\big]\right). \tag{4.2}$$

Construct for $j = 2, 3, 4 \ldots$ higher order score functions $S_{j,k}(X_k)$ by Gram Schmidt orthonormalization of powers $S_{1,k}^j(X_k)$.

Step 1. Example. Score functions of FEV, Age and Smoke (Fig. 2).

Step 2. LP Score Comoments. Fix k_1, k_2. Compute for $j_i, j_2 = 0, 1, 2, 3, 4$

$$\text{LP}(j_1, j_2; k_1, k_2) = \mathbb{E}\left[S_{j_1,k_1}(X_{k_1})\, S_{j_2,k_2}(X_{k_2})\right] \tag{4.3}$$

defining zero-th order score function $S_{0,k}(X_k) = Q(; X_k)$. Above definition of zero order used to compute comoments. Alternative definition (made explicitly if needed) for copula density estimation is $S_{0,k}(X_k) = 1$. There is an extensive literature on L moments

and L comoments which is extended by our concept of LP comoments of score functions constructed for each variables. Another related concept is Gini correlation.

Correlation: For $j_1, j_2 > 0$, our comoments are correlations. For one order 0 and other order j positive, define LP score correlation.

$$\begin{aligned} \mathrm{LPR}(0, j; k_1, k_2) &= \mathrm{LP}(0, j; k_1, k_2) / \mathrm{LP}(0, 1; k_2, k_2) \\ \mathrm{LPR}(j; 0; k_1, k_2) &= \mathrm{LP}(j, 0; k_1, k_2) / \mathrm{LP}(1, 0; k_1, k_1) \end{aligned} \tag{4.4}$$

Step 2. Example. Comoments of FEV, Age, Smoke (Fig. 4).

Step 3. Nonparametric Conditional Expectation of $X_2, F^{\mathrm{mid}}(X_2)$. Conditional means $\mathbb{E}[X_2|X_1 = Q_{X_1}(u)]$, $\mathbb{E}[F^{\mathrm{mid}}(X_2; X_2)|X_1 = Q_{X_1}(u)]$ are non-linear functions of X_1 which we can compute (or estimate) by a linear combination of the orthonormal score functions $S_{j,1}(X_1)$ (Fig. 7, 9):

$$\mathbb{E}[X_2|X_1] \;=\; \mathbb{E}[X_2] + \sum_{j=1}^{4} S_{j,1}(X_1)\, \mathrm{LP}(j, 0; X_1, X_2) \tag{4.5}$$

$$\mathbb{E}\left[(F^{\mathrm{mid}}_{X_2}(X_2|X_1) - .5)/\sigma_{\mathrm{mid}}\right] \;=\; \sum_{j=1}^{4} S_{j,1}(X_1)\, \mathrm{LP}(j, 1; X_1, X_2) \tag{4.6}$$

For AIC selection (Fig. 6) of influential score function, define for $m = 1, 2, 3, 4$ and $h = 0, 1, 2, 3, 4$ "score test" measures of dependence (information) useful to detect highly dependent variables and novel associations in high dimensions:

$$\mathrm{CR}(m, h; X_2 \mid X_1) \;=\; \sum_{j=1}^{4} \mid \mathrm{LP}(j, h; X_1, X_2) \mid^2 \tag{4.7}$$

$$\mathrm{CCR}(m, H; X_2 \mid X_1) \;=\; \sum_{h=1}^{H} \mathrm{CR}\,(m, h; X_2 \mid X_1) \tag{4.8}$$

$$\mathrm{CINFOR}(X_1, X_2) \;=\; \mathrm{CCR}(4, 4; X_2, X_1) \tag{4.9}$$

"Naive" least squares estimator copula density function is (Fig. 8):

$$\mathrm{cop}_{L_2}(u_1, u_2; X_1, X_2) \;=\; 1 + \sum_{j_1, j_2} S_{j_1,1}(X_1) S_{j_2,2}(X_2)\, \mathrm{LP}(j_1, j_2; X_1.X_2) \tag{4.10}$$

"Exact" Maximum entropy (exponential model) copula density estimator identifies influential product score functions $S_{j_1,1}(X_1)S_{j_2,2}(X_2)$, models log copula density as a linear

combination of these selected double score functions, and computes estimators of parameters by Newton-Raphson method.

Step 3. Example. Scatter diagrams with conditional means plot, conditional comparison densities, copula density function.

Step 4. Verify Conditional Expectation Formula when X_1 is binary $0-1$ Define $p = \Pr[X_1 = 1], q = \Pr[X_1 = 0]$. Two sample problem observes (X_1, X_2) where X_2 continuous. Define $\mathbb{E}[X_2|X_1 = 1] = \mu_1$, $\mathbb{E}[X_2|X_1 = 0] = \mu_0$, pooled mean $\mathbb{E}[X_2] = \mu$. Verify $\mu = q\mu_0 + p\mu_1$, $\mu_0 - \mu = p(\mu_0 - \mu_1)$ and

$$F^{\text{mid}}(0; X_1) = q/2, \qquad F^{\text{mid}}(1; X_1) = 1 - .5p \tag{4.11}$$
$$S_{1,1}(X_1 = 0) = -\sqrt{p/q}, \qquad S_{1,1}(X_1 = 1) = \sqrt{q/p}.$$

Also note that, $\text{LP}(0, 1; X_1, X_2) = \sqrt{pq}(\mu_1 - \mu_0)$ and

$$S_{1,1}(X_1)\,\text{LP}(1, 0; X_1, X_2) = \begin{cases} -\sqrt{p/q}\sqrt{pq}(\mu_1 - \mu_0) = p(\mu_0 - \mu_1) & \text{at } X_1 = 0 \\ q(\mu_1 - \mu_0) & \text{at } X_1 = 1. \end{cases}$$

Next verify at $X_1 = 0$, $\mathbb{E}[X_2|X_1] - \mathbb{E}[X_2] = \mu_0 - \mu = p(\mu_0 - \mu_1)$. Similarly $\mathbb{E}[X_2|X_1 = 1] = \mu_1 - \mu = q(\mu_1 - \mu_0)$. We have calculated both sides of the formula for conditional mean and verified that they are equal.

Traditional Pooled Variance Two Sample t-statistics: $T = R/\sqrt{1 - R^2}$ where $R = \text{Corr}(X_1, X_2)$, verify

$$R \;=\; p(\mu_1 - \mu)/\sqrt{pq}\sigma(X_2) \;=\; \sqrt{pq}(\mu_1 - \mu_0)/\sigma(X_2). \tag{4.12}$$

Note that, for X_2 binary $\sigma(X_2) = \Pr(X_2 = 1)\Pr(X_2 = 0)$.

Wilcoxon Statistics: Formula for conditional mean of $F^{\text{mid}}(X_2)$ follows from above formula: $\mu = .5$, $\mu_1 = \mathbb{E}\left[F^{\text{mid}}(X_2; X_2)|X_1 = 1\right]$ = average over sample corresponding to $X_1 = 1$ of mid-ranks in pooled sample, $\sigma(F^{\text{mid}}(X_2; X_2)) = \sigma_{\text{mid}}(X_2)$.

Step 5. Estimating CMPI **given two covariates.** Given dependent response variable X_1, covariates X_2, X_3 (often X_3 binary) define,

$$\text{CMPI}(X_1 \mid X_2, X_3) \;=\; \mathbb{E}[F^{\text{mid}}(X_1; X_1) \mid X_2, X_3] \;=\; .5 \;+\; \sigma_{\text{mid}}(X_1)\mathbb{E}[S_{1,1}(X_1) \mid X_2, X_3]. \tag{4.13}$$

Our definition of CMPI in terms of conditional mean of $F^{\text{mid}}(Y)$ should be compared with interpreting regression as conditional mean of X_1 given covariates X_2 and X_3. Estimator of CMPI can be found using the formula that conditional expectation $\mathbb{E}(g(X_1)|X_2, X_3)$ when $\mathbb{E}|g(X_1)|^2 < \infty$ can be represented as a linear combination of product score functions $S_h = S_{j_2,2}(X_2)S_{j_3,3}(X_3), h = (j_2, j_3)$. We use orders 0 to 4. Express estimator of $\mathbb{E}[g(X_1)|X_2, X_3]$ by $\sum_h \theta_h S_h$. A "naive" estimator takes $\theta_h = \mathbb{E}[g(X_1)S_h]$ An "exact" estimator computes coefficients θ_h from normal equations

$$E[g(X_1)S_h] = E\left[\, E[g(X_1)|X_2, X_3]S_h \,\right]$$

A parsimonious set of indices is selected using stepwise multiple linear regression with interaction in conjunction with AIC criterion. The naive estimator of CMPI$(X_1; X_2, X_3)$ (Fig. 9) is

$$\sum_{j_2,j_3=0}^{4} S_{j_2,2}(X_2)S_{j_3}(X_3)\,\mathrm{LP}(1, j_2, j_3; X_1, X_2, X_3) \tag{4.14}$$

Note we define zero-th order score function $S_{0,k}(X_k) = 1$. "Exact" Logistic regression estimation of CMPI is recommended using the score functions selected at L_2 estimation phase.

Step 6. Classification Dependence Problem X_1 Binary $0-1$. Goal is to estimate $\Pr[X_1 = 1|X_2, X_3]$. Define standardization with $p = \Pr[X_1 = 1]$ and

$$I^* = I^*[X_1 = 1] \,=\, (I[X_1 = 1] - p)/\sqrt{p(1-p)} = \sqrt{p/(1-p)}\,\big[I[X_1 = 1]/p \,- 1\big].$$

Naive estimator is

$$\mathbb{E}[I^*|X_2.X_3] \,=\, \sum_{j_2,j_3=0}^{4} S_{j_2,2}(X_2)\, S_{j_3,3}(X_3)\, \mathbb{E}\left[I^* S_{j_2,2}(X_2)S_{j_3,3}(X_3)\right]. \tag{4.15}$$

Use Logistic regression for exact estimation of $\Pr(X_1 \mid X_2, X_3)$.

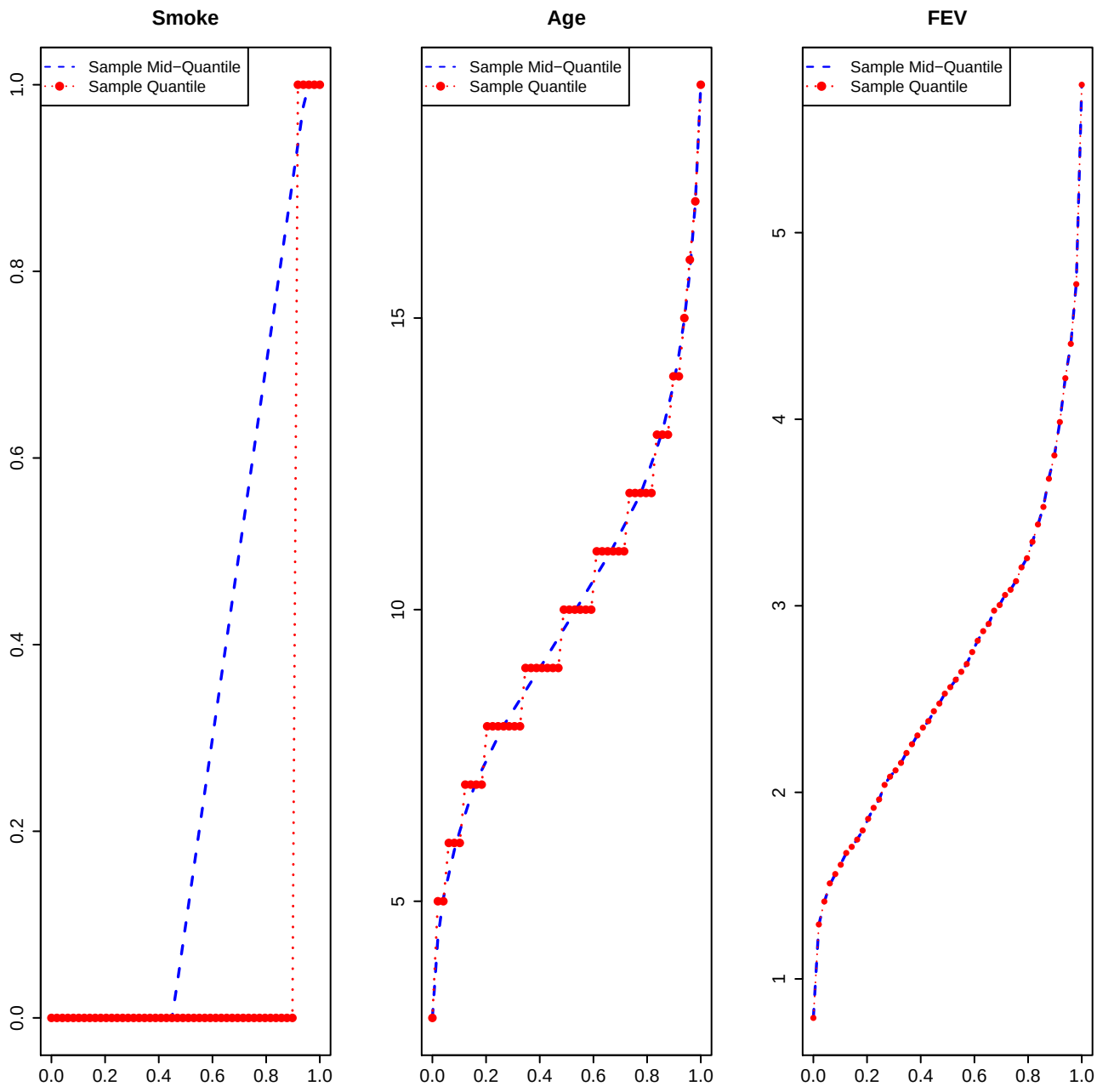


Figure 1: *Sample Quantile Functions*: $\widetilde{Q}$ and $\widetilde{Q}^{\text{mid}}$ for variables Smoke, Age and FEV.

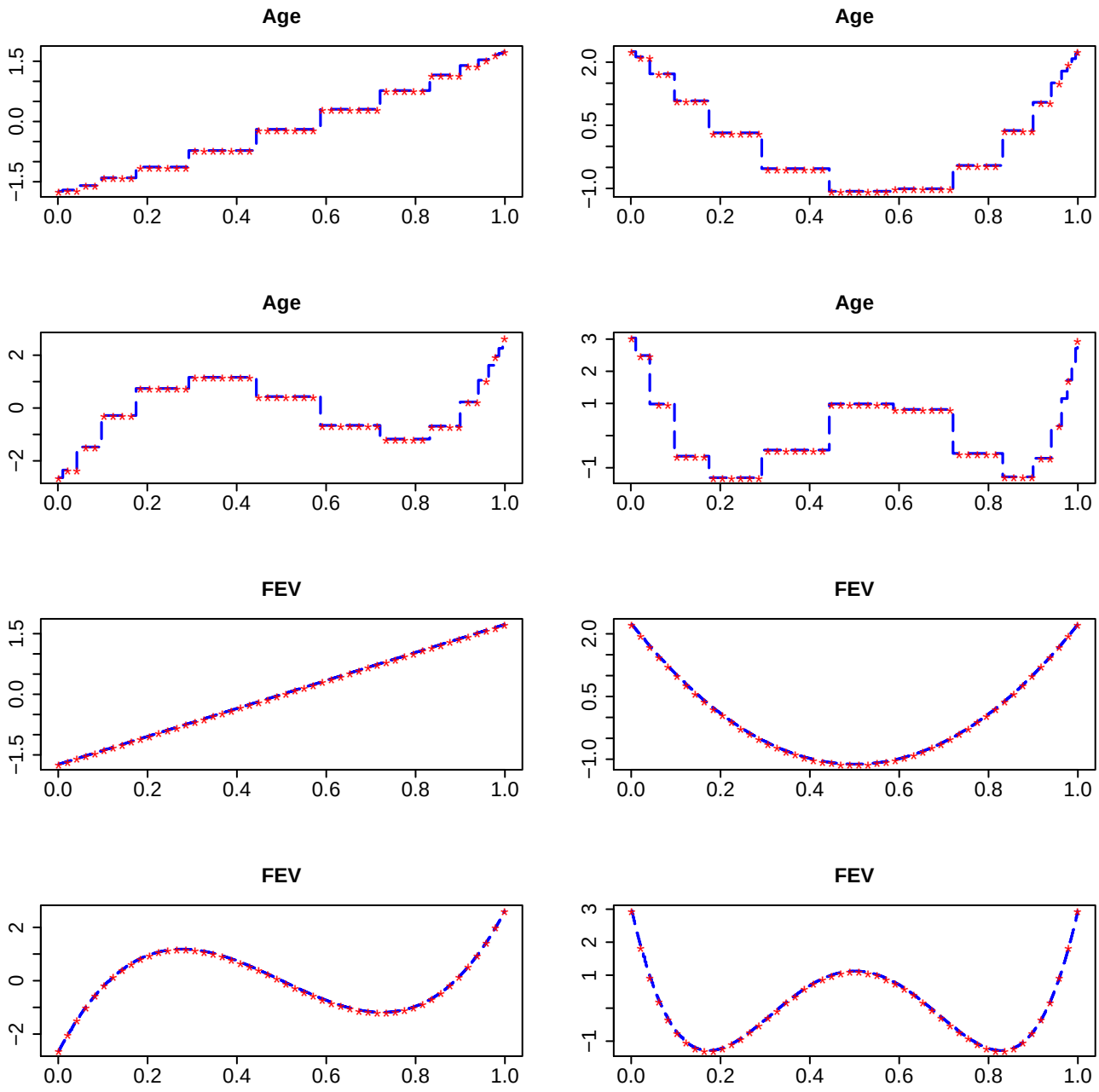


Figure 2: *Sample Mid-distribution based Orthonormal Score Functions*: S_1, S_2, S_3, S_4 for Age and FEV.

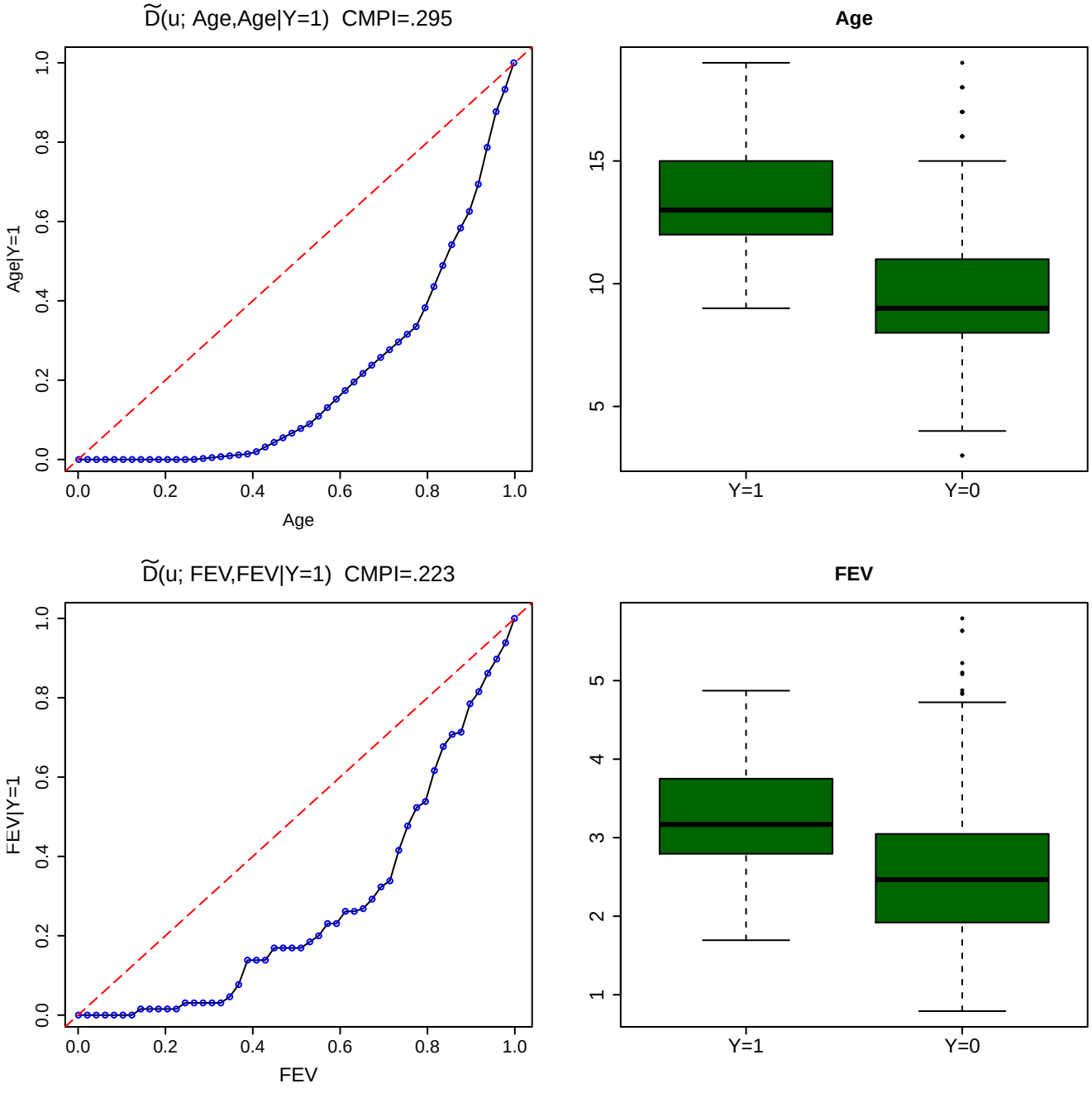
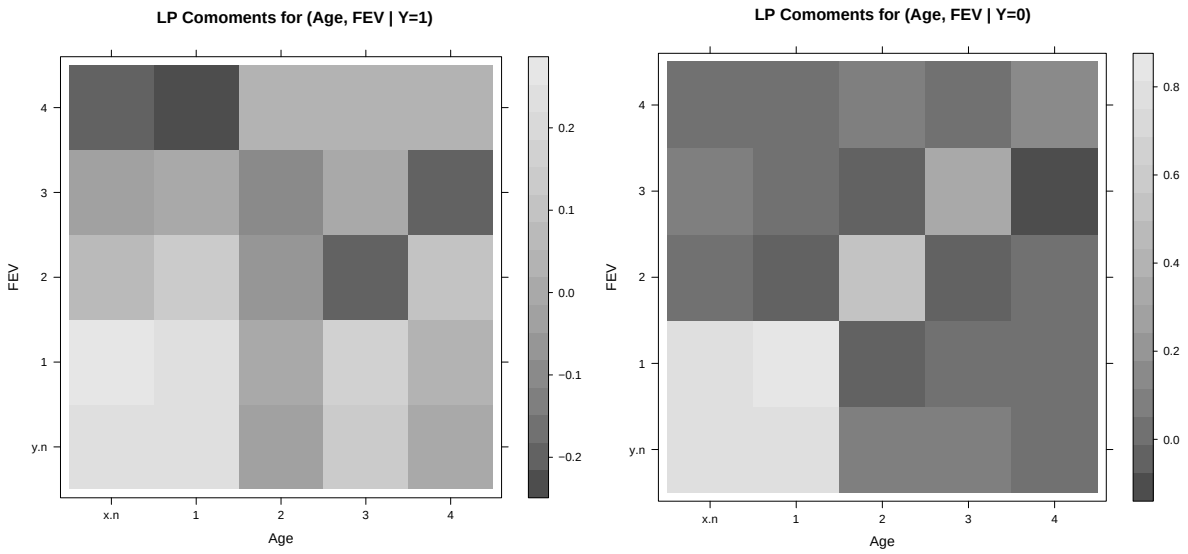


Figure 3: Sample Comparison distribution and the corresponding Comparison Mid-Probability Index (CMPI). Top left figure CMPI(Age | Smoke) and bottom left CMPI(FEV | Smoke).



(a) LP Comoment Matrix: Smoker (b) LP Comoment Matrix: Non-Smoker

Age	FEV.S0	FEV.S1	FEV.S2	FEV.S3	FEV.S4
Age.S0	**0.249**	**0.253**	0.0798	-0.024	**-0.211**
Age.S1	**0.238**	**0.231**	0.119	-0.0046	**-0.216**
Age.S2	-0.0245	0.0123	-0.081	-0.0923	0.0198
Age.S3	0.143	0.182	**-0.211**	0.0096	0.028
Age.S4	0.0094	0.0392	0.0855	-0.1889	0.0392

(c)

Age	FEV.S0	FEV.S1	FEV.S2	FEV.S3	FEV.S4
Age.S0	**0.7816**	**0.7869**	-0.0023	0.0855	0.0385
Age.S1	**0.7834**	**0.8137**	-0.0335	0.0187	0.0324
Age.S2	0.0520	-0.0477	0.5485	-0.0181	0.0692
Age.S3	0.0770	0.0136	-0.0651	0.3270	-0.0128
Age.S4	0.0163	0.0095	0.0451	-0.0780	0.1630

(d)

Figure 4: *Heat plot and original Comoment Matrix.* (c) LP comoments for Y=1; (d) LP comoments for Y=0. Few large LP comoments are shown in bold.

Age	Age.S0	Age.S1	Age.S2	Age.S3	Age.S4
Age.S0	1	.993	.0073	.129	.043
Age.S1	0.994	1.016	0	0	0
Age.S2	0.074	0	1.016	0	0
Age.S3	0.130	0	0	1.016	0
Age.S4	.043	0	0	0	1.016

(a)

Age	Age.S0	Age.S1	Age.S2	Age.S3	Age.S4
Age.S0	1	.969	.076	.205	.052
Age.S1	0.969	1	0	0	0
Age.S2	0.076	0	1	0	0
Age.S3	0.205	0	0	1	0
Age.S4	0.052	0	0	0	1

(b)

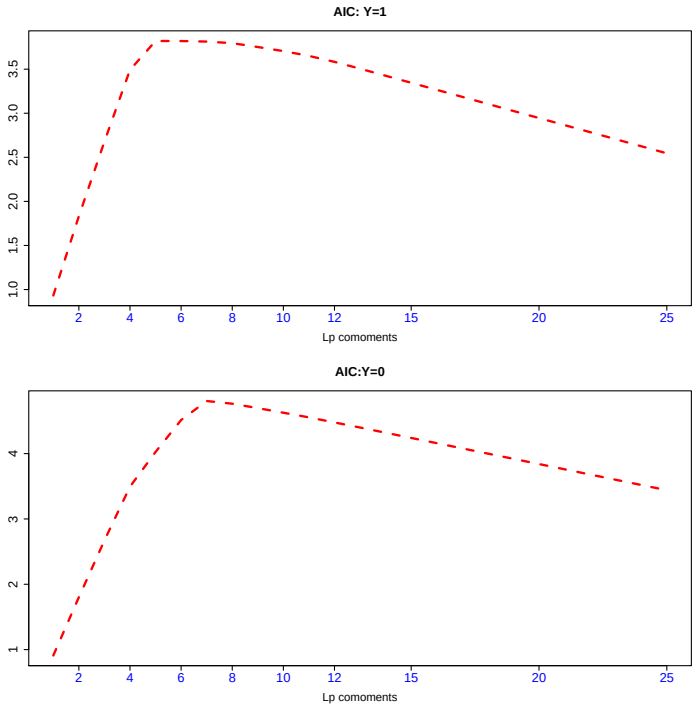
Age	FEV.S0	FEV.S1	FEV.S2	FEV.S3	FEV.S4
FEV.S0	1	.982	.087	.191	-.007
FEV.S1	.982	1.016	0	0	0
FEV.S2	.087	0	1.016	0	0
FEV.S3	.191	0	0	1.016	0
FEV.S4	-0.007	0	0	0	1.016

(c)

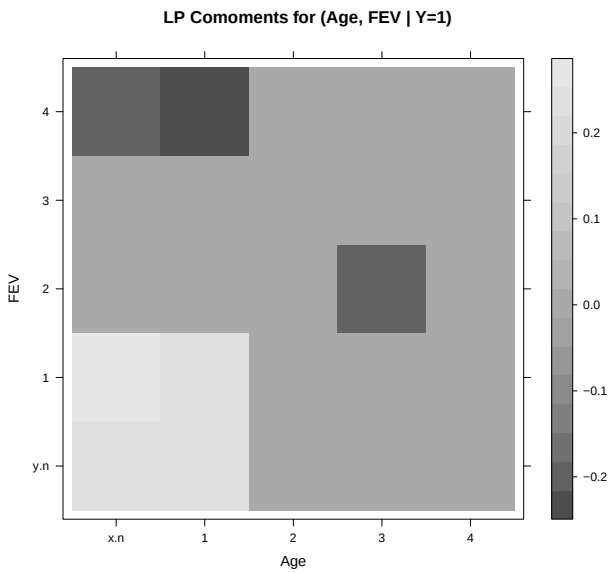
Age	FEV.S0	FEV.S1	FEV.S2	FEV.S3	FEV.S4
FEV.S0	1	.960	.168	0183	.082
FEV.S1	.960	1.002	0	0	0
FEV.S2	.168	0	1.002	0	0
FEV.S3	.183	0	0	1.002	0
FEV.S4	0.082	0	0	0	1.002

(d)

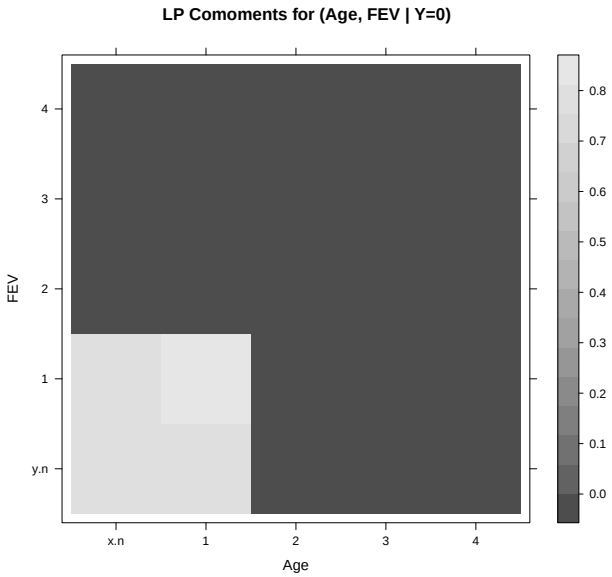
Figure 5: (a) LP moments of Age of Smokers; (b) LP moments of Age of Non-smokers; (c) LP moments of FEV of Smokers; (d) LP moments of FEV of Non-smokers.



(a) AIC

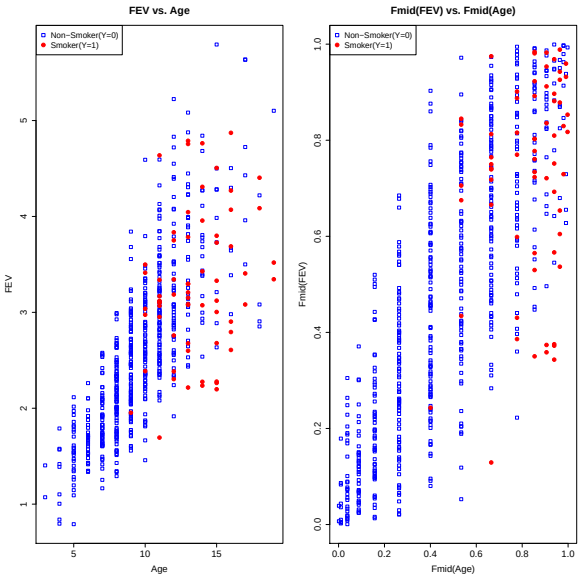


(b) LP Comoment Matrix: Smoker

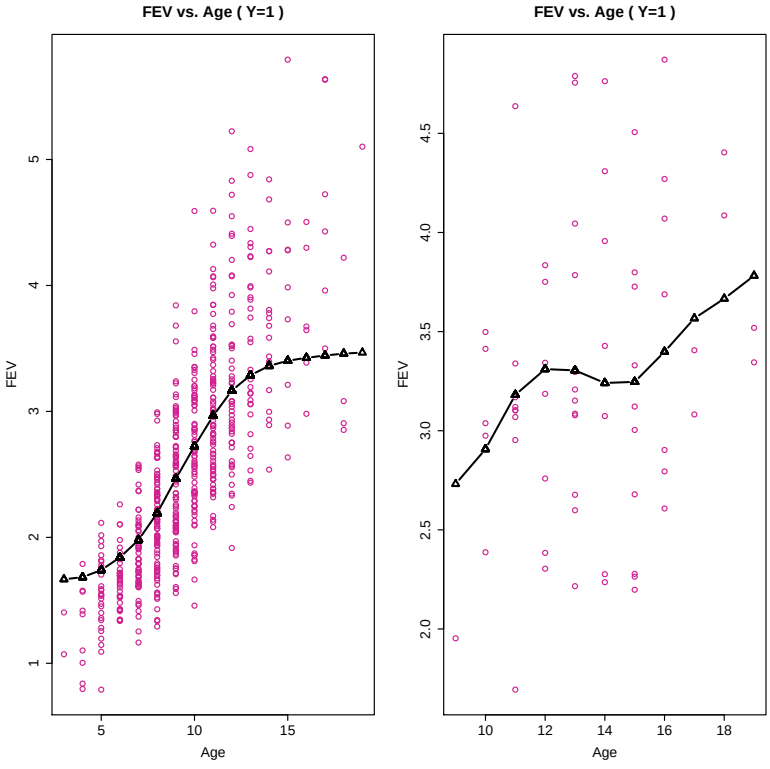


(c) LP Comoment Matrix: Non-Smoker

Figure 6: *Data Adaptive thresholding using AIC.*



(a) Scatter Plot



(b) Nonparametric Conditional mean curve

Figure 7: Scatter plot and nonparametric estimator of conditional mean; Compare with linear regression.

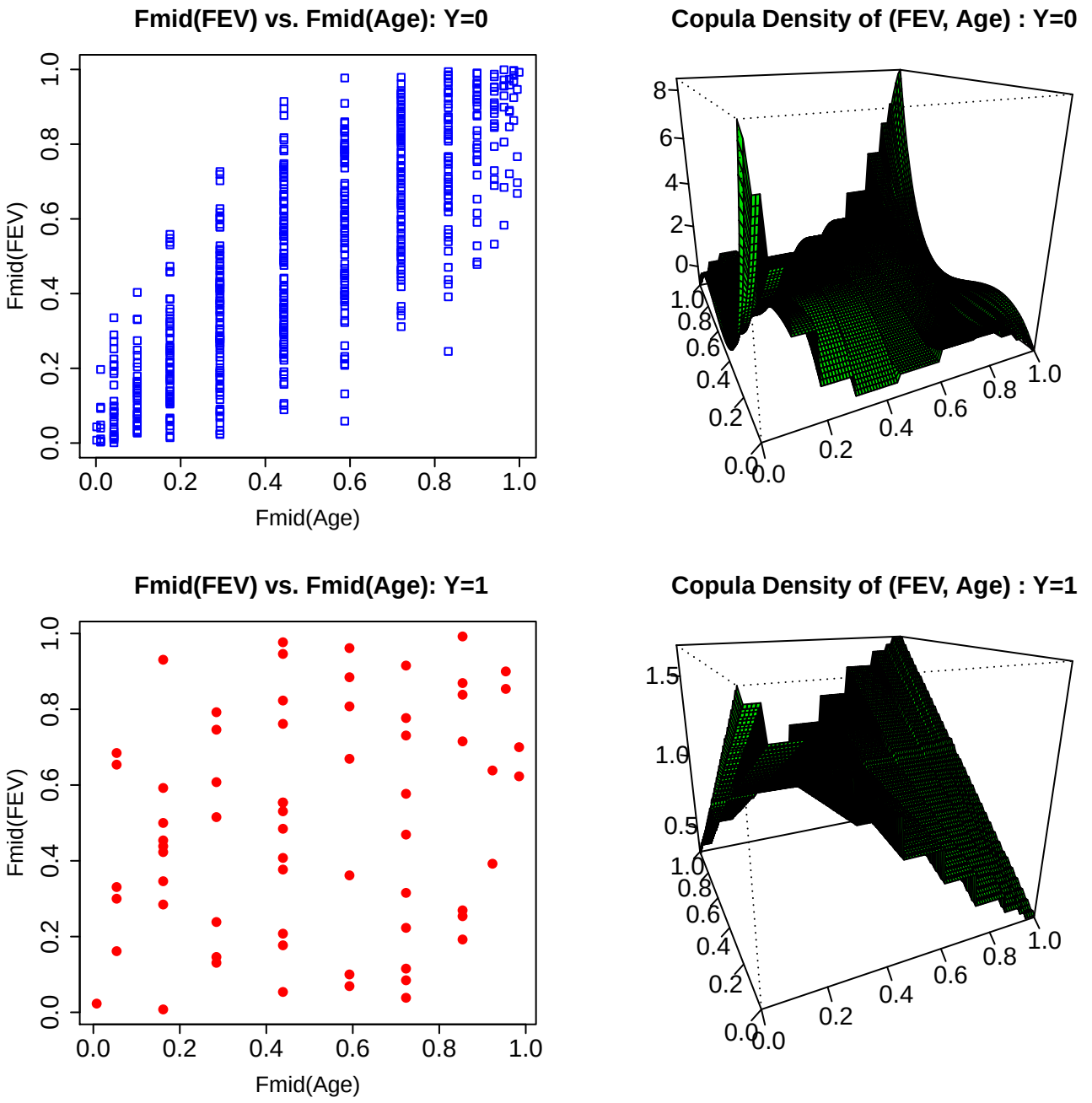


Figure 8: L_2 copula estimation for (Age,FEV) for smoker and non-smoker class, which gives a complete picture of the underlying dependence and how it changes. Note that Age is discrete and FEV is continuous random variable, which we precisely achieve via Eq. (3.1). Also note that for Non-smokers ($Y = 0$) there is a considerable "tail dependency" exists (top left panel), accurately captured by our estimated copula (top right panel).

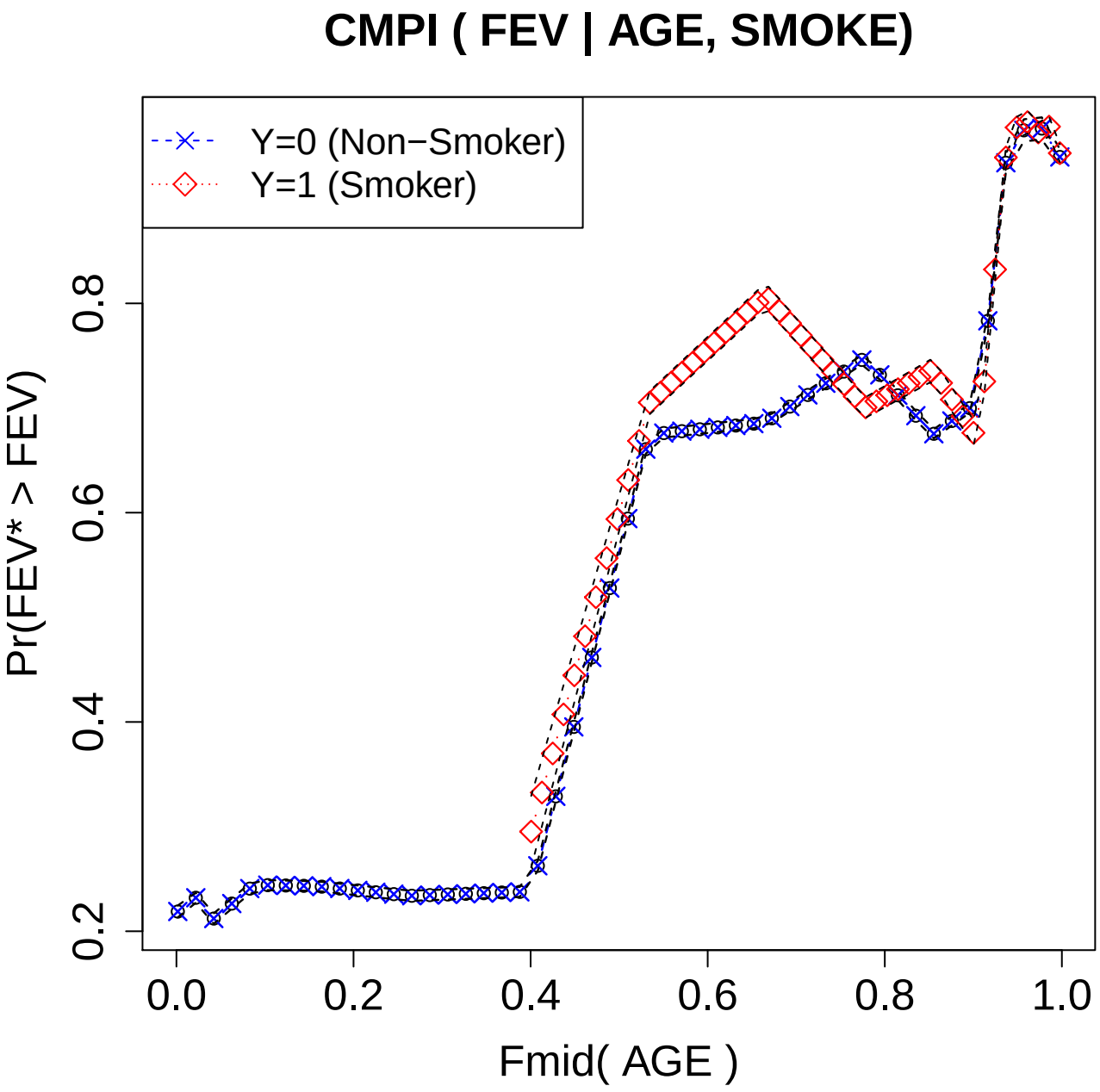


Figure 9: CMPI(FEV | AGE, SMOKE). Comparing Nonparametric conditional mean of F^{mid}(FEV) over two class, smoker and non-smoker. Y-axis represent Pr(FEV* > FEV) as a function of mid-rank of Age.