Mixture models with symmetric errors

Cristina Butucea[1]

[1]Université Paris-Est Marne-la-Vallée

Van Dantzig Seminar, Delft 2016

Outline

1. Introduction

B. and Vandekerkhove (2013) *Scandinavian J. Statist.*

Mixture of probability densities:

$$X \sim \begin{cases} f(\cdot - a), & \text{with probability } p \\ f(\cdot - b), & \text{with probability } 1-p \end{cases},$$

where f is a probability density, symmetric around 0.

X has probability density

$$g = p \cdot f(\cdot - a) + (1-p) \cdot f(\cdot - b).$$

Identifiability and estimation results:
[1] Bordes, Mottelet, Vandekerkhove (2006) Ann. Statist.
[2] Hunter, Wang, Hettmansperger (2006) Ann. Statist.

B., Ngueyep Tzoumpe, Vandekerkhove (2015) *Bernoulli, on line*

Mixture of regression models

$$Y = \begin{cases} a(X) + \varepsilon, & \text{with probability } \pi(X) \\ b(X) + \varepsilon, & \text{with probability } 1 - \pi(X) \end{cases},$$

where ε centered with symmetric probability density (conditional on X).

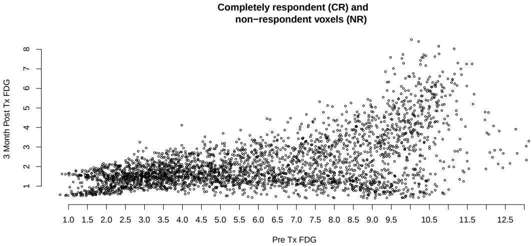


Figure : Display of the original PET-radiotherapy data from Bowen et al. (2012)

2. Mixture of symmetric probability densities

We observe $X_1, \ldots, X_n$ i.i.d. having common probability density function (p.d.f.)

$$g(x) = p \cdot f(x-a) + (1-p) \cdot f(x-b), \quad x \in \mathbb{R},$$

where $p \in (0,1)$, a, $b \in \mathbb{R}$ and $f : \mathbb{R} \to \mathbb{R}_+$ is a symmetric p.d.f. around 0 axis.

Denote by $\theta = (p,\ a,\ b)$ the scalar parameter.

Goal: recover θ and the function f in our semiparametric model, from $g = K_\theta f$.

Previously,
[1] Bordes, Mottelet, Vandekerkhove (2006) Ann. Statist.
[2] Hunter, Wang, Hettmansperger (2006) Ann. Statist.
work with c.d.f

$$G_{\theta,F}(x) = p \cdot F(x-a) + (1-p) \cdot F(x-b).$$

Write $G = K_\theta F$, i.e. inverse problem with partially known operator.

Identifiability

They prove the **identifiability** of θ over the set $\Theta = [0, \frac{1}{2}) \times \mathbb{R}^2 \backslash \Delta$, $\Delta = \{(x, x) : x \in \mathbb{R}\}$ and F c.d.f. of a symmetric distribution: If

$$G_{\theta_1, F_1}(x) = G_{\theta_2, F_2}(x), x \in \mathbb{R}$$

then $\theta_1 = \theta_2$ and $F_1 \equiv F_2$.

Iterative inversion procedure:
Recall that $G_{\theta,F}(x + b) = p \cdot F(x - a + b) + (1 - p) \cdot F(x)$ giving

$$\begin{aligned}
F(x) &= \frac{1}{1-p} G(x+b) - \frac{p}{1-p} F(x-a+b) \\
&= \frac{1}{1-p} G(x+b) - \frac{p}{(1-p)^2} G(x-a+2b) + \frac{p^2}{(1-p)^2} F(x-2a+2b) \\
&= \ldots =: K_\theta^{-1} G(x),
\end{aligned}$$

for $\frac{p}{1-p} < 1$.

Key fact: F is the cdf of a symmetric distribution

iff

$F(x) = SF(x)$, for all x, with $SF(x) = 1 - F(-x)$.

Estimation is based on the fact that

$$G(x) = [K_t S K_t^{-1} G](x), \text{ for all } x, \text{ iff } t = \theta.$$

Therefore, a contrast can be build

$$T(t) = \int (G - K_t S K_t^{-1} G)^2 dG$$

and

$$\theta = \arg\inf_t \int (G - K_t S K_t^{-1} G)^2 dG.$$

The procedure:
-truncate the iterative algorithm at N, $K_{t,N}^{-1}$;
-estimate the contrast

$$T_N(t) = \int (G - K_{t,N} S K_{t,N}^{-1} G)^2 dG;$$

by $\hat{T}_{N,n}$;
-minimize that estimator to get $\hat{\theta}$:

$$\hat{\theta} = \arg\inf_t \hat{T}_{N,n}(t).$$

Main results: $n^{1/4-\alpha}(\hat{\theta} - \theta) = o(1)$ a.s. and $\|\hat{F}_n - F\|_\infty = o_{a.s.}(n^{-1/4+\alpha})$, for some $\alpha > 0$.

Contrast function

We go to Fourier domain: $f^*(u) = \int_{\mathbb{R}} e^{iux} f(x) dx$.

Key fact: f symmetric iff $f^* \in \mathbb{R}$ iff $Im(f^*) \equiv 0$.

We have,

$$g^*(u) = pe^{iua} f^*(u) + (1-p)e^{iub} f^*(u) = M(\theta, u) \cdot f^*(u),$$

where $M(\theta, u) = pe^{iua} + (1-p)e^{iub}$.

We suppose that $0 < P_* \leq p \leq P^* < \frac{1}{2}$ and then

$$0 < 1 - 2P^* \leq |M(\theta, u)| \leq 1, \text{ for all } u.$$

Thus, our inverse problem is well-posed! The exact inversion goes:

$$g \longrightarrow g^* \longrightarrow f^* = \frac{g^*}{M(\theta,\cdot)} \longrightarrow f.$$

Equivalently, $f = \mathcal{F}^{-1}\left[\frac{\mathcal{F}[g](u)}{M(\theta,u)}\right]$.

Our procedure:

-define a new contrast $S(t)$ based on the Fourier transform;

-estimate it by $\hat{S}_n(t)$ at parametric rate;

-minimize it to get $\hat{\theta}_n = \arg\inf_t \hat{S}_n(t)$

-estimate f by a deconvolution-type estimator that uses $\hat{\theta}_n$.

If f is a symmetric pdf such that f^* belongs to $\mathbb{L}_1$ and $\mathbb{L}_2$ and if θ belongs to T a compact set included in Θ, then

$$Im \frac{g^*(u)}{M(t,u)} = 0, \text{ for all } u, \quad \text{iff} \quad t = \theta.$$

We build the contrast function

$$S(t) = \int_{\mathbb{R}} \left(Im \frac{g^*(u)}{M(t,u)} \right)^2 dW(u), \quad t \in \mathbb{R}$$

where W is the cdf of a continuous distribution with finite 3rd order moments.

Rk: W helps computing the integrals with Monte-Carlo AND allows less restrictive assumptions on f.

Proposition: $S(t) \geq 0$ for all t and $S(t) = 0$ iff $t = \theta$.

Estimators

Estimation of the contrast function:

$$S(t) = -\frac{1}{4} \int_{\mathbb{R}} \left(\frac{g^*(u)}{M(t,u)} - \frac{\overline{g^*}(u)}{M(t,-u)} \right)^2 dW(u).$$

Recall that $g^*(u) = E(e^{iuX})$ and put

$$Z_k(t,u) = \frac{e^{iuX_k}}{M(t,u)} - \frac{e^{-iuX_k}}{M(t,-u)}.$$

Thus

$$\hat{S}_n(t) = -\frac{1}{4n(n-1)} \sum_{k \neq j} \int Z_k(t,u) Z_j(t,u) dW(u).$$

Rk: do not use the plug-in estimator!
Our estimator of θ is

$$\hat{\theta}_n = \arg\min_t \hat{S}_n(t).$$

Estimator of f by kernel-deconvolution like procedure:

$$\hat{f}_n^*(u) = \frac{1}{n} \sum_{k=1}^{n} \frac{e^{iuX_k} K^*(uh_n)}{M(\hat{\theta}_{n,-k}, u)}$$

where K is a kernel and $\hat{\theta}_{n,-k}$ is the previous leave-one-out estimator of θ.

Theorem: If $W : \mathbb{R} \to \mathbb{R}^+$ is a continuous cdf such that $\int |u|^3 dW(u) < \infty$ then

$$\hat{\theta}_n \to \theta, \text{ in probability}$$

and

$$\sqrt{n}(\hat{\theta}_n - \theta) \to N(0, \Sigma), \text{ in distribution,}$$

where Σ is an explicit covariance matrix depending on θ and on W.

Rk: loss of asymptotic efficacity due to W, but less "expensive" assumptions on f.

If, moreover, f belongs to a Sobolev class $S(\beta, L)$ with smoothness $\beta > 1/2$ and $L > 0$, then $\hat{f}_n$ with $h = cn^{-1/(2\beta)}$, $c > 0$ and K symmetric kernel in $\mathbb{L}_1$ and $\mathbb{L}_2$, such that $supp(K^*) \subset [-1, 1]$, then

$$\lim\sup_{n\to\infty} \sup_{f\in S(\beta,L)} \sup_{\theta\in T} E_{\theta,f}[n^{-\frac{2\beta-1}{2\beta}} |\hat{f}_n(x) - f(x)|^2] \leq C^*,$$

where $C^* = C^*(\beta, L, P^*, \int K^2)$. Moreover,

$$\lim\inf_{n\to\infty} \inf_{\tilde{f}_n} \sup_{f\in S(\beta,L)} \sup_{\theta\in T} E_{\theta,f}[n^{-\frac{2\beta-1}{2\beta}} |\tilde{f}_n(x) - f(x)|^2] \geq C_* > 0,$$

where the infimum is taken over all estimators $\tilde{f}_n$ of f.

Rk. the nonparametric rates are those in the direct problem and the lower bounds are directly deduced from there.
Rk. the well-posed inverse problem implies that there is no loss in the nonparametric rate.

Gassiat, Rousseau (2013, arxiv:1302.2345)

$$Y_i = a_{S_i} + \varepsilon_i,$$

where S_i's take values $\{1, ..., K\}$ with probabilities $p_1, ..., p_K$ and are dependent.
From marginal bi-variate distributions $((Y_1, Y_2))$, identifiability and estimation of

$$K, \quad a_1, ..., a_K \text{ and } p_1, ..., p_K,$$

under some assumptions.

Here, these assumptions are not verified!
If $K > 2$, our method provides an estimator, but no identifiability results are known for $K > 3$, sufficient conditions are known for $K = 3$ (Bordes et al., 2006).
Balabdaoui and B. (2014) identifiability of mixture of probability densities that are Pólya functions.

In the multivariate case (a, $b \in \mathbb{R}^d$), it is sufficient to use the marginal densities in order to identify and estimate θ.

3. Mixture of regression functions with symmetric errors

B., Ngueyep Tzoumpe, Vandekerkhove (2015) *Bernoulli,* to appear
$(X_1, Y_1), ..., (X_n, Y_n)$ i.i.d. such that

$$Y_i = \begin{cases} a(X_i) + \varepsilon_i, & \text{with probability } \pi(X_i) \\ b(X_i) + \varepsilon_i, & \text{with probability } 1 - \pi(X_i) \end{cases},$$

where ε_i i.i.d., centered with symmetric conditional probability density.

The conditional probability density of $Y/X = x$ is

$$g_x(y) = \pi(x) f_x(y - a(x)) + (1 - \pi(x)) f_x(y - b(x)),$$

where $\varepsilon/X = x$ has symmetric probability density f_x for all x.

Rk. We can also apply the method to
1) $\sigma(X_i)\varepsilon_i$, i.e. $f_x(y) = \frac{1}{\sigma(x)} f\left(\frac{y}{\sigma(x)}\right)$;
2) $f_x(y) = \sum_{k=1}^{K} \lambda_k(x) f_k(x)$, f_k is symmetric for all k.

Local and global identifiability

For given x_0 in $supp(\ell)$ (pdf of X_i, $i = 1, ..., n$), we want to estimate $\theta(x_0) = (\pi(x_0), a(x_0), b(x_0))$ and f_{x_0}.

Local indentifiability for fixed x_0;
revisit the proof by Bordes et al. to get it on the set $[P_*, P^*] \subset (0, 1)$ and a compact set in $(x, y) : x < y$.
So, no restriction to $\pi(x_0) < 1/2$! Label switching to get $a(x_0) < b(x_0)$.

Global identifiability We assume the curves a and b are transversal, following Huang, Li, Wang (2013) JASA.
Suppose a, b are $\mathcal{C}^1$ such that

$$(a(x) - b(x))^2 + \|\dot{a}(x) - \dot{b}(x)\|^2 \neq 0, \text{ for all } x.$$

Contrast function

In Fourier domain, $g_x^*(u) = M(\theta(x), u) f_x^*(u)$, for all u.

The new **contrast** is based on the fact that

$$Im(g_x^*(u) \cdot \bar{M}(t, u)) = 0, \text{ for all real number } u \text{ iff } t = \theta(x).$$

Contrast function

$$S(t) = \int Im^2(g_x^*(u) \cdot \bar{M}(t, u)) \cdot \ell^2(x) dW(u),$$

for x in $supp(\ell)$.

We write

$$S(t) = -\frac{1}{4} \int \left(g_x^*(u) \cdot \bar{M}(t, u) - \bar{g_x^*}(u) \cdot M(t, u)\right)^2 \ell^2(x) dW(u).$$

Smoothing is needed. We choose kernel smoothing!

We put

$$Z_{k,x}(t,u,h) = \left(e^{iuY_k} \bar{M}(t,u) - e^{-iuY_k} M(t,u) \right) \frac{1}{h} K\left(\frac{X_k - x}{h} \right)$$

and

$$S_n(t) = -\frac{1}{4n(n-1)} \sum_{k \neq j} \int Z_{k,x}(t,u,h) Z_{j,x}(t,u,h) dW(u),$$

and

$$\hat{\theta}_n = \arg\inf_t S_n(t).$$

Nonparametric rates for estimating S will follow for θ.

Kernel estimator for f using $\hat{\theta}_n$ - under the assumptions of the former paper ($\pi < 1/2$).

Results

A1. We assume that the functions π, a, b, ℓ belong to a Hölder smoothness class $L(\alpha, M)$ with α, $M > 0$.

A2. Assume that $f_{\mathbf{x}}(\cdot) \in \mathbb{L}_1 \cap \mathbb{L}_2$ for all $\mathbf{x} \in \mathbb{R}^d$. In addition, we require that there exists a w-integrable function φ such that

$$|f^*_{\mathbf{x}}(u) - f^*_{\mathbf{x}'}(u)| \leq \varphi(u)\|\mathbf{x} - \mathbf{x}'\|^\alpha, \quad (\mathbf{x}, \mathbf{x}') \in \mathbb{R}^d \times \mathbb{R}^d, \; u \in \mathbb{R}.$$

Remark. Note that for the scaling model, if f is the $\mathcal{N}(0,1)$ p.d.f. and $\sigma(\cdot)$ is bounded and Hölder α-smooth, we have:

$$|f^*_{\mathbf{x}}(u) - f^*_{\mathbf{x}'}(u)| \leq \frac{u^2}{2}|\sigma^2(\mathbf{x}) - \sigma^2(\mathbf{x}')| \leq C\frac{u^2}{2}\|\mathbf{x} - \mathbf{x}'\|^\alpha.$$

A3. We assume that the kernel K is such that $\int |K| < \infty$, $\int K^4 < \infty$ and that it satisfies also the moment condition

$$\int \|\mathbf{x}\|^\alpha |K(\mathbf{x})| d\mathbf{x} < \infty.$$

A4. The weight function w is a p.d.f. such that

$$\int (u^4 + \varphi(u)) w(u) du < \infty.$$

The following results will hold true under the additional assumption on the kernel (see **A3**): $\int \mathbf{x}^j K(\mathbf{x}) d\mathbf{x} = 0$, for all j such that $|j| \leq k$.

Proposition For each $t \in \Theta$ and $\mathbf{x}_0 \in supp(\ell)$ fixed, suppose $\theta_0 \in \overset{\circ}{\Theta}$ and that assumptions **A1-A4** hold. Then, the empirical contrast function $S_n(\cdot)$ satisfies

$$E\left[(S_n(t) - S(t))^2\right] \leq C_1 h^{2\alpha} + C_2 \frac{1}{nh^d},$$

if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$, where constants C_1, C_2 depend on Θ, K, w, α and M but are free from n, h, t and $\mathbf{x}_0$.

Theorem (Consistency) Let suppose that assumptions of the previous Proposition hold. The estimator $\hat{\theta}_n$ converges in probability to $\theta(\mathbf{x}_0) = \theta_0$ if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.

In the asymptotic variance we will use the following notation:

$$\dot{J}(\theta_0, u) := Im\left(-\dot{M}(\theta_0, u)\bar{M}(\theta_0, u)\right) f^*_{\mathbf{x}_0}(u)\ell(\mathbf{x}_0), \quad (1)$$

and

$$V(\theta_0, u_1, u_2) := 4 \cdot \int Im\left(e^{iu_1 y}\bar{M}(\theta_0, u_1)\right) \cdot Im\left(e^{iu_2 y}\bar{M}(\theta_0, u_2)\right) g_{\mathbf{x}_0}(y)dy. \quad (2)$$

Theorem (Asymptotic normality) Suppose that assumptions of the Proposition hold. The estimator $\hat{\theta}_n$ of θ_0, with $h \to 0$ such that $nh^d \to \infty$ and such that $h^{2\alpha+d} = o(n^{-1})$, as $n \to \infty$, is asymptotically normally distributed:

$$\sqrt{nh^d}(\hat{\theta}_n - \theta_0) \to N(0, \mathcal{S}) \quad \text{in distribution,}$$

where $\mathcal{S} = \frac{1}{4}\mathcal{I}^{-1}\Sigma\mathcal{I}$, with

$$\mathcal{I} = -\frac{1}{2}\int \dot{J}(\theta_0, u)\dot{J}(\theta_0, u)^\top w(u)du,$$

and

$$\Sigma := \int\int \dot{J}(\theta_0, u_1)\dot{J}^\top(\theta_0, u_2)V(\theta_0, u_1, u_2)w(u_1)w(u_2)du_1du_2,$$

for $\dot{J}$ defined in (1) and V in (2).

Theorem (Minimax rates) Suppose **A1**-**A4** and consider $\mathbf{x}_0 \in supp(\ell)$ fixed such that $\ell(\mathbf{x}_0) \geq L_* > 0$ for all $\ell \in L(\alpha, M)$ and $\theta_0 = \theta(\mathbf{x}_0) \in \overset{\circ}{\Theta} \setminus \{1/2\}$. The estimator $\hat{\theta}_n$ of θ_0, with $h \asymp n^{-1/(2\alpha+d)}$, as $n \to \infty$, is such that

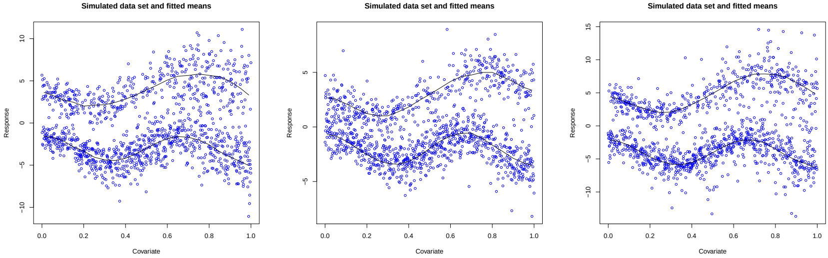
$$\sup E[\|\hat{\theta}_n - \theta_0\|^2] \leq C n^{-\frac{2\alpha}{2\alpha+d}},$$

where the supremum is taken over all the functions π, a, b, ℓ and f^* checking assumptions **A1**-**A2**. Moreover,

$$\inf_{T_n} \sup E[\| T_n - \theta_0\|^2] \geq c n^{-\frac{2\alpha}{2\alpha+d}},$$

where C, $c > 0$ depend only on α, M, Θ, K and w, and the infimum is taken over the set of all the estimators T_n (measurable function of the observations $(X_1, \ldots, X_n)$) of θ_0.

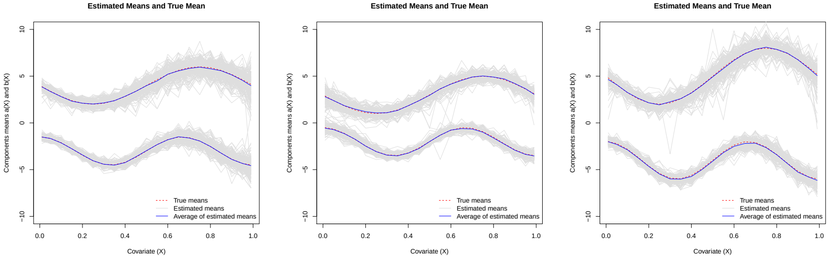
4.1 Synthetic data



(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

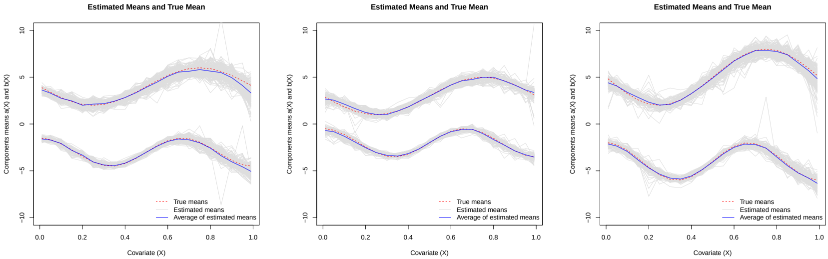
Figure : Examples of a simulated dataset of size 1200 with different distribution errors

Estimators by Huang, Li, Wang (2013) under the assumption of Gaussian errors:



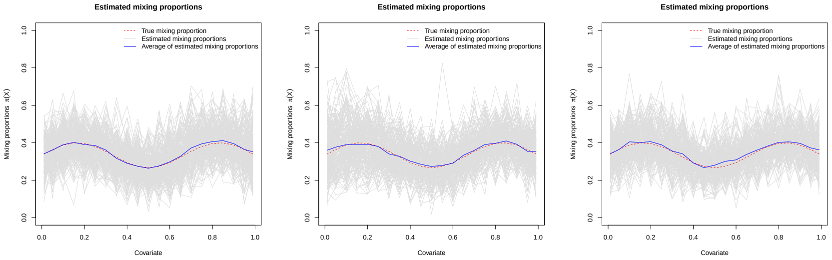
(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

Figure : Mean Curves estimated with NMRG (100 repeated samples)



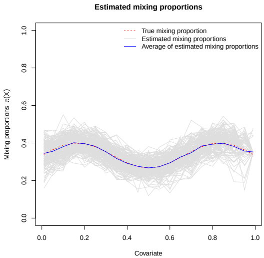
(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

Figure : Mean Curves estimated with NMR-SE (100 repeated samples)

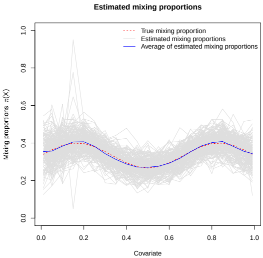


(a) Gaussian distribution (b) Student distribution (c) Laplace distribution

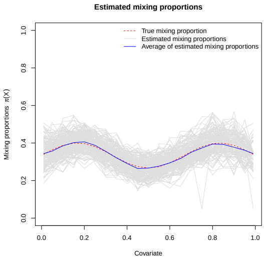
Figure : Mixing proportions estimated with NMRG



(a) Gaussian distribution



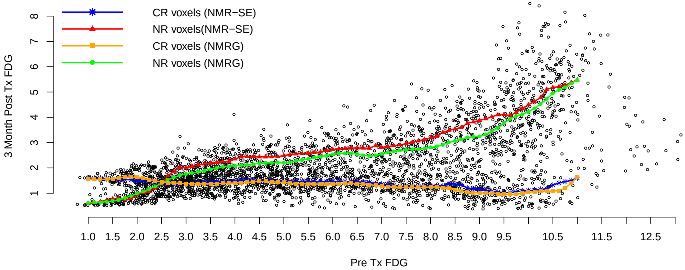
(b) Student distribution



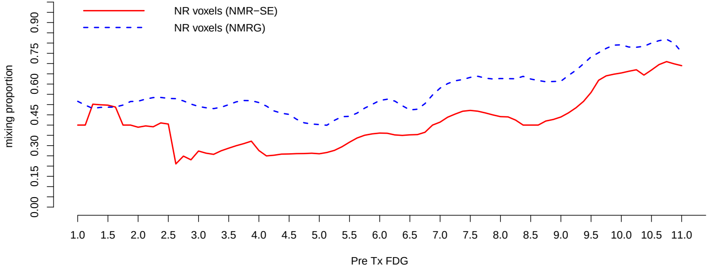
(c) Laplace distribution

Figure : Mixing proportions curves estimated with NMR-SE

4.2 Real data



(a) Scatter of plots of pre-treatment FDG PET vs. post-treatment FDG PET and estimated location functions for the completely respondent and non-respondent voxel subpopulations



(a) Estimated mixing proportions for the completely (CR) and non-respondent (NR) voxel subpopulation

Figure : Location and mixing proportion function estimation by using NMR-SE and NMRG methods

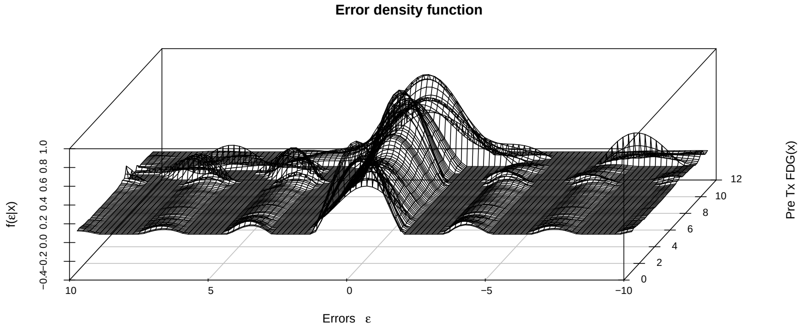


Figure : Density Estimates of the errors for the different levels of PET Tx FDG values