Mixture models with symmetric errors

Cristina Butucea

Université Paris-Est Marne-la-Vallée

Van Dantzig Seminar, Delft 2016
Outline

1 Introduction

2 Mixture of symmetric probability densities
   - Model and Identifiability
   - Contrast function
   - Estimators and Results

3 Mixture of regression functions with symmetric errors
   - Model
   - Contrast function
   - Estimators and Results

4 Numerical performance
   - Synthetic data
   - Real data
1. Introduction


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:
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 \( \varepsilon \) 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 $\theta$ and the function $f$ in our semiparametric model, from $g = K_\theta f$.

Previously,


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.
They prove the **identifiability** of $\theta$ over the set $\Theta = [0, \frac{1}{2}) \times \mathbb{R}^2 \setminus \Delta$, $\Delta = \{(x, x) : x \in \mathbb{R}\}$ 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

$$F(x) = \frac{1}{1 - p} G(x + b) - \frac{p}{1 - p} F(x - a + b)$$

$$= \frac{1}{1 - p} G(x + b) - \frac{p}{(1 - p)^2} G(x - a + 2b) + \frac{p^2}{(1 - p)^2} F(x - 2a + 2b)$$

$$= \ldots =: K_{\theta}^{-1} G(x),$$

for $\frac{p}{1 - p} < 1$. 
**Key fact:** $F$ is the cdf of a symmetric distribution iff

$$F(x) = SF(x), \text{ for all } x, \text{ with } SF(x) = 1 - F(-x).$$

Estimation is based on the fact that

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

Therefore, a contrast can be build

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

and

$$\theta = \arg \inf_t \int (G - K_tSK_t^{-1}G)^2 dG.$$
The procedure:
- truncate the iterative algorithm at $N, K_{t,N}^{-1}$;
- estimate the contrast

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

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 $\text{Im}(f^*) \equiv 0$.

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
$$0 < 1 - 2P^* \leq |M(\theta, u)| \leq 1,$$ for all $u$. 
Thus, our inverse problem is well-posed! The exact inversion goes:

\[
g \rightarrow g^* \rightarrow f^* = \frac{g^*}{M(\theta, \cdot)} \rightarrow f.
\]

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 $L_1$ and $L_2$ and if $\theta$ belongs to $T$ a compact set included in $\Theta$, then

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

We build the contrast function

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

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:

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

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 \( \theta \) 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 $\theta$.

**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 $\Sigma$ is an explicit covariance matrix depending on $\theta$ and on $W$.

Rk: loss of asymptotic efficacy due to $W$, but less ”expensive” assumptions on $f$. 

C. Butucea (Université Paris-Est Marne-la-Vallée)
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 \( L_1 \) and \( L_2 \), such that \( \text{supp}(K^*) \subset [-1, 1] \), then

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

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

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

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 \( \{1, \ldots, K\} \) with probabilities \( p_1, \ldots, p_K \) and are dependent.

From marginal bi-variate distributions \( ((Y_1, Y_2)) \), identifiability and estimation of \( K, \ a_1, \ldots, a_K \) and \( p_1, \ldots, p_K \), 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 \( \theta \).
3. Mixture of regression functions with symmetric errors


$(X_1, Y_1), ..., (X_n, Y_n)$ 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 $\varepsilon_i$ 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 $\text{supp}(\ell)$ (pdf of $X_i$, $i = 1, \ldots, n$), we want to estimate $
(0, x_0) = (\pi(x_0), a(x_0), b(x_0))$ and $f_{x_0}$.

**Local identifiability** 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 $C^1$ 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

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

Contrast function

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

for \( x \) in \( \text{supp}(\ell) \).

We write

\[
S(t) = -\frac{1}{4} \int (g_x^*(u) \cdot \bar{M}(t, u) - \bar{g}_x^*(u) \cdot M(t, u))^2 \ell^2(x) dW(u).
\]

Smoothing is needed. We choose kernel smoothing!
We put
\[ Z_{k,x}(t, u, h) = \left( e^{iuY_k} \tilde{M}(t, u) - e^{-iuY_k} M(t, u) \right) \frac{1}{h} K \left( \frac{X_k - x}{h} \right) \]
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 \( \theta \).

Kernel estimator for \( f \) using \( \hat{\theta}_n \) - under the assumptions of the former paper (\( \pi < 1/2 \)).
A1. We assume that the functions $\pi, a, b, \ell$ belong to a Hölder smoothness class $L(\alpha, M)$ with $\alpha, M > 0$.

A2. Assume that $f_x(\cdot) \in L_1 \cap L_2$ for all $x \in \mathbb{R}^d$. In addition, we require that there exists a $w$-integrable function $\varphi$ such that

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

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 $\alpha$-smooth, we have:

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

$$\int \|x\|^\alpha |K(x)| dx < \infty.$$ 

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): $\int x^j K(x) dx = 0$, for all $j$ such that $|j| \leq k$.

**Proposition** For each $t \in \Theta$ and $x_0 \in \text{supp}(\ell)$ fixed, suppose $\theta_0 \in \hat{\Theta}$ 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 $\Theta, K, w, \alpha$ and $M$ but are free from $n, h, t$ and $x_0$. 
**Theorem (Consistency)** Let suppose that assumptions of the previous Proposition hold. The estimator $\hat{\theta}_n$ converges in probability to $\theta(x_0) = \theta_0$ if $h \to 0$ and $nh^d \to \infty$ as $n \to \infty$.

In the asymptotic variance we will use the following notation:

$$
\dot{J}(\theta_0, u) := \text{Im} \left( -\dot{M}(\theta_0, u) \bar{M}(\theta_0, u) \right) f_{x_0}^*(u) \ell(x_0),
$$

(1)

and

$$
V(\theta_0, u_1, u_2) := 4 \cdot \int \text{Im} \left( e^{iu_1 y} \bar{M}(\theta_0, u_1) \right) \cdot \text{Im} \left( e^{iu_2 y} \bar{M}(\theta_0, u_2) \right) g_{x_0}(y) dy.
$$

(2)
**Theorem (Asymptotic normality)** Suppose that assumptions of the Proposition hold. The estimator \( \hat{\theta}_n \) of \( \theta_0 \), 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:

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

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

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

and

\[
\Sigma := \int \int J(\theta_0, u_1) J^\top(\theta_0, u_2) V(\theta_0, u_1, u_2) w(u_1) w(u_2) du_1 du_2,
\]

for \( J \) defined in (1) and \( V \) in (2).
**Theorem (Minimax rates)** Suppose A1-A4 and consider $x_0 \in \text{supp}(\ell)$ fixed such that 

\[ \ell(x_0) \geq L_* > 0 \] 

for all $\ell \in L(\alpha, M)$ and $\theta_0 = \theta(x_0) \in \Theta \setminus \{1/2\}$. The estimator $\hat{\theta}_n$ of $\theta_0$, with $h \asymp n^{-1/(2\alpha+d)}$, as $n \to \infty$, is such that 

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

where the supremum is taken over all the functions $\pi, a, b, \ell$ and $f^*$ checking assumptions A1-A2. Moreover,

\[ \inf \sup E[\|T_n - \theta_0\|^2] \geq cn^{-\frac{2\alpha}{2\alpha+d}}, \]

where $C, c > 0$ depend only on $\alpha, M, \Theta, K$ 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 $\theta_0$. 

---

C. Butucea (Université Paris-Est Marne-la-Vallée)  
Mixture models with symmetric errors  
Van Dantzig Seminar, Delft 2016  
23 / 31
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)
Estimated Means and True Mean

Components means $a(X)$ and $b(X)$

True means
Estimated means
Average of estimated means

(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
Estimated mixing proportions

- True mixing proportion
- Estimated mixing proportions
- Average of estimated mixing proportions

(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