Monografías del Seminario Matemático García de Galdeano **33**, 317–323 (2006)

ORTHOGONAL SERIES DENSITY ESTIMATION IN MIXTURE MODELS

Denys Pommeret

Abstract. This paper concerns estimation of mixture densities. It is the continuation of the work of Pommeret [5] on mixture models in two directions: first we consider orthogonal series density estimates within the frame of a wide class of mixture models. Second, we illustrate the methods using real data sets and we compare them with other density estimators.

Keywords: L^1 and L^2 norms, mean integrated square error, mixture density, orthogonal polynomials.

AMS classification: 62E17.

§1. Introduction

Mixture models play an important role in many statistical problems such as for example in biostatistics, psychometry or econometrics. Many works have been devoted to estimation of the density: in finite mixture, nonparametric maximum likelihood estimation (see [4]) give estimates for both mixing distribution and number of components. In parametric models, when number of components is fixed, the EM algorithm introduced by Dempster *et al.* [2] has been widely used and extended in the literature. Recently, in [5], polynomials as density estimates have been proposed within the frame of mixture exponential families. The aim of this paper is to extend this work to any distributions family such that densities admit an orthogonal series expansion. Moreover, we are interested in applying the method to real data sets.

We consider the following modelisation: let $X_1,\dots,X_q$ be an i.i.d. sample from a mixture with density

$$g(x)=\begin{cases}\int f(x,\theta)\Pi(d\theta), & \text{continuous mixture},\\ \sum f(x,\theta)\Pi(\theta), & \text{discrete mixture},\end{cases} \tag{1}$$

where Π is a probability distribution and $\{f(.,\theta)\,;\,\theta\in\Theta\}$ are parent mixed density functions. We are interesting in finding an orthogonal series estimate for g. We assume that f and g are densities with respect to a common known measure μ and we denote by $B=\{P_n;n\in\mathbb{N}\}$ an orthogonal basis of $L^2(\mu)$. Then the two expansions of g and f in the basis B permit the construction of a kth order estimate of g. An expression of the distance between g and its estimate in L^1 and L^2 norms is given. It provides also bounds for the difference between the mixture distribution function and its estimate. This method is illustrated through few data set: Notice data ([9]), Sibship data ([8]), Bortkewisch's data ([10]) and Accident data ([7]).

The paper is organized as follows: in Section 2 we introduce orthogonal series. In Section 3 we derive an expression of the density estimator. In Section 4 we briefly treat error bounds. Section 5 is devoted to illustrations when the mixing density, Π, is unknown.

§2. General expansions

We assume now that $f(.,\theta)$ and g are densities with respect to a given probability measure μ. We write $\{P_n; n \in \mathbb{N}\}$ a basis of orthonormal polynomials with respect to μ; that is,

$$\int P_n(x)P_k(x)\mu(dx) = \begin{cases} 1, & \text{if } n = k, \\ 0, & \text{if } n \neq k. \end{cases}$$

It is required that the basis B is dense in $L^2(\mu)$. We assume that all densities $f(.,\theta)$ are square integrable with respect to μ; that is, $\int f(x,\theta)^2\mu(dx) < \infty$. Thus we can write

$$f(x,\theta) = \sum_{n\in\mathbb{N}} a_n(\theta)P_n(x), \tag{2}$$

where $a_n(\theta) = \int f(x,\theta)P_n(x)\mu(dx)$. We will write it simply a_n instead of $a_n(\theta)$ when no confusion can arise. We will denote by $\mathbb{E}_\Pi$ the expectation with respect to Π. By abuse of notation we write $\mathbb{E}_\Pi(a_n)$ instead of $\mathbb{E}_\Pi(a_n(\theta))$. Let us mention an important consequence of (2):

Lemma 1. *Let g be a mixture density defined by (1). If the series $\sum_{n\in\mathbb{N}} \mathbb{E}_\Pi(a_n)|P_n(x)|$ converges then we have*

$$g(x) = \sum_{n\in\mathbb{N}} \mathbb{E}_\Pi(a_n)P_n(x). \tag{3}$$

Proof. Combining (1) with (2) we obtain

$$g(x) = \int \sum_{n\in\mathbb{N}} a_nP_n(x)\Pi(dm) = \sum_{n\in\mathbb{N}} \mathbb{E}_\Pi(a_n)P_n(x). \quad \square$$

As a direct consequence we have the following expansion of the difference between the mixture density and its parent:

Proposition 2. *Fix $\theta_0 \in \Theta$. If the series $\sum_{n\in\mathbb{N}} \mathbb{E}_\Pi(a_n)|P_n(x)|$ converges, then we have*

$$g(x) - f(x,\theta_0) = \sum_{n\geq 1} (\mathbb{E}_\Pi(a_n) - a_n(\theta_0))P_n(x).$$

§3. Estimation

From the expansion given in Proposition 2 we may deduce kth order approximations of the mixture density g, namely

$$g^{[k]}(x,\theta_0) = f(x,\theta_0) + \sum_{1\leq n\leq k} \{\mathbb{E}_\pi(a_n) - a_n(\theta_0)\}P_n(x),$$

for some fixed $\theta_0 \in \Theta$. Note that, for $k = 0$, $g^{[0]}(x,\theta_0) = f(x,\theta_0)$.

In practice, statisticians may be confronted with a mixture of known mixed distributions f but with unknown mixing distribution Π. Then, estimating the quantities $\mathbb{E}_\Pi(a_n)$ and reporting them in $g^{[k]}$ yields an estimated approximation, say $\widehat{g}^{[k]}$, of the mixture density. We have

$$\widehat{g}^{[k]}(x,\theta_0) = f(x,\theta_0) + \sum_{1\le n\le k} \{\widehat{\mathbb{E}_\pi(a_n)} - a_n(\theta_0)\} P_n(x),$$

where $\widehat{\mathbb{E}_\pi(a_n)}$ are convergent empirical estimates. Note that a_n is deduce from the knowledge of f. In that case, the mean integrated square error (MISE) may be used to evaluate the quality of this approximation. The MISE is defined by

$$\text{MISE} = \mathbb{E}\{\|g(.) - \widehat{g}^{[k]}(.,\theta_0)\|^2\}.$$

We have the following property:

Proposition 3. *If $\widehat{\mathbb{E}_\pi(a_n)}$ are convergent estimates of $\mathbb{E}_\pi(a_n)$ then the* MISE *tends to $\|g(.) - g^{[k]}(.,M)\|^2$, a.s.*

Proof.

$$\text{MISE} = \|g(.) - g^{[k]}(.,M)\|^2 + \sum_{n=1}^{k} (\mathbb{E}_\Pi(a_n) - \widehat{\mathbb{E}_\Pi(a_n)})^2,$$

and the MISE tends to $\|g(.) - g^{[k]}(.,M)\|^2$ at the same rate as $\widehat{\mathbb{E}_\Pi(a_n)}$. $\square$

Example 1 (Poisson mixture)**.** If $f(x,\theta) = \exp(-\theta)\theta^x/x!$, then, writing $\theta_0 = \mathbb{E}_\Pi(\theta)$, we have (see [5]):

$$g^{[2]}(x,\theta_0) = f(x,\theta_0) + \frac{(x-\theta_0)^2 - x}{2\theta_0^2} \operatorname{Var}(\theta).$$

§4. Error bounds

Write G and $G^{[k]}(.,\theta_0)$ the distribution functions associated to g and $g^{[k]}(.,\theta_0)$ respectively; that is, $G(x) = \int_{-\infty}^{x} g(y)\mu(dy)$ and $G^{[k]}(x,\theta_0) = \int_{-\infty}^{x} g^{[k]}(y,\theta_0)\mu(dy)$. We have a general result:

Proposition 4. *Under the assumptions of Proposition 2 we have:*

$$|G(x) - G^{[k]}(x,\theta_0)| \le \sum_{n\ge k+1} |\mathbb{E}_\Pi(a_n) - a_n(\theta_0)| \int_{-\infty}^{x} |P_n(y)|\mu(dy),$$

$$\left\| g(.) - g^{[k]}(.,\theta_0) \right\|^2 = \sum_{n\ge k+1} (\mathbb{E}_\Pi(a_n) - a_n(\theta_0)^2,$$

$$\int |g(x) - g^{[k]}(x,\theta_0)|\mu(dx) \le \sum_{n\ge k+1} |\mathbb{E}_\Pi(a_n) - a_n(\theta_0)| \int |P_n(x)|\,\mu(dx).$$

Proof. It is immediate from Proposition 2 and from the μ-orthogonality of the polynomials $\{P_n(.)\,;\, n\in\mathbb{N}\}$. $\square$

Number of notices	0	1	2	3	4	5	6	7	8	9
Frequencies	162	267	271	185	111	61	27	8	3	1

Table 1: Notice data

Notices	0	1	2	3	4	5	6	7	8	9
Empirical	0.148	0.244	0.247	0.169	0.101	0.056	0.025	0.007	0.003	0.001
2-component	0.147	0.248	0.239	0.175	0.104	0.052	0.022	0.009	0.003	0.001
Orthogonal	0.142	0.253	0.244	0.172	0.101	0.052	0.023	0.008	0.003	0.001

Table 2: Notice data set.

§5. Illustrations

5.1. Notice data

Data in Table 1 consist of the numbers of death notices for women aged 80 years and over from the *Times Newspaper* for each day between 1910–1912. These data have been analysed by Titterington *et al.* [9]. Assuming a two-component mixture, these authors obtained $\theta_1 = 1.2561$, $\theta_2 = 2.6634$ and $\Pi(\theta_1) = 0.3599$. Using our orthogonal series approach we get:

$$\widehat{g}^{[2]}(x) = \exp(-m)\frac{m^x}{x!}\frac{(1+v((x-m)^2-x))}{2m^2},$$

where $m = 2.157$ and $v = 0.448$. Table 2 gives estimated values from these two methods.

5.2. Sibship data

The data in Table 3 are taken from Sokal and Rohlf [8]. They consist of frequencies of males in 6115 sibship of size 12 in Saxony (1876–85).

The nature of the data set suggests the use of a binomial NEF $(B)_{12}$ as the model. However, Gelfand and Dalal [3] proved that there is a significant overdispersion; that is, the estimated variance significantly exceeds the theoretical one. Such overdispersed data could be fitted better by truncated Poisson model. Figure 1 shows that truncated orthogonal series estimation fit better than Poisson density. We have considered for $x \in \{0,\dots,12\}$:

$$\widehat{g}^{[2]}(x) = C\exp(-\hat{m})\frac{\hat{m}^x}{x!}\frac{(1+\hat{v}((x-\hat{m})^2-x))}{2\hat{m}^2},$$

where C is a constant for normalization.

5.3. Accident data

The data in Table 4 are taken from [7]. They consist of frequencies of accident counts issued by La Royale Belge Insurance Company.

Males	0	1	2	3	4	5	6	7	8	9	10	11	12
Observed	3	24	104	286	670	1033	1343	1112	829	478	181	45	7

Table 3: Sibship data

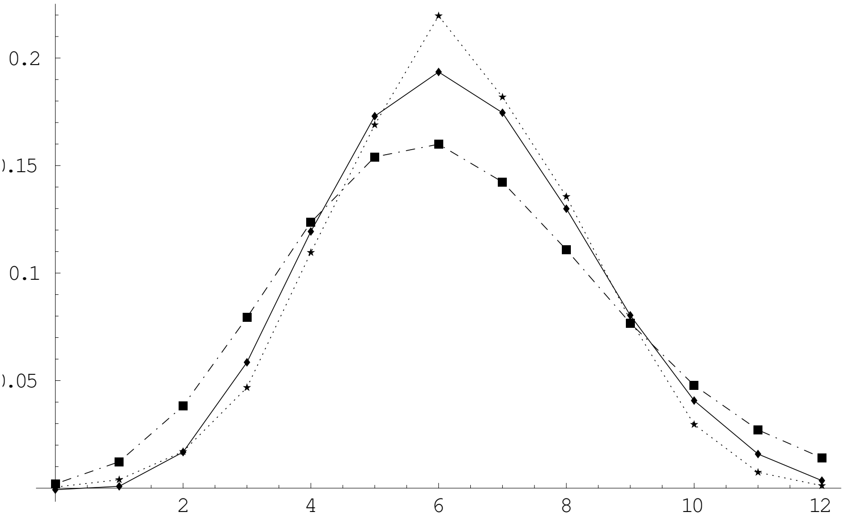

Figure 1: Densities for Sibship data: empirical density (−★−), second order orthogonal series (−♦−) and Poisson approximation (−■−)

The estimate density given by Bohning [1] is a two-component Poisson mixture with additional mass at zero. More precisely the author proposed

$$\widehat{g}(x) = p_0\delta_0(x) + p_1 \exp(-m_1)\frac{m_1^x}{x!} + p_2 \exp(-m_2)\frac{m_2^x}{x!},$$

with $p_0 = 0.42$, $p_2 = 0.57$, $p_3 = 0.009$, $m_1 = 0.34$, $m_2 = 2.55$ and where $\delta_0(x)$ is 1, if $x = 0$, and 0, otherwise. Table 5 compare these results with our method.

5.4. Bortkewitsch data

The data in Table 6 are taken from [10]. They consist of frequencies of Prussian soldiers killed by horse-kicks .

Although these counts are historically associated to the Poisson distribution, Preece *et al.* [6] showed that the negative binomial distribution may be derived as a model for these data. Applying our method with orthogonal series of second order we obtain:

$$\widehat{g}^{[2]}(x) = \exp(-0.7)\frac{(0.7)^x}{x!}\frac{(1+0.06((x-0.7)^2-x))}{0.98},$$

Figure 2 shows that empirical and orthogonal series densities coincide while Poisson one doesn't fit well the data.

Number of accidents	0	1	2	3	4	5	6	7
Observed freq.	7840	1317	239	42	14	4	4	1

Table 4: Accident data set.

Number of accidents	0	1	2	3	4	5	6	7
Empirical prob.	0.828	0.139	0.025	0.004	0.001	0.0004	0.0004	0.0001
Bohning's results	0.830	0.139	0.025	0.004	0.001	0.0006	0.0002	0.00009
Orthogonal series	0.837	0.119	0.036	0.006	0.0006	0.00004	$2.5\ 10^{-6}$	$1.0\ 10^{-7}$

Table 5: Accident data set.

Number of deaths	0	1	2	3	4	5+
Observed freq.	144	91	32	11	2	0

Table 6: Bortkewisch's data set.

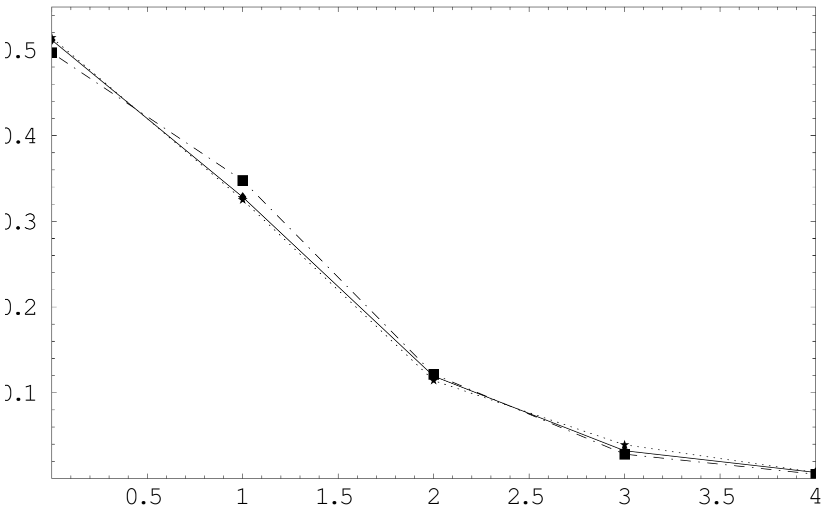

Figure 2: Bortkewitsch data. Empirical density (−★−), orthogonal series estimation (−◆−) and Poisson estimation estimation (−■−)

References

[1] BOHNING, D. *Computer-Assisted Analysis of Mixtures and Applications*. Chapman & Hall, 2000.

[2] DEMPSTER, A. P., LAORD, N. M., AND RUBIN, D. B. Maximum likelihood estimation from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society B, 39* (1977), 1–38.

[3] GELFAND, A. E., AND DALAL, S. R. A note on overdispersed exponential families. *Biometrika 77* (1990), 55–64.

[4] LAIRD, N. M. Nonparametric maximum likelihood estimation of a mixing distribution. *Journal of the American Statistical Association 73* (1978), 805–811.

[5] POMMERET, D. Distance of a mixture form its parent distribution. *Sankhya*, (2006), to appear.

[6] PREECE, D. A., ROSS, G. J. S. AND KIRBY, S. P. J. Bortkewitsch's horse-kicks and the generalized linear model. *The Statistician 37* (1988), 313–318.

[7] SIMAR, L. Maximum likelihood estimation of a compound Poisson process. *Annals of Statistics 4* (1976), 1200–1209.

[8] SOKAL, R. R. AND ROHLF, F. J. *Introduction to Biostatistics*. Freeman, San Francisco, 1973.

[9] TITTERINGTON, D. M., SMITH, A. F. M., AND MAKOV, U. E. *Statistical Analysis of Finite Mixture Distributions*. Wiley, 1985.

[10] VON BORTKEWITSCH, L. *Das Gesetz der kleinen Zahlen*. B.G. Teubner, Leipzig, 1898.

Denys Pommeret
CREST-ENSAI
Rue Blaise pascal, BP 37203
35170 Bruz Cedex — France
pommeret@ensai.fr