Modeling Probability Density through Ultraspherical Polynomial Transformations

Terhi Mäkinen*[1] and Lasse Holmström†[2]

[1]Finnish Meteorological Institute, P.O.Box 503, 00101 Helsinki, Finland

[2]University of Oulu, Department of Mathematical Sciences, P.O.Box 3000, 90014 University of Oulu, Finland

Abstract

We present a method for fitting parametric probability density models using an integrated square error criterion on a continuum of weighted Lebesgue spaces formed by ultraspherical polynomials. This approach is inherently suitable for creating mixture model representations of complex distributions and allows fully autonomous cluster analysis of high-dimensional data sets. The method is also suitable for extremely large sets, allowing *post facto* model selection and analysis even in the absence of the original data. Furthermore, the fitting procedure only requires the parametric model to be pointwise evaluable, making it trivial to fit user-defined models through a generic algorithm.

*terhi.makinen@fmi.fi, corresponding author

†lasse.holmstrom@oulu.fi

Work of LH supported by grant no. 24301034 from the Academy of Finland.

1 Introduction

As the complexity of data sets produced in scientific enterprises, engineering projects and internet-based applications keeps increasing, efficient tools are needed to uncover the information they contain. An example of such a tool is density estimation that can be used to model the probability distribution that supposedly generated the data. A density estimate can be used to visualize the salient features of the data and identify potentially interesting clusters in it. Density estimates can also be used in classification to recognize patterns in the data and they can be employed in simulation based analyses. For a general introduction to density estimation methods and their applications, see for example Klemelä (2009); Silverman (1986); Wand and Jones (1995); Scott (1992).

One approach to density estimation is to estimate the parameters thought to describe the probability distribution from which the data originated. Thus, given the data $\mathbf{x}_1, \ldots, \mathbf{x}_N \in \mathbb{R}^k$, one assumes that they were generated from a distribution with density function $f(\cdot|\boldsymbol{\theta})$ specified by a vector of parameters $\boldsymbol{\theta}$. An example would be fitting a normal (Gaussian) density to the data in which case $\boldsymbol{\theta}$ would comprise the components of the mean and the covariance matrix of the distribution. The standard way to estimate $\boldsymbol{\theta}$ is to maximize the likelihood $\prod_{i=1}^{N} f(\mathbf{x}_i|\boldsymbol{\theta})$ of the data. The maximizer $\hat{\boldsymbol{\theta}}$ is the maximum likelihood estimate (MLE) of $\boldsymbol{\theta}$ and the density estimate is $f(\cdot|\hat{\boldsymbol{\theta}})$. This approach is called "parametric" as it is based on the estimation of a set of unknown parameters that describe a family of model distributions.

A different approach is taken in "non-parametric" estimation, where the density from which the data were generated is not assumed to be a member of fixed parametrically defined family of models but rather to belong to a function space defined by some regularity conditions imposed on its member functions. In such estimation the data are allowed to "speak for themselves" and the estimate usually takes the form $\hat{f}(\cdot|\lambda)$, where λ is a tuning parameter that controls how much the individual data points are allowed to affect the result. A simple example in the case of univariate data is the histogram, a bar plot of the data normalized so that the area under the plot equals 1. Here the width of the bars acts as a tuning parameter in that a small width shows the contributions of individual data points whereas a large width smooths the data to show only its coarse, overall structure. The accuracy of a non-parametric density estimate $\hat{f}(\cdot|\lambda)$ of a density f is typically measured by an integrated error $\int_{\mathbb{R}^k} |\hat{f}(\mathbf{x}|\lambda) - f(\mathbf{x})|^p d\mathbf{x}$. From a mathematical point of view the L^1-error that corresponds to the choice $p = 1$ is most satisfactory as the integral is then always defined provided only that the estimate is integrable and the error is also invariant under reasonable transformations of the data (Devroye, 1987; Devroye and Györfi, 1985; Holmström and Klemelä, 1992). However, by far the most common choice is the L^2-error corresponding to $p = 2$ because it facilitates much more straightforward mathematical analyses.

Density estimation using mixture models is an approach somewhere in between the parametric and non-parametric methods (e.g. McLachlan and Peel (2000); Mengersen et al. (2011)). The mixture is a linear combination of parametrically defined component densities but the potentially large and even unspecified number of components gives the method a non-parametric flavor.

From a theoretical point of view, the parametric approach is more efficient than the non-parametric approach because it leads to a faster decrease of the error as a function of the sample size N. This however requires that the estimated unknown density indeed does belong to the hypothesized parametric family. Given the highly complex data sets the analyst faces today, any strict parametric assumptions may well be risky and therefore the flexibility of the non-parametric approach may well compensate for its lesser efficiency.

An interesting variation of these estimation principles is the L_2E-method described in Scott (2001). There, instead of MLE one fits a parametric model $f(\cdot|\boldsymbol{\theta})$ by approximately minimizing the integrated squared error $\int_{\mathbb{R}^k} (f(\mathbf{x}|\boldsymbol{\theta}) - f(\mathbf{x}))^2 d\mathbf{x}$. For earlier work on this idea, see the references in Scott (2001). A particularly interesting aspect of this method is its robustness against outliers, that is, its resistance to data points that do not originate from the density of interest but rather from a different distribution that corrupts the data at hand.

The method we propose is related to L_2E in that a parametric model is fit by effectively minimizing the L^2-distance. However, whereas L_2E uses the raw data $\mathbf{x}_1, \ldots, \mathbf{x}_N$ as such to approximate the part of the error that involves the unknown density f, we prefer to model the data non-parametrically before the parameters are fit. Thus, our proposal could be described as smooth L_2E. Orthogonal series expansion using a polynomial basis is the non-parametric estimation method employed. This results in a minimization problem involving a sum of squares that can be efficiently solved using for example the Levenberg-Marquardt algorithm.

The proposed method exhibits robustness similar to L_2E. In addition, the benefits particular to our method include the ability to efficiently model extremely large data sets through a single pass, completely parallelizable assimilation algorithm. Combined with the inherent suitability of orthogonal polynomials for representing multidimensional data, this leads to natural applications in data compression and feature extraction. For example, in classification of weather radar data, the difficulty of modeling multidimensional data sets without *a priori* knowledge about the actual distributions involved has favored the use of unrealistically simple empirical Bayes classifiers even in cases where their assumptions about coordinate independence clearly are invalid. As a preliminary demonstration of the potential of the proposed density estimation method we describe how to achieve a much better classifier design in a real world case that is practically unassailable by the more conventional approaches. The new method also appears to work particularly well in fitting mixtures of densities. Its resistance to outliers in the data guides the fitting algorithm to find clusters reliably in a decreasing order of importance. Comparisons with the EM algorithm, the standard fitting procedure for mixture models, suggests that our method can perform competitively or even exceed the EM performance when the difficulty of the task increases.

The rest of the paper is organized into two main sections. The first section describes the technical underpinnings of the method and the second provides examples of its application to the analysis of synthetic and empirical data sets. Some technical derivations related to ultraspherical polynomials are given in Appendix A. A proof of a consistency theorem is given in Appendix B.

2 The method

2.1 An outline of the approach

The basic idea is to connect the observed data and a suitable parametric density model through an intermediary entity that can efficiently interface with both. An L^2 function space is a natural choice for such an entity since orthogonal function expansions can be used to approximate both a parametrically defined probability density and the unknown density from which the data arose. This, however, requires that both expansions can be constructed in the same space in an effective manner. We propose to use a scale of weighted L^2 spaces that, combined with a linear transformation between any two of them, has the desired property.

In a univariate setting, let $\alpha > -1$ and consider the weight function $w^{(\alpha)}(x) = (1 - x^2)^{\alpha}$, $x \in [-1, 1]$, and the associated weighted Lebesgue space L^2_{α} that consists of functions f for which

$$\|f\|^2_{\alpha} = \int_{-1}^{1} f(x)^2 w^{(\alpha)}(x) dx < \infty. \tag{1}$$

Clearly, $L^2_{\alpha} \subset L^2_{\beta}$ when $\alpha < \beta$ and L^2_{α} includes all bounded functions on $[-1, 1]$, for example the polynomials. As explained in Appendix A, the ultraspherical polynomials $G^{(\alpha)}_m$, $m \in \mathbb{N}$, constitute an orthogonal basis for L^2_{α}, with the inner product between two basis functions given by

$$\langle G_m, G_n \rangle_{\alpha} = \int_{-1}^{1} w^{(\alpha)}(x) G^{(\alpha)}_m(x) G^{(\alpha)}_n(x) dx = \gamma^{(\alpha)}_m \delta_{mn}. \tag{2}$$

Two bases $\{G^{(\alpha)}_m\}$ and $\{G^{(\beta)}_m\}$ are related by an explicitly defined linear transformation (cf. (24)). As we describe below, certain choices of the weight parameter α are advantageous for each of the two subtasks, approximation of a parametric model and non-parametric representation of the data-generating density. Therefore, combined with the linear transformation between bases, the family $\{L^2_{\alpha}\}$ of weighted Lebesgue spaces provides a natural framework for the proposed density estimation method. The details of our approach will be explained in the context of univariate data only. The extension to the k-variate case is straightforward when one uses tensor product basis functions of the form $G^{(\alpha)}_{m_1,\ldots,m_d}(\mathbf{x}) = \prod_{i=1}^{k} G^{(\alpha)}_{m_i}(x_i)$,

$(m_1, \ldots, m_k) \in \mathbb{N}^k, \mathbf{x} \in \mathbb{R}^k$.

Note that there is no loss of generality in assuming that the density functions considered have support in $[-1,1]$ since one can always achieve this through a suitable change of variables. If the data-generating density can be assumed to have a compact support, a simple affine transformation to $[-1,1]$ can be applied. If a density $g : \mathbb{R} \to \mathbb{R}$ defined on the whole real line must be considered instead, one can use for example a transformation like $\phi : [-1,1] \to \mathbb{R}$, $\phi(x) = \kappa x/\sqrt{1-x^2}$, where $\kappa > 0$ is a user-defined scale factor. Then $f = (g \circ \phi)\phi'$ is the transformed density on $[-1,1]$ and after the desired model is fit, one can transform back to the original domain $\mathbb{R}$. This is in fact the transformation (with $\kappa = 1$) used in the numerical examples of Section 3. Note that this particular a choice of ϕ is pertinent in the present context because of the similarity between ϕ' and the weight function $w^{(\alpha)}$. This provides for example a simple formula for the L^2-norm of the original g,

$$\begin{aligned}\int_{-\infty}^{\infty} g(z)^2 dz &= \int_{-1}^{1} (\phi'(x))^{-1} f(x)^2 dx \\ &= \frac{1}{\kappa}\int_{-1}^{1} (1-x^2)^{\frac{3}{2}} f(x)^2 dx \\ &= \frac{1}{\kappa}\textstyle\sum_{n=0}^{m} \gamma_n^{(\frac{3}{2})}\left(\sum_{m=0}^{\infty} C_{mn}^{(\alpha,\frac{3}{2})} c_m^{(\alpha)}\right)^2\end{aligned} \tag{3}$$

where $f = \sum_{m=0}^{\infty} c_m^{(\alpha)} G_m^{(\alpha)}$ for some $\alpha > -1$ and the coefficients $C_{mn}^{(\alpha,\frac{3}{2})}$ link the bases in L_α^2 and $L_{3/2}^2$ (cf. Appendix A). In a multivariate setting a suitable transformation can be applied separately to each coordinate.

2.2 The non-parametric model

Consider a density $f \in L_\alpha^2$ and let

$$f = \sum_{m=0}^{\infty} d_m G_m^{(\alpha)}. \tag{4}$$

By (2),

$$d_m = \frac{1}{\gamma_m^{(\alpha)}} \int_{-1}^{1} w^{(\alpha)}(x) G_m^{(\alpha)}(x) f(x) dx = \mathbb{E}\left\{\frac{w^{(\alpha)}(X) G_m^{(\alpha)}(X)}{\gamma_m^{(\alpha)}}\right\}, \tag{5}$$

where the distribution of the random variable X has density f. Given a random sample $X_1, \ldots, X_N \sim f$ and defining $\xi_m^{(\alpha)}(x) = w^{(\alpha)}(x) G_m^{(\alpha)}(x)/\gamma_m^{(\alpha)}$ this allows us to calculate an estimator for d_m through

$$\hat{d}_m = \frac{1}{N} \sum_{i=1}^{N} \xi_m^{(\alpha)}(X_i). \tag{6}$$

The estimator is unbiased, $\mathbb{E}(\hat{d}_m) = d_m$.

Consider then the M-dimensional random vector $\hat{\mathbf{d}} = (\hat{d}_m) \equiv (\hat{d}_0, \ldots, \hat{d}_{M-1})^{\mathsf{T}}$. We have

$$\hat{\mathbf{d}} = \frac{1}{N} \sum_{i=1}^{N} \boldsymbol{\xi}_i^{(\alpha)},$$

where $\boldsymbol{\xi}_i^{(\alpha)} = (\xi_0^{(\alpha)}(X_i), \ldots, \xi_{M-1}^{(\alpha)}(X_i))^{\mathsf{T}}$ and therefore, by the central limit theorem, for large N the distribution of $\hat{\mathbf{d}}$ is approximately multivariate normal. A sample estimate of the covariance matrix of $\hat{\mathbf{d}}$ is given by

$$\widehat{\boldsymbol{\Sigma}} = \frac{1}{N(N-1)} \sum_{i=1}^{N} (\boldsymbol{\xi}_i^{(\alpha)} - \bar{\boldsymbol{\xi}}_N^{(\alpha)})(\boldsymbol{\xi}_i^{(\alpha)} - \bar{\boldsymbol{\xi}}_N^{(\alpha)})^{\mathsf{T}}, \tag{7}$$

where $\bar{\boldsymbol{\xi}}_N^{(\alpha)} = (1/N) \sum_{i=1}^{N} \boldsymbol{\xi}_i^{(\alpha)}$ is the sample mean.

The above considerations are valid at least when the functions $\xi_m^{(\alpha)}$ are bounded on $[-1, 1]$. For ultraspherical polynomials this requirement is satisfied for all $\alpha \geq 0$. However, our strategy for estimating a parametric model for the density f is to fit the model in a finite dimensional polynomial subspace of L_0^2, the ordinary non-weighted space of square integrable functions on $[-1, 1]$ (cf. Section 2.4). The optimal target density then is the orthogonal projection of f onto the polynomial subspace which suggests that $\{G_m^{(0)}\}$ is the orthogonal basis one should use to represent f.

The ultraspherical polynomials $G_m^{(0)}$ are the classical Legendre polynomials P_m with the orthogonality factor in (2) given by

$$\gamma_m^{(0)} = \gamma_m = \frac{2}{2m+1}. \tag{8}$$

The expansion (4) becomes the Legendre series of f,

$$f = \sum_{m=0}^{\infty} d_m P_m. \tag{9}$$

Since now $w^{(0)}(x) = P_0(x) = 1$ for all $x \in [-1, 1]$, we have by (5) that

$$d_0 = \frac{1}{\gamma_0} \int_{-1}^{1} f(x)dx = \frac{1}{\gamma_0} = \frac{1}{2}. \tag{10}$$

Since also $\hat{d}_0 = 1/2$, the first Legendre series coefficient d_0 does not have to be estimated or taken into account when minimizing the L^2-distance between the non-parametric estimate of the density and a fitted parametric model, another reason for choosing $\alpha = 0$ (see Section 2.4).

To minimize the expected L^2-distance between the underlying density f and its non-parametric estimate, the empirical coefficients $\hat{d}_m$ need to be shrunk in some way. We have adopted a simple truncated estimator approach with the Hart (1985) cut-off criterion. For an M-term Legendre polynomial estimator, the Hart criterion is equivalent to finding an $M \geq 2$ that minimizes

$$H(M) = \sum_{m=1}^{M-1} \gamma_m \left[2\widehat{\mathrm{Var}}(\hat{d}_m) - \hat{d}_m^2 \right], \tag{11}$$

where diagonal elements of (7) provide estimates for the variances of the coefficients $\hat{d}_m$. The non-parametric estimate of f is then defined as

$$\hat{f}(\cdot|M) = \sum_{m=0}^{M-1} \hat{d}_m P_m, \tag{12}$$

where the role of the smoothing parameter is played by M. The criterion (11) is both trivial to evaluate and is easily generalized for multivariate data. A number of alternative approaches exists as described, e.g., in Efromovich (2010). The specific choice of a shrinking scheme is non-critical to the working of the proposed method as long as the quality of the estimate is adequately controlled. Also note that yet another benefit of using the Legendre basis is that the non-parametric density estimate $f(\cdot|M)$ is always normalized since by (10),

$$\int_{-1}^{1} \hat{f}(x|M)dx = \left\langle P_0, \hat{f}(\cdot|M) \right\rangle_0 = \gamma_0 \hat{d}_0 = \gamma_0 d_0 = 1.$$

With large data sets, for evaluating $\hat{d}_m$ and $\mathrm{Var}(\hat{d}_m)$ we strongly recommend some variant of the single-pass, parallelizable algorithm suggested by Chan et al. (1979), depending on computational constraints either in the simple form, or as a highly stable pairwise adder that requires of the order of $\log_2(N)$ additional storage units per cumulant. This allows the partitioning of data in an arbitrary manner which is a necessary requirement for parallel processing but also provides an advantage in the management of data that are divided into particular sets, as the cumulants of a set of sets are simply the sums of cumulants of individual sets. Again, while the choice of algorithm for calculating the coefficients is not germane to the method itself, the availability of a stable, single-pass and parallelizable algorithm for data assimilation gives the method the widest possible applicability.

2.3 The parametric model

The weighted Lebesgue space $L^2_{-1/2}$ of Chebyshev polynomials of the first kind is ideal for fitting a parametric model $f(\cdot|\boldsymbol{\theta})$ because the expansion coefficients in (4) can be approximated without explicit integration and Runge's phenomenon is minimized (Berrut and Trefethen, 2004). Thus, denote by T_m the Chebyshev polynomials of the first kind (see e.g. Abramowitz and Stegun (1972), p. 889), $T_m = G_m^{(-1/2)}$, and suppose that

$$f(\cdot|\boldsymbol{\theta}) = \sum_{m=0}^{\infty} c_m(\boldsymbol{\theta}) T_m. \tag{13}$$

Denoting the kth Chebyshev node of $T_M(x)$ by

$$x_k = \cos\left(\pi \frac{2k+1}{2M}\right) \tag{14}$$

we have the discrete orthogonality equation

$$\sum_{k=0}^{M-1} T_m(x_k) T_n(x_k) = \gamma_m^M \delta_{mn}\,, \quad m, n = 0, \ldots, M-1, \tag{15}$$

where $\gamma_m^M = (M/2)(1 + \delta_{m,0})$. Assuming point-wise convergence in (13), which holds for example for a continuous $f(\cdot|\boldsymbol{\theta})$, it follows that the M first expansion coefficients can obtained from a discrete cosine transformation (DCT) of $f(\cdot|\boldsymbol{\theta})$,

$$c_m(\boldsymbol{\theta}) = \frac{1}{\gamma_m^M} \sum_{k=0}^{M-1} f(x_k|\boldsymbol{\theta}) T_m(x_k), \quad m = 0, \ldots, M-1. \tag{16}$$

This use of the Chebyshev polynomial basis sets minimal requirements for a user-defined model since only its evaluation with given parameters at predetermined points is required.

2.4 Coupling the models

Non-parametric modeling creates a Legendre expansion that approximates the density of the underlying data-generating distribution, and parametric modeling creates a Chebyshev expansion of a user-provided, parametrized density model. In principle, both of these can be transformed to an arbitrary ultraspherical basis if the application calls for a particular choice. In the absence of such preferences, the best choice for matching the expansions is the Legendre basis.

It is important to notice that transforming the parametric model to an expansion in an arbitrary basis is no more computationally intensive than transforming it to a Chebyshev expansion. For example, if $\mathbf{T} = (T_m(x_k))$ is the matrix performing the DCT of (16), then the Chebyshev coefficients are given by the column vector $\mathbf{c}(\boldsymbol{\theta}) = (c_m(\boldsymbol{\theta})) = \mathbf{T}\mathbf{f}(\boldsymbol{\theta})$ where $\mathbf{f}(\boldsymbol{\theta}) = (f(x_k)|\boldsymbol{\theta}))$. If $\mathbf{C}$ is the transformation between the M first Chebyshev and Legendre basis functions, then the Legendre coefficients $\mathbf{d}(\boldsymbol{\theta}) = (d_m(\boldsymbol{\theta}))$ are given by

$$\mathbf{d}(\boldsymbol{\theta}) = \mathbf{C}^{\mathsf{T}}\mathbf{c}(\boldsymbol{\theta}) = \mathbf{C}^{\mathsf{T}}\mathbf{T}\mathbf{f}(\boldsymbol{\theta}),$$

where both $\mathbf{C}^{\mathsf{T}}$ and $\mathbf{T}$ are constant matrices that can be combined into a single transformation $\mathbf{L} = \mathbf{C}^{\mathsf{T}}\mathbf{T}$, yielding the Legendre coefficients as

$$\mathbf{d}(\boldsymbol{\theta}) = \mathbf{L}\mathbf{f}(\boldsymbol{\theta}). \tag{17}$$

It appears that the complete formula for the matrix $\mathbf{C}$ has not been published before and we therefore include its derivation in the Appendix A (Theorem A.4).

Now that both the model and the data are expressed in the same polynomial basis, we can easily measure their difference using the L^2-distance,

$$\begin{aligned} \|f(\cdot|\boldsymbol{\theta}, M) - \hat{f}(\cdot|M)\|_0^2 &= \int_{-1}^{1} \Big[f(x|\boldsymbol{\theta}, M) - \hat{f}(x|M)\Big]^2 dx \\ &= \sum_{m=0}^{M-1} \gamma_m (d_m(\boldsymbol{\theta}) - \hat{d}_m)^2, \end{aligned} \tag{18}$$

where $f(\cdot|\boldsymbol{\theta}, M)$ is the M-term Chebyshev approximation of $f(\cdot|\boldsymbol{\theta})$ obtained from the coefficients (16). The minimization of (18) is a standard procedure. For models that only provide function values at given points, Powell's method (Powell, 1964) is an effective choice; for models that also provide partial derivatives of the model with respect to its parameters the Levenberg-Marquardt algorithm (Levenberg, 1944; Marquardt, 1963) can be applied. Once a parameter vector $\hat{\boldsymbol{\theta}}$ that minimizes (18) has been found, the final density estimate is $f(\cdot|\hat{\boldsymbol{\theta}})$. Note here the asymptotic relationship to Scott's L_2E: for a large sample size N, and hence large M, we have that $f(\cdot|\boldsymbol{\theta}, M) \approx f(\cdot|\boldsymbol{\theta})$ and $\hat{f}(\cdot|M) \approx f$.

The goodness of the fit of the model can be estimated through χ^2-statistics,

$$\chi^2_\nu = \|\widehat{\boldsymbol{\Sigma}}_1^{-1/2}(\mathbf{d}_1(\hat{\boldsymbol{\theta}}) - \hat{\mathbf{d}}_1)\|^2, \tag{19}$$

where the first fixed Legendre coefficient has been left out from $\mathbf{d}(\hat{\boldsymbol{\theta}})$ and $\hat{\mathbf{d}}$, that is, $\mathbf{d}_1(\hat{\boldsymbol{\theta}}) = (d_1(\hat{\boldsymbol{\theta}}), \ldots, d_{M-1}(\hat{\boldsymbol{\theta}}))^\mathsf{T}$, $\hat{\mathbf{d}}_1 = (\hat{d}_1, \ldots, \hat{d}_{M-1})^\mathsf{T}$, and $\widehat{\boldsymbol{\Sigma}}_1$ is the sample covariance matrix of $\hat{\mathbf{d}}_1$ (cf. (7)). The degrees of freedom are given by $\nu = M - 1 - p$, where p is the number of components of $\boldsymbol{\theta}$.

The following theorem is a consistency result for an estimator $\hat{\boldsymbol{\theta}}_N = \hat{\boldsymbol{\theta}}$. Its proof can be found in the Appendix B.

Theorem 2.1. *Consider a compact parameter space* $\boldsymbol{\Theta} \subset \mathbb{R}^p$ *and let* $\{f(\cdot|\boldsymbol{\theta})|\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$ *be a family of probability density functions on* $[-1, 1]$ *for which both* $(x, \boldsymbol{\theta}) \mapsto f(x, \boldsymbol{\theta})$ *and* $(x, \boldsymbol{\theta}) \mapsto \frac{\partial}{\partial x} f(x, \boldsymbol{\theta})$ *are continuous. Assume that for the density* f *underlying the data one has* $f = f(\cdot|\boldsymbol{\theta}_0)$ *for a unique* $\boldsymbol{\theta}_0 \in \boldsymbol{\Theta}$*. Assume further that in the Legendre expansion (9)* $d_m \neq 0$ *for infinitely many* m*, that is,* f *is not a polynomial, and that the Hart criterion (11) is applied by selecting the optimal* M *from a set* $\{2, 3, \ldots, R(N)\}$*, where* $R(N) \to \infty$ *and* $R(N) = o(\sqrt{N})$ *as the sample size* N *tends to infinity. Then* $\hat{\boldsymbol{\theta}}_N \xrightarrow{P} \boldsymbol{\theta}_0$ *as* $N \to \infty$ *(convergence in probability), where ,* $\hat{\boldsymbol{\theta}}_N$ *is obtained by minimizing (18) over* $\boldsymbol{\Theta}$*.*

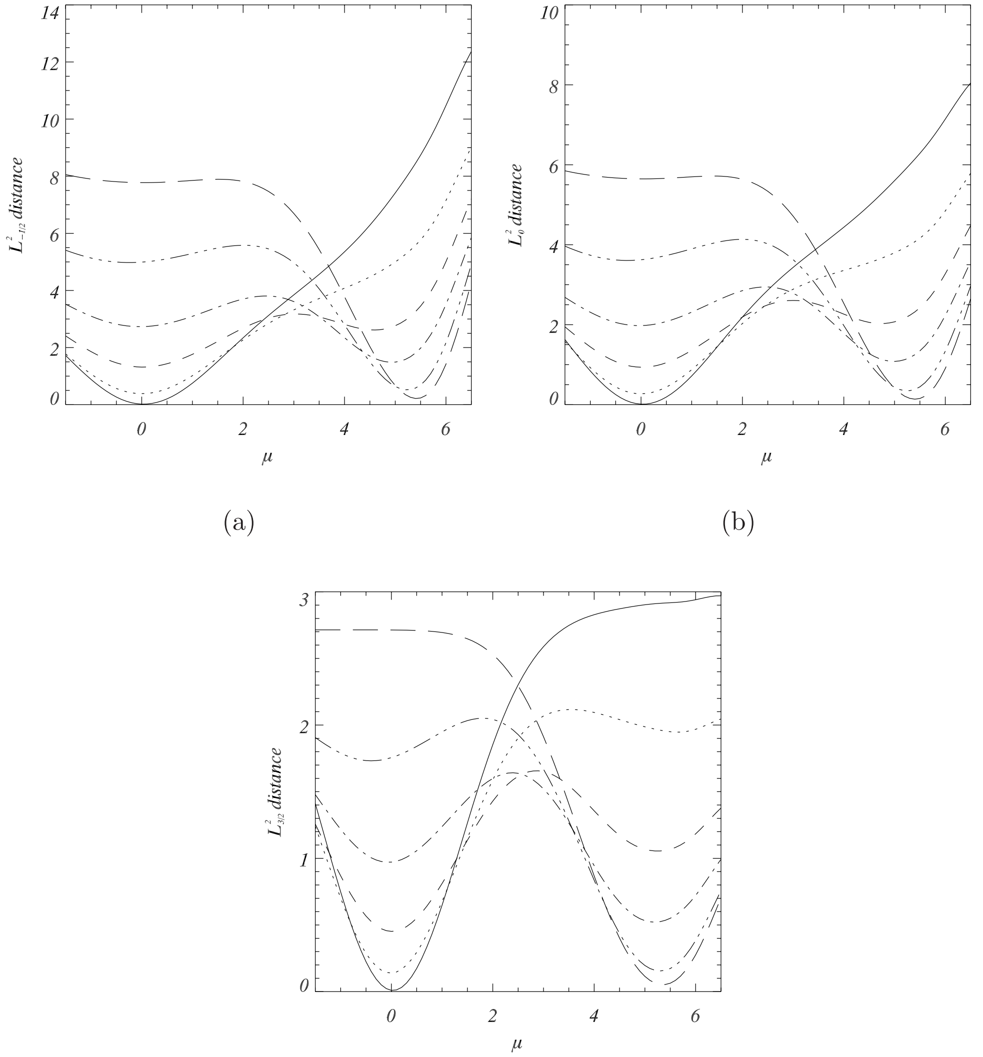
3 Numerical examples

The utility of the proposed method can be assessed through specific numerical examples. In this section the robustness of the estimator is first examined, followed by an application to cluster analysis and to construction of an empirical Bayes classifier for weather radar data.

3.1 Robustness of fit

As discussed in Scott (2001), in contrast to maximum likelihood, minimization of the integrated square error (i.e. L_2E) in parametric modeling leads to an inherently robust estimation method. L_2E does not require any tuning parameters typically found in robust likelihood algorithms, making it a desirable alternative provided that it can be computed efficiently. The method suggested in the present article works in the spirit of L_2E so it is of interest to examine if it actually exhibits similar robustness against outliers. A test on synthetic data similar to the one used by Scott was therefore performed. A random sample of size $N = 100$ was generated from two normal distributions, $N(0, 1)$ and $N(5, 1)$, with mixing ratios between 0 and 100% at 20% intervals. The data were projected to $[-1, 1]$ by using (3) with $\kappa = 1$, and the underlying probability density was approximated by a Legendre series $\hat{f}(\cdot|M)$, as described in Section 2.2. Likewise, the density function $f(\cdot|\mu)$ of $N(\mu, 1)$ was approximated by a Chebyshev series in a manner explained in Section 2.3. To test the robustness of the method in different weighted Lebesgue spaces, the L^2_α-distance between $\hat{f}(\cdot|M)$ and $f(\cdot|\mu)$ as a function of μ was calculated both in the Chebyshev ($\alpha = -1/2$) and Legendre ($\alpha = 0$) spaces, and also in the special space $L^2_{3/2}$ (cf. (3)). The results are shown in Figure 1.

As expected, the choice of the Lebesgue space has a moderate effect on the overall shape of the distance function. In a sense, the special case $\alpha = 3/2$ is the most objective choice by the virtue of coinciding with the L^2 metric in $\mathbb{R}$ (cf. (3)). However, the behavior of the minima and thus also the parametric density estimate is not affected by the metric. Just like L_2E, instead of yielding a linear interpolation between the two models like MLE does, the



(a)

(b)

(c)

Figure 1: The L^2_α-distance between a normal model $N(\mu, 1)$ and a non-parametric estimate of the data-generating density when a sample of size $N = 100$ was drawn from two normal distributions $N(0, 1)$ and $N(5, 1)$ at mixing ratios between 0 and 100% at 20% intervals. The panels (a) - (c) shown distance as a function of μ in the spaces $L^2_{-1/2}$, L^2_0 and $L^2_{3/2}$, respectively. The solid line corresponds to a sample from $N(0, 1)$ only, the line with long dashes to a sample from $N(5, 1)$ only, and the remaining four curves correspond to mixing ratios 20%, 40%, 60%, and 80%, respectively.

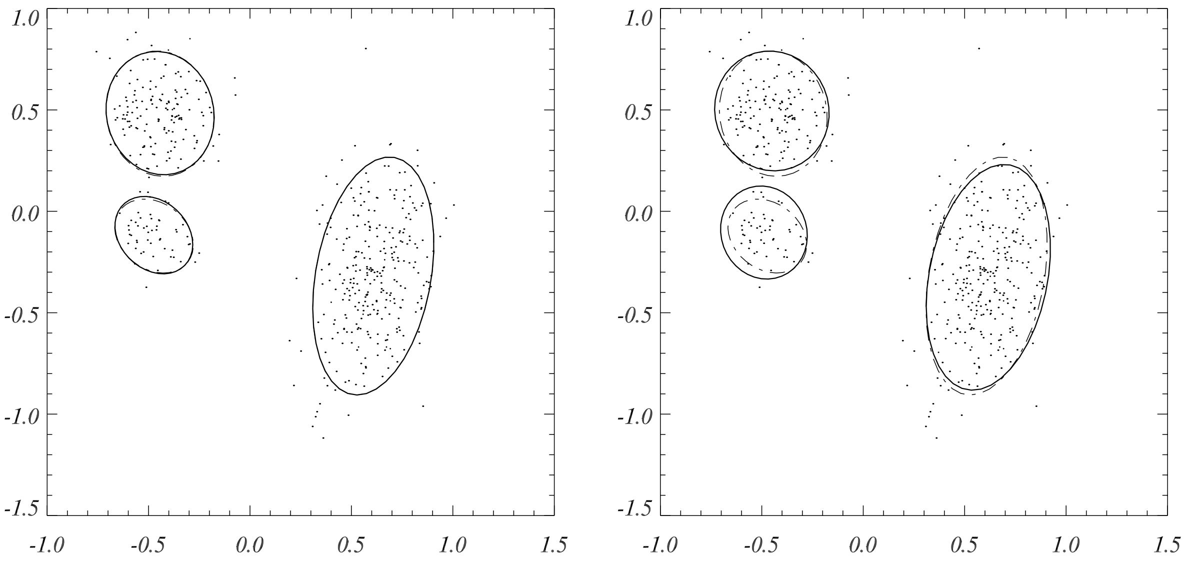


Figure 2: Cluster analysis, "easy" test. 500 points were drawn from three bivariate normal distributions and the resulting data set was analyzed by the method in this paper and by the Matlab `fitgmdist` algorithm (EM). Left: The result from the EM algorithm (solid line). The exact solution is congruent with the EM one. Right: The result by the method of this paper (solid line) and the exact solution (dash-dotted line)

proposed method gives an estimate for either one or the other distribution, depending on the mixing ratio (cf. Scott (2001)). This property of our method opens up some interesting applications.

3.2 Cluster analysis

For tasks where only the global minimum of a function is of particular interest, the existence of local minima is a significant problem. However, if the objective is to approximate a complex function with a mixture of simple ones, then local minima may actually facilitate the task. A relatively simplistic approach is to converge on the nearest minimum and "fill" it with a prototype fitting function, and then iterate this procedure until a convergence criterion is satisfied. In the context of the proposed method, the fitting of model functions is done to

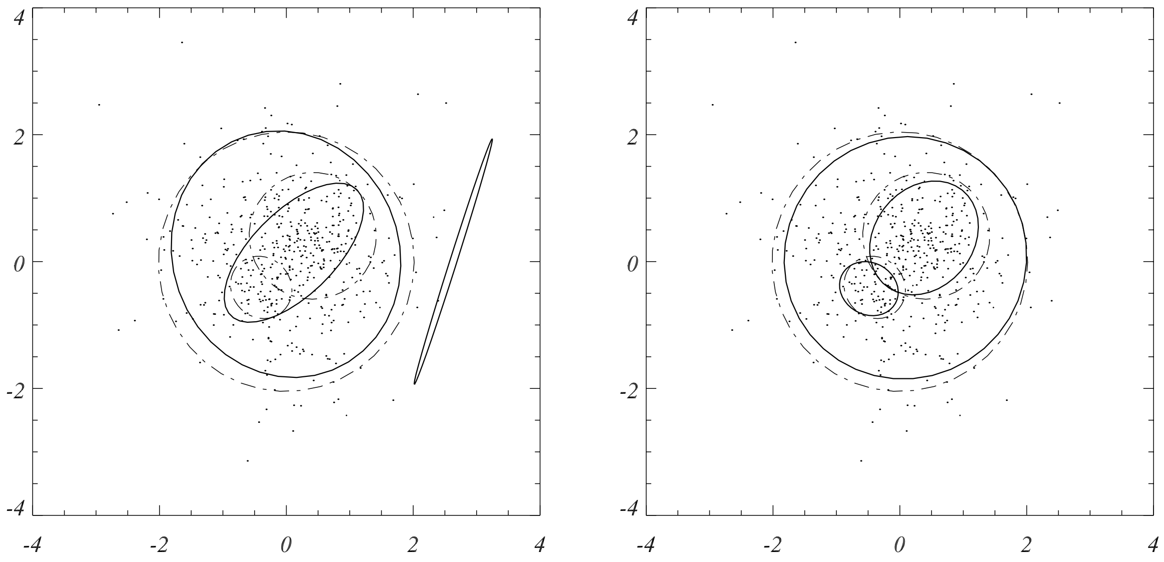


Figure 3: Cluster analysis, “hard” test. 8000 points points were drawn from three bivariate normal distributions and the resulting data set was analyzed by the method in this paper and by the Matlab `fitgmdist` algorithm (EM). Left: The result from the EM algorithm (solid line) and the exact solution (dash-dotted line). Right: The result by the method of this paper (solid line) and the exact solution (dash-dotted line). Only a fraction of the points is shown to keep the plot readable.

Test	i	Solution	p_i	$\mu_{i,x}$	$\mu_{i,y}$	$\sigma_{i,x}$	$\sigma_{i,y}$	ρ_i
easy	0	exact	0.6000	0.6079	-0.3189	0.1495	0.2934	0.2676
		OPT	0.5852	0.6167	-0.3254	0.1531	0.2783	0.2559
		EM	0.6000	0.6079	-0.3189	0.1495	0.2934	0.2676
	1	exact	0.3000	-0.4435	0.4815	0.1329	0.1545	-0.0605
		OPT	0.2991	-0.4502	0.4948	0.1411	0.1476	-0.0397
		EM	0.2980	-0.4430	0.4840	0.1334	0.1520	-0.0690
	2	exact	0.1000	-0.4730	-0.1219	0.0970	0.0910	-0.2422
		OPT	0.1157	-0.4894	-0.1046	0.1070	0.1147	-0.0688
		EM	0.1020	-0.4738	-0.1174	0.0959	0.0954	-0.2404
hard	0	exact	0.6000	0.0000	-0.0011	1.0079	1.0225	0.0038
		OPT	0.7216	0.0875	0.0592	0.9563	0.9553	0.0184
		EM	0.6541	-0.0054	0.1166	0.9058	0.9720	-0.0670
	1	exact	0.3000	0.4165	0.4066	0.5011	0.4995	-0.0020
		OPT	0.2057	0.3813	0.3730	0.4289	0.4486	0.2017
		EM	0.3347	0.1197	0.1396	0.5493	0.5482	0.6387
	2	exact	0.1000	-0.3959	-0.4071	0.2422	0.2457	-0.0848
		OPT	0.0727	-0.4879	-0.4268	0.2314	0.2129	-0.1326
		EM	0.0111	2.6328	0.0010	0.3085	0.9682	0.9955

Table 1: Results for fitting a three component normal mixture in the easy and hard cases. Here p_i is the mixing ratio, $\mu_{i,x}$ and $\mu_{i,y}$ are the components of the mean, $\sigma_{i,x}$ and $\sigma_{i,y}$ are the marginal variances and ρ_i is the correlation coefficient. For each test, the table gives the exact parameter values as well as the values obtained with the method described in this article (OPT) and the Matlab `fitgmdist` algorithm (EM).

the Legendre expansion, thus making the computational cost effectively independent of the size of the underlying data set. As a direct consequence of this and the robustness property of the proposed fitting algorithm, our method can be used for computationally efficient cluster analysis. The relative computational advantage compared to algorithms that manipulate the data directly is proportional to the sample size involved. The fitting algorithm in pseudocode for general mixtures in $\mathbb{R}^k$ is displayed in Algorithm 3.1. A theorem on the consistency of the estimators $\hat{\boldsymbol{\theta}}_i$ of the mixture component parameters in the one-dimensional case with a fixed number of components is formulated in Appendix B.

Algorithm 3.1 Fitting mixtures

```
Given data in $\mathbb{R}^k$, transform them to $[-1,1]^k$
Let $\boldsymbol{\theta}_i$ be the parameter vector of the $i$th fitted mixture component $f(\cdot|\boldsymbol{\theta}_i)$
$i = 1$
Let $\hat{f}_1(\cdot|M)$ be the Legendre series estimate of the underlying density in $[-1,1]^k$
repeat
    Let $f(\cdot|\boldsymbol{\theta}_i)$ be the $i$th mixture component function transformed to $[-1,1]^k$ and let $f(\cdot|\boldsymbol{\theta}_i, M)$ be the Legendre series approximation with the same basis functions as in $\hat{f}_1(\cdot|M)$
    Let $\hat{\boldsymbol{\theta}}_i$ be the minimizer of $\|f(\cdot|\boldsymbol{\theta}_i, M) - \hat{f}_i(\cdot|M)\|_0^2$
    Let $\hat{f}_{i+1}(\cdot|M) = \hat{f}_i(\cdot|M) - f(\cdot|\hat{\boldsymbol{\theta}}_i, M)$
    $i = i + 1$
until convergence
$K = i$
The fitted mixture is $\sum_{i=1}^{K} f(\cdot|\hat{\boldsymbol{\theta}}_i)$ transformed to $\mathbb{R}^k$
```

In order to assess the performance of our method, two synthetic data sets were constructed, representing “easy” and “hard” problems. In both tests data were drawn from three two-dimensional normal distributions in mixing ratios $(p_1, p_2, p_3) = (0.6, 0.3, 0.1)$. In Algorithm 3.1 we therefore have $\boldsymbol{\theta}_i = (p_i, \boldsymbol{\mu}_i, \boldsymbol{\Sigma}_i)$, where $\boldsymbol{\mu}_i$ is the mean and $\boldsymbol{\Sigma}_i$ is the covariance matrix of the ith mixture component. In the easy test these distributions were well apart from each other with the total number of data points $N = 500$. In the hard case they were on top of each other, with $N = 8000$ to provide statistical significance. The solutions given by our method were compared to those from the Matlab `fitgmdist` Expectation-Maximization (EM) algorithm (Dempster et al., 1977) which is the default choice for the task. The results, along with expected (exact) solutions, are listed in Table 1 and shown in figures 2 and 3 for the easy and hard cases, respectively.

In the easy case the EM algorithm produced the essentially correct solution whereas the result obtained with the proposed method was less accurate. In the hard case our solution was still quite close to the exact one but the EM solution that had the highest likelihood out of 100 runs of the algorithm merged the two weaker distributions into one and produced a completely superfluous third one. The failure of the EM algorithm to analyze the hard case correctly was indeed anticipated since the algorithm relies on the spatial separation of data points for estimating their respective latent variables, an approach that becomes less efficient as component distributions overlap (cf. Redner and Walker (1984)).

While the solutions given by both approaches are roughly comparable, there are additional advantages beyond data set size considerations that the proposed method has over the EM algorithm. First, the number of clusters must be explicitly given to the EM algorithm whereas our method can determine it directly from the data. This can be achieved simply by progressively fitting each local minimum, one component at a time, until either the sum of the mixing ratios approaches unity or the χ^2_ν in (19) becomes sufficiently small. Thus, unlike the EM approach that requires manual inspection and judicious interpretation of the solution, our method implements completely autonomous cluster analysis.

Second, the method is deterministic if a reasonable guess, like a maximum likelihood estimate, for the initial state of the minimization algorithm is used. This is unlike the EM algorithm where the validity of the solution depends on random initialization, thus making it necessary to repeat the whole EM procedure until a solution with an acceptable likelihood is produced. This may add a large overhead to computational requirements; in the tests described here the EM algorithm had a success rate of 55% for the easy case and only 4%

for the hard case.

It must be noticed that the proposed method always produces a solution that is a fair approximation of the original distribution, even if all the underlying clusters are not identified. In the hard case, halving the number of data points to 4000 yielded a solution with just two components because the close proximity of the underlying distributions made the difference between two and three components statistically insignificant.

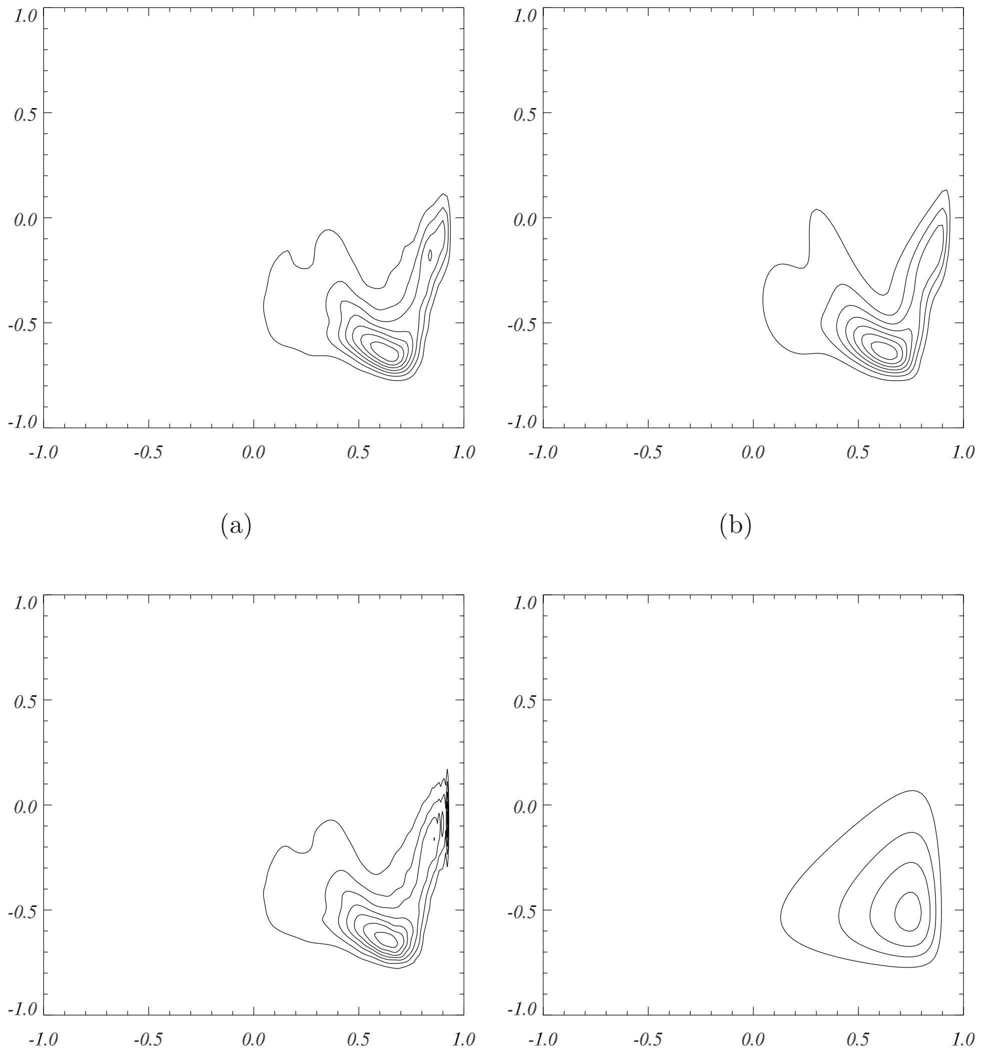
While it was not implemented for this study, it is also trivial to create a mixed-model version of the fitting algorithm. Instead of using just one prototype function, like the Gaussian, to fill the local minima, an extended version of the algorithm would try all functions from an user-provided set and choose the one that fits the data best.

In conclusion, our method produces solutions that are comparable to those of the EM algorithm in numerical accuracy while simultaneously providing significant methodological and computational advantages, making the method an attractive choice for cluster analysis.

3.3 Classifier design

A major hurdle in the construction of density based classifiers has been the relative difficulty of modeling high-dimensional data sets. Because of this, an overly simple Bayes approach that assumes independent feature dimensions may still be the default choice for example in weather radar data analyses, even for cases where the independence assumption is either questionable or obviously invalid. Here we outline a real world case of supervised learning in which the described method was used to model and analyze a very large and high-dimensional data set with the purpose of producing optimal components for an empirical Bayes classifier.

New dual polarimetric weather radars produce copious amounts of data which creates an immediate need for automatic classification of observed targets. This task is relatively well established for common meteorological phenomena but much less so for other target classes like birds or insects. Simple multivariate models like Principal Component Analysis (PCA) and neural networks have previously been used as the framework for supervised learning. However, both approaches have significant shortcomings in this particular context and the



(a)

(b)

(c)

(d)

Figure 4: The marginal density of a dual polarimetric radar target class (arctic birds) in the subspace that maximizes its distance from all other classes in the training set, shown as a Legendre expansion (a) and a mixture of five multivariate normal distributions based on the Legendre expansion (b). For comparison, a kernel density estimate (c) and a MLE model (d) are also shown. The MLE model is a multivariate normal density that looks here distorted because it has been transformed to the rectangle $[-1, 1] \times [-1, 1]$.

method described in this article was specifically developed to overcome these issues.

The training set for the project in question consists of about 1000 radar images manually selected and labeled into 30 classes. The total number of data points in the set is about one million. The channels coming directly from the radar are augmented by applying a number of filters that quantify the presence of particular types of texture in the images. In the current configuration this brings the dimensionality of the data set up to 58.

The data were first turned into class specific Legendre expansions. For each class, a low-dimensional optimally resolving subspace, maximizing target class separation was then determined by simple search through possible combinations of feature vector dimensions. This task was made easy by the properties of multivariate orthogonal polynomial expansions for which a marginal density is obtained through a simple selection of expansion coefficients: to integrate out variables x_i, $i \in J$ one simply drops the coefficients $d_{m_1,\ldots,m_k}$ for which $m_i > 0$ for some $i \in J$. This follows because $\int_{-1}^{1} P_{m_i}(x_i)dx_i = \langle P_0, P_{m_i} \rangle_0 = 0$ when $m_i > 0$. The L^2-distances between class models were evaluated as weighted euclidean distances of coefficient vectors (cf. (21)). Once the optimal subspace was found, the marginal density of each class was turned into a minimal parametric model to be used as a class density in an empirical Bayes classifier for the class in question.

Concerning the dimensionality of the Legendre expansion, it would be both computationally infeasible and statistically meaningless to calculate the coefficients of a full 58-dimensional model. Taking into account the cut-off criterion for discarding insignificant coefficients allows the dimensionality to be reduced to a size comparable to the lengths of one-dimensional expansions. In this application where the objective was to find a low-dimensional class model, the complexity of the Legendre expansion could be reduced even further to the expected maximal dimensionality of the final models. Thus, based on existing knowledge about the task, the expansion used in the analysis consisted only of terms that each involved at most five different variables x_i.

A typical class density, corresponding to the training class “arctic birds”, is shown in Figure 4, plotted in its own optimal two-dimensional subspace, in which the complement class “NOT-arctic-birds” would be approximately centered and Gaussian, thus allowing easy classification. Comparing the minimal mixture model that was derived from the Legendre expansion to a kernel density estimate calculated directly from the bivariate data and used here for visualization purposes, demonstrates good congruence of the minimal model with the original data. PCA is the only other commonly used approach that would be comparable to our method in the ability to search for optimal subspaces and the amount of computing time required to produce a solution, but the resulting MLE models are very poor approximations of the actual distributions, a known issue in the context of radar data that makes PCA unsuitable for the task.

In conclusion, the method does not just solve the task of creating density models for Bayes classifiers but makes it a relatively trivial and effortless exercise, with the added benefits that the solution is easy to inspect and verify, and allows for easy manipulation of original class definitions without a need to re-assimilate the data – none of which are true for example with neural networks. In addition to such practical computational advantages, one might also conjecture that the usefulness of the proposed method in high dimensional problems is increased by the potential capability of orthogonal series methods to resist the curse of dimensionality (see e.g. Prakasa Rao (1983, Section 3.3), Krzyżak and Pawlak (1982)), Klemelä (2009, Section 16.4)).

4 Summary

We have described a method for estimating the probability density of potentially large and multivariate data sets by parametric models through ultraspherical polynomials and given specific examples of potential applications for the method. Similarly to L_2E, the proposed method lies between parametric and non-parametric approaches to density estimation where a parametric model is fitted by minimizing an integrated square error. The key idea is to model the data first non-prametrically and to then fit the parametric model in a finite dimensional space spanned by polynomials. The advantages of the proposed method over more conventional approaches include:

- Single pass, perfectly parallelizable data input that allows processing of very large data sets.
- Economical handling of high dimensional data sets offered by the properties of orthogonal series expansions.
- Compressing the data into an orthogonal series expansion implies that the subsequent parametric fitting step is independent of the sample size. The Levenberg-Marquardt algorithm can be used for efficient parameter estimation.
- In pattern recognition applications, efficient computation of marginal densities and L^2-distances greatly facilitate feature extraction and classifier design.

The relevant algorithms are easy to implement and we believe that the proposed method has potential to be of great generic utility in practical data analyses.

A Ultraspherical polynomials

The method described in this work is based on the weighted Lebesgue spaces L^2_α on $[-1, 1]$, defined by the weight functions $w^{(\alpha)}(x) = (1 - x^2)^\alpha$, $\alpha > -1$, and their orthogonal bases defined by ultraspherical or Gegenbauer polynomials. Because the required properties of the ultraspherical polynomials are not readily found in the related literature, we provide the necessary details here.

Definition 1. Given the Jacobi polynomials $P_m^{(\alpha\beta)}$ (see, e.g., Szegő (1959, Ch. IV)), we define the normalized ultraspherical polynomials $G_m^{(\alpha)}$ as

$$G_m^{(\alpha)}(x) = P_m^{(\alpha\alpha)}(x)/P_m^{(\alpha\alpha)}(1)$$

for $\alpha > -1$, $m \in \mathbb{N}$ and $x \in [-1, 1]$.

Remark. The above definition is chosen for simple consistency between related spaces. The classical definition of Gegenbauer polynomials uses a parametrization that differs by a constant $-1/2$ compared to the definition above, as well as a different normalization factor.

Theorem A.1. *The polynomials $G_m^{(\alpha)}$ are orthogonal on the interval $[-1, 1]$ with respect to the weight function $w^{(\alpha)}(x) = (1 - x^2)^\alpha$ and satisfy the orthogonality equation*

$$\begin{aligned} \int_{-1}^{1} w^{(\alpha)}(x) G_m^{(\alpha)}(x) G_n^{(\alpha)}(x) dx &= \gamma_m^{(\alpha)} \delta_{mn}, \\ \gamma_m^{(\alpha)} &= \frac{2^{2\alpha+1}[\Gamma(\alpha+1)]^2 m!}{(2m+2\alpha+1)\Gamma(m+2\alpha+1)}. \end{aligned} \tag{20}$$

Proof. This follows directly from the definition of $G_m^{(\alpha)}$ and known properties of Jacobi polynomials. □

Remark. In (20), $\gamma_m^{(\alpha)}$ has a special point at $m = 0$, $\alpha = -1/2$ which must be evaluated by setting m to zero before taking the limit, yielding

$$\lim_{\alpha \to -1/2} \gamma_0^{(\alpha)} = \lim_{\alpha \to -1/2} \frac{2^{2\alpha+1}[\Gamma(\alpha+1)]^2}{\Gamma(2\alpha+2)} = \pi.$$

Corollary A.2. *For functions $f, g \in L_\alpha^2$ defined as combinations of the basis functions $G_m^{(\alpha)}$, $f = \sum_{m=0}^{\infty} a_m G_m^{(\alpha)}$ and $g = \sum_{m=0}^{\infty} b_m G_m^{(\alpha)}$, the inner product and L^2-distance have the form*

$$\begin{aligned} \langle f, g \rangle_\alpha &= \int_{-1}^{1} w^{(\alpha)}(x) f(x) g(x) dx \\ &= \textstyle\sum_{m=0}^{\infty} \gamma_m^{(\alpha)} a_m b_m, \\ \|f - g\|_\alpha^2 &= \int_{-1}^{1} w^{(\alpha)}(x) (f(x) - g(x))^2 dx \\ &= \textstyle\sum_{m=0}^{\infty} \gamma_m^{(\alpha)} (a_m - b_m)^2. \end{aligned} \tag{21}$$

Proof. This follows directly from (20). □

Theorem A.3. *The polynomials $G_m^{(\alpha)}$ satisfy the recurrence relation*

$$\begin{aligned} &G_0^{(\alpha)}(x) = 1\,, \quad G_1^{(\alpha)}(x) = x, \\ &(m+2\alpha) G_m^{(\alpha)}(x) = (2m+2\alpha-1) x G_{m-1}^{(\alpha)}(x) - (m-1) G_{m-2}^{(\alpha)}(x). \end{aligned} \tag{22}$$

Proof. This follows directly from the definition of $G_m^{(\alpha)}$ and known properties of Jacobi polynomials. □

Remark. Based on (22) and the uniqueness of orthogonal polynomials, the set $\{G_m^{(\alpha)}\}$ contains as special cases the Chebyshev polynomials T_m, Legendre polynomials P_m and at the limit $\alpha \to \infty$ the set of monomials: for $x \in [-1, 1]$,

$$\begin{aligned} G_m^{(-1/2)}(x) &= T_m(x), \\ G_m^{(0)}(x) &= P_m(x), \\ \lim_{\alpha\to\infty} G_m^{(\alpha)}(x) &= x^m. \end{aligned}$$

Theorem A.4. *Let $G_m^{(\alpha)}$ and $G_n^{(\beta)}$ be normalized ultraspherical polynomials as defined above. Then the connection coefficients $C_{mn}^{(\alpha\beta)}$,*

$$G_m^{(\alpha)} = \sum_{n=0}^{\infty} C_{mn}^{(\alpha\beta)} G_n^{(\beta)}, \tag{23}$$

have the form

$$\begin{aligned} C_{mn}^{(\alpha\beta)} &= 0 \quad \textit{if } n > m \textit{ or } m+n \textit{ is odd}, \\ C_{00}^{(\alpha\beta)} &= C_{11}^{(\alpha\beta)} = 1, \\ C_{mn}^{(\alpha\beta)} &= \frac{m!}{n!k!}\frac{d_h(\alpha)}{d_h(\beta)} r_{mn}(\alpha,\beta)\,, \quad m \geq 2, \\ d_h(\alpha) &= \frac{2^{h+2\alpha}}{\sqrt{\pi}}\Gamma(\alpha+1)\frac{\Gamma(h+\alpha+\frac{1}{2})}{\Gamma(h+2\alpha+1)}, \\ r_{mn}(\alpha,\beta) &= (\alpha-\beta)_k \frac{2n+2\beta+1}{m+n+2\beta+1}\frac{\Gamma(n+2\beta+1)}{\Gamma(m+2\alpha+1)}\frac{\Gamma(h+2\alpha+1)}{\Gamma(h+2\beta+1)}, \end{aligned} \tag{24}$$

where $k = (m-n)/2$ and $h = (m+n)/2$ and the Pochhammer symbol $(\alpha)_k = \alpha(\alpha+1)\cdots(\alpha+k-1)$ represents the rising factorial.

Proof. Equation (23) can be expanded by applying (22), which yields

$$\begin{gathered} (m+2\alpha)\textstyle\sum_n C_{mn}^{(\alpha\beta)} G_n^{(\beta)} = \\ (2m+2\alpha-1)\textstyle\sum_n C_{m-1,n}^{(\alpha\beta)} \left[\dfrac{n+2\beta+1}{2n+2\beta+1} G_{n+1}^{(\beta)} + \dfrac{n}{2n+2\beta+1} G_{n-1}^{(\beta)}\right] \\ -\,(m-1)\textstyle\sum_n C_{m-2,n}^{(\alpha\beta)} G_n^{(\beta)}. \end{gathered}$$

Comparing the coefficients of $G_n^{(\beta)}$ for each n yields a recurrence relation for the elements of $C^{(\alpha\beta)}$,

$$\begin{aligned}
C_{mn}^{(\alpha\beta)} &= 0 \quad \text{if } n > m \text{ or } m+n \text{ is odd}, \\
C_{00}^{(\alpha\beta)} &= C_{11}^{(\alpha\beta)} = 1, \\
(m+2\alpha)C_{m0}^{(\alpha\beta)} &= \frac{2m+2\alpha-1}{2\beta+3}C_{m-1,1}^{(\alpha\beta)} - (m-1)C_{m-2,0}^{(\alpha\beta)}, \\
(m+2\alpha)C_{mn}^{(\alpha\beta)} &= (n+2\beta)\frac{2m+2\alpha-1}{2n+2\beta-1}C_{m-1,n-1}^{(\alpha\beta)} \\
&+ (n+1)\frac{2m+2\alpha-1}{2n+2\beta+3}C_{m-1,n+1}^{(\alpha\beta)} \\
&- (m-1)C_{m-2,n}^{(\alpha\beta)} \\
&\quad \text{for } m \geq 2, n \geq 1,\ n = m, m-2, \ldots.
\end{aligned} \tag{25}$$

Since there is only one non-zero term in the recurrence relation for diagonal elements, the respective relation reduces to

$$\begin{aligned}
C_{mm}^{(\alpha\beta)} &= \delta_m(\alpha,\beta)C_{m-1,m-1}^{(\alpha\beta)}, \\
\delta_m(\alpha,\beta) &= \frac{(m+2\beta)(2m+2\alpha-1)}{(m+2\alpha)(2m+2\beta-1)}.
\end{aligned} \tag{26}$$

Similarly, combining the terms in (25) for element $C_{m0}^{(\alpha\beta)}$ and then progressively for the rest of the elements of the respective side diagonal $C_{m+k,k}^{(\alpha\beta)}$, $k \geq 0$ and repeating this for $m = 2, 4, \ldots$ yields an anti-diagonal recurrence relation

$$\begin{aligned}
C_{mn}^{(\alpha\beta)} &= \rho_{mn}(\alpha,\beta)C_{m-1,n+1}^{(\alpha\beta)}, \\
\rho_{mn}(\alpha,\beta) &= \frac{m(n+1)}{m-n}\,\frac{2n+2\beta+1}{n+2\beta+1}\,\frac{(m-n+2\alpha-2\beta-2)}{(m+2\alpha)(2n+2\beta+3)}
\end{aligned} \tag{27}$$

for $m \geq 2$ and $n = m-2, m-4, \ldots$. The formulas (24) for all $C_{mn}^{(\alpha\beta)}$, $m \geq 2$, then directly follow from

$$C_{mn}^{(\alpha\beta)} = \prod_{i=2}^{h}\delta_i(\alpha,\beta)\prod_{j=0}^{k}\rho_{h+j,h-j}(\alpha,\beta). \tag{28}$$

□

Remark. The coefficients in (24) have particular forms for certain limits and special values. They can be determined by taking the respective limits in (26) and (27) and then reconstructing $C_{mn}^{(\alpha\beta)}$ through (28). For completeness,

$$\begin{aligned}
\lim_{\alpha\to\infty} d_h(\alpha) &= 1,\\
\lim_{\alpha\to\infty} r_{mn}(\alpha,\beta) &= \frac{1}{2^k}\frac{2n+2\beta+1}{m+n+2\beta+1}\frac{\Gamma(n+2\beta+1)}{\Gamma(h+2\beta+1)},\\
\lim_{\beta\to\infty} r_{mn}(\alpha,\beta) &= \frac{1}{(-2)^k}\frac{\Gamma(h+2\alpha+1)}{\Gamma(m+2\alpha+1)},\\
\lim_{\beta\to-1/2} r_{m0}(\alpha,\beta) &= \frac{(\alpha+1)_k}{2k!}\frac{\Gamma(k+2\alpha+1)}{\Gamma(m+2\alpha+1)},\\
\lim_{\beta\to-1/2}\lim_{\alpha\to\infty} r_{m0}(\alpha,\beta) &= \frac{1}{2^{k+1}k!}.
\end{aligned}$$

Remark. Given a finite M, the coefficients $C_{mn}^{(\alpha\beta)}$, $m,n = 1,\ldots,M-1$ define change of coordinates between the bases $\{G_m^{(\alpha)}\}$ and $\{G_n^{(\beta)}\}$, that is, if

$$\sum_{m=0}^{M-1} a_m G_m^{(\alpha)} = \sum_{n=0}^{M-1} b_n G_n^{(\beta)},$$

then $b_n = \sum_{m=0}^{M-1} C_{mn}^{(\alpha\beta)} a_m$.

Lemma A.5. *Let $\alpha \geq \beta$. Then $C_{mn}^{(\alpha\beta)} \geq 0$ for all m, $n \in \mathbb{N}$.*

Proof. The sign of $C_{mn}^{(\alpha\beta)}$ is determined by $(\alpha-\beta)_k$ in (24). Since $\alpha \geq 0$ impies that $(\alpha)_k \geq 0$ for all $k \in \mathbb{N}$, Lemma A.5 directly follows. □

Lemma A.6. $\sum_{n=0}^{\infty} C_{mn}^{(\alpha\beta)} = 1$ *for all $m \in \mathbb{N}$.*

Proof. Summing over n in (25) yields

$$(m+2\alpha)\sum_{n=0}^{\infty} C_{mn}^{(\alpha\beta)} = (m+2\alpha)\sum_{n=0}^{m} C_{mn}^{(\alpha\beta)} = (2m+2\alpha-1)\sum_{n=0}^{m-1} C_{m-1,n}^{(\alpha\beta)} - (m-1)\sum_{n=0}^{m-2} C_{m-2,n}^{(\alpha\beta)}$$

Since $C_{00}^{(\alpha\beta)} = \sum_{n=0}^{1} C_{1n}^{(\alpha\beta)} = 1$, Lemma A.6 directly follows by induction. □

Theorem A.7. $\left|G_m^{(\alpha)}(x)\right| \leq 1$ *for all $x \in [-1,1]$ and $\alpha \geq -1/2$.*

Proof. As mentioned above, $G_n^{(-\frac{1}{2})} = T_n$. Since Chebyshev polynomials satisfy the equation $T_n(\cos(\theta)) = \cos(n\theta)$ it follows that for $x \in [-1,1]$ and $\alpha \geq -1/2$,

$$\begin{aligned}
\left|G_m^{(\alpha)}(x)\right| &= \left|\textstyle\sum_{n=0}^{m} C_{mn}^{(\alpha,-\frac{1}{2})} G_n^{(-\frac{1}{2})}(x)\right|\\
&\leq \textstyle\sum_{n=0}^{m} \left|C_{mn}^{(\alpha,-\frac{1}{2})} G_n^{(-\frac{1}{2})}(x)\right| = \textstyle\sum_{n=0}^{m} C_{mn}^{(\alpha,-\frac{1}{2})} \left|G_n^{(-\frac{1}{2})}(x)\right|\\
&= \textstyle\sum_{n=0}^{m} C_{mn}^{(\alpha,-\frac{1}{2})} \left|\cos(n\arccos(x))\right| \leq \textstyle\sum_{n=0}^{m} C_{mn}^{(\alpha,-\frac{1}{2})} = 1
\end{aligned}$$

by applying (23), the triangle inequality, Lemma A.5, the Chebyshev identity, the definition of $\cos(x)$ and Lemma A.6, respectively. □

B Proof of Theorem 2.1

The proof of Theorem 2.1 has the usual structure of a consistency argument for a minimizer or M-estimator commonly used for example in connection with non-linear econometric models. A good account of the relevant theory can be found in Pötscher and Prucha (1997).

Proof. Let $(\Omega, \mathcal{F}, \mathbb{P})$ be the underlying probability space. The estimator $\hat{\boldsymbol{\theta}}_N : \Omega \to \boldsymbol{\Theta}$ is obtained by minimizing

$$G_N(\omega, \boldsymbol{\theta}) = \|f(\cdot|\boldsymbol{\theta}, M) - \hat{f}(\cdot|M)\|_0^2 = \int_{-1}^{1} [f(x|\boldsymbol{\theta}, M) - \hat{f}(x|M)]^2 dx \tag{29}$$

with respect to $\boldsymbol{\theta}$ (cf. (18)). To simplify notation we leave out ω from the right hand side of (29) and also in most of what follows. Note that M is in fact a random variable with values in $\{2, 3, \ldots, R(N)\}$ that implements the Hart cut-off (cf. (11)). It is clear that $G_N(\cdot, \boldsymbol{\theta})$ is measurable when $\boldsymbol{\theta}$ is fixed. It also follows easily from the assumptions and the Lebesgue dominated convergence theorem that for each $\omega \in \Omega$ the function $G_N(\omega, \cdot)$ is continous on $\boldsymbol{\Theta}$. Lemma 3.4 of Pötscher and Prucha (1997) then guarantees that the minimizer $\hat{\boldsymbol{\theta}}_N$ of (29) can chosen to be measurable, that is, a random variable.

Define next

$$G(\boldsymbol{\theta}) = \|f(\cdot|\boldsymbol{\theta}) - f\|_0^2 = \int_{-1}^{1} [f(x|\boldsymbol{\theta}) - f(x)]^2 dx. \tag{30}$$

It follows again from the assumptions and the dominated convergence theorem that G is a continuos function on $\boldsymbol{\Theta}$. Also, $\boldsymbol{\theta}_0$ is the unique minimizer of G because $f = f(\cdot|\boldsymbol{\theta}_0)$. We will show that

$$\max_{\boldsymbol{\theta} \in \boldsymbol{\Theta}} |G_N(\cdot, \boldsymbol{\theta}) - G(\boldsymbol{\theta})| \xrightarrow{P} 0. \tag{31}$$

It then follows from Lemma 3.1 of Pötscher and Prucha (1997) that $\hat{\boldsymbol{\theta}}_N \overset{P}{\to} \boldsymbol{\theta}_0$, as claimed. Note that by standard arguments the maximum on the left hand side of (31) is a random variable because of continuity in $\boldsymbol{\theta}$, measurability of $G_N(\cdot,\boldsymbol{\theta})$, and the compactness of $\boldsymbol{\Theta}$ (e.g. Lemma A3 of Pötscher and Prucha (1997)).

Following the proof of Hart (1985) we show first that the cut-off $M \overset{P}{\to} \infty$ as $N \to \infty$. Thus, let $L \geq 2$ be an integer. We show that

$$\lim_{N\to\infty} \mathbb{P}(M > L) = 1. \tag{32}$$

Select $l \geq L$ such that $d_l \neq 0$ and let N be so large that $l+1 \leq R(N)$. Then, for any $J \in \{2, 3, \ldots, L\}$ we have by (11) that

$$H(l+1) - H(J) = \sum_{m=J}^{l} \gamma_m \left[2\widehat{\mathrm{Var}}(\hat{d}_m) - \hat{d}_m^2 \right] \geq \gamma_l \left[2\widehat{\mathrm{Var}}(\hat{d}_l) - \hat{d}_l^2 \right]. \tag{33}$$

Here, by the strong law of large numbers, $\widehat{\mathrm{Var}}(\hat{d}_l) \to 0$ and $\hat{d}_l^2 \to d_l^2 > 0$ with probability 1 and therefore (32) holds.

Let us then establish the consistency of $\hat{f}(\cdot|M)$. Using the orthogonality of Legendre polynomials we have

$$\begin{aligned} \|\hat{f}(\cdot|M) - f\|_0^2 &= \int_{-1}^{1} \left[\hat{f}(x|M) - f(x) \right]^2 dx \\ &= \sum_{m=0}^{M-1} \gamma_m (\hat{d}_m - d_m)^2 + \sum_{m=M}^{\infty} \gamma_m d_m^2. \end{aligned}$$

Here $\sum_{m=M}^{\infty} \gamma_m d_m^2 \overset{P}{\to} 0$ because $M \overset{P}{\to} \infty$ and $\sum_{m=0}^{\infty} \gamma_m d_m^2 = \|f\|_0^2 < \infty$. To prove that the first term also converges to zero in probability we estimate its expectation,

$$\mathbb{E}\left\{ \sum_{m=0}^{M-1} \gamma_m (\hat{d}_m - d_m)^2 \right\} \leq \mathbb{E}\left\{ \sum_{m=0}^{R(N)-1} \gamma_m (\hat{d}_m - d_m)^2 \right\} = \sum_{m=0}^{R(N)-1} \gamma_m \mathrm{Var}(\hat{d}_m).$$

Here

$$\mathrm{Var}(\hat{d}_m) = \frac{1}{N\gamma_m^2} \mathrm{Var}(P_m(X_1)) \leq \frac{1}{N\gamma_m^2}$$

because $|P_m(x)| \leq 1$ for all $x \in [-1, 1]$ (Theorem A.7). Substituting $\gamma_m = 2/(2m+1)$ we then get after some simple computations that

$$\mathbb{E}\left\{ \sum_{m=0}^{M-1} \gamma_m (\hat{d}_m - d_m)^2 \right\} \leq \frac{R(N)^2}{2N} \underset{N\to\infty}{\longrightarrow} 0$$

because $R(N) = o(\sqrt{N})$. It now follows from Markov's inequality that

$$\sum_{m=0}^{M-1} \gamma_m (\hat{d}_m - d_m)^2 \xrightarrow{P} 0$$

and therefore

$$\|\hat{f}(\cdot|M) - f\|_0^2 = \int_{-1}^{1} \Big[\hat{f}(x|M) - f(x)\Big]^2 \, dx \xrightarrow{P} 0. \tag{34}$$

Let us now turn to (31). Combining the integrals, using the identity $a^2 - b^2 = (a-b)(a+b)$ and the Schwarz inequality we get

$$\begin{aligned} &|G_N(\omega, \boldsymbol{\theta}) - G(\boldsymbol{\theta})| \le \\ &\quad \|[f(\cdot|\boldsymbol{\theta}, M) - f(\cdot|\boldsymbol{\theta})] - [\hat{f}(\cdot|M) - f]\|_0 \|f(\cdot|\boldsymbol{\theta}, M) + f(\cdot|\boldsymbol{\theta}) - \hat{f}(\cdot|M) - f\|_0. \end{aligned} \tag{35}$$

Let us first show that the second factor in this upper bound remains uniformly bounded for $\boldsymbol{\theta} \in \boldsymbol{\Theta}$ as $N \to \infty$. Using the triangle inequality it is bounded by

$$\begin{aligned} &\|f(\cdot|\boldsymbol{\theta}, M)\|_0 + \|f(\cdot|\boldsymbol{\theta})\|_0 + \|\hat{f}(\cdot|M)\|_0 + \|f\|_0 \le \\ &\quad \|f(\cdot|\boldsymbol{\theta}, M)\|_0 + \|f(\cdot|\boldsymbol{\theta})\|_0 + \|\hat{f}(\cdot|M) - f\|_0 + 2\|f\|_0. \end{aligned}$$

By (34) the third term converges to zero in probability and the fourth term is a constant. Further,

$$\|f(\cdot|\boldsymbol{\theta}, M)\|_0^2 = \sum_{m=0}^{M-1} \gamma_m d_m(\boldsymbol{\theta})^2 \le \sum_{m=0}^{\infty} \gamma_m d_m(\boldsymbol{\theta})^2 = \|f(\cdot|\boldsymbol{\theta})\|_0^2$$

and because by the assumptions of the theorem $(x, \boldsymbol{\theta}) \mapsto f(x|\boldsymbol{\theta})$ is bounded, the second factor in (35) has an upper bound $C + \|\hat{f}(\cdot|M) - f\|_0 \xrightarrow{P} C$ where $C > 0$ is a constant.

The first factor in (35) is bounded by

$$\|f(\cdot|\boldsymbol{\theta}, M) - f(\cdot|\boldsymbol{\theta})\|_0 + \|\hat{f}(\cdot|M) - f\|_0.$$

Therefore, to prove (31) it is by (34) enough to show that the first term converges to zero in probability uniformly in $\boldsymbol{\theta}$. By the continuity of $x \mapsto \frac{\partial}{\partial x} f(x, \boldsymbol{\theta})$ it follows from Theorem 6.2 of DeVore and Lorentz (1993) that for each $\boldsymbol{\theta} \in \boldsymbol{\Theta}$,

$$\|f(\cdot|\boldsymbol{\theta}, M) - f(\cdot|\boldsymbol{\theta})\|_0 \le \frac{\pi}{2M} \|\frac{\partial}{\partial x} f(\cdot, \boldsymbol{\theta})\|_0$$

and by the continuity of $(x,\boldsymbol{\theta}) \mapsto \frac{\partial}{\partial x} f(x,\boldsymbol{\theta})$ the right hand side is bounded by C'/M where the constant C' is independent of $\boldsymbol{\theta}$. The proof is then complete because $M \overset{P}{\to} \infty$. □

Remark. Theorem 2.1 assumes that the true density f underlying the data corresponds to a unique member $f(\cdot|\boldsymbol{\theta}_0)$ of the parametric family $\{f(\cdot|\boldsymbol{\theta})|\boldsymbol{\theta} \in \boldsymbol{\Theta}\}$. This may not always be the case, a prime example being mixture models where numbering of the components alone is a source of non-identifiability. The proof Theorem 2.1 can be extended to such non-identifiable situations using, Lemma 4.2 of Pötscher and Prucha (1997) instead of their Lemma 3.1. Then we only assume that $\boldsymbol{\Theta}_0 = \{\boldsymbol{\theta} \in \boldsymbol{\Theta}|f(\cdot|\boldsymbol{\theta}) = f\}$ is non-empty and consistency of the minimization estimator holds in the sense that $\min_{\boldsymbol{\theta}\in\boldsymbol{\Theta}_0} \|\hat{\boldsymbol{\theta}}_N - \boldsymbol{\theta}\| \xrightarrow{P} 0$.

Let then $\{g(\cdot|\boldsymbol{\psi})|\boldsymbol{\psi} \in \boldsymbol{\Psi}\}$ be a parametric family of densities on $[-1,1]$, $\boldsymbol{\Psi} \subset \mathbb{R}^p$, and consider a mixture model

$$\sum_{i=1}^{K} p_i g(\cdot|\boldsymbol{\psi}_i) = \sum_{i=1}^{K} f(\cdot|\boldsymbol{\theta}_i),$$

where $0 \le p_1,\ldots,p_K \le 1$, $\sum_{i=1}^K p_i = 1$, $\boldsymbol{\psi}_i,\ldots,\boldsymbol{\psi}_K \in \boldsymbol{\Psi}$, and $\boldsymbol{\theta}_i = (p_i,\boldsymbol{\psi}_i)$, $f(\cdot|\boldsymbol{\theta}_i) = p_i g(\cdot|\boldsymbol{\psi}_i)$. Suppose that Algorithm 3.1 is used to fit such a mixture model. Then, at the ith stage one finds a parameter vector $\boldsymbol{\theta}_i = \hat{\boldsymbol{\theta}}_{iN}$ which minimizes

$$\|f(\cdot|\boldsymbol{\theta}_i, M) - \hat{f}_i(\cdot|M)\|_0^2, \tag{36}$$

where

$$\hat{f}_i(\cdot|M) = \hat{f}(\cdot|M) - \sum_{j=1}^{i-1} f(\cdot|\hat{\boldsymbol{\theta}}_j).$$

Suppose that the density f underlying the is of the form $f = \sum_{i=1}^K f(\cdot|\boldsymbol{\theta}_{i0})$ for some $\boldsymbol{\theta}_{01},\ldots,\boldsymbol{\theta}_{0K} \in \boldsymbol{\Psi}$. Minimization of (36) can then be thought to approximate minimization of

$$\|f(\cdot|\boldsymbol{\theta}_i) - f_i\|_0^2, \tag{37}$$

where

$$f_i = f - \sum_{j=1}^{i-1} f(\cdot|\boldsymbol{\theta}_{0j}).$$

Theorem B.1. *Suppose that both* $(x, \boldsymbol{\psi}) \mapsto g(x, \boldsymbol{\psi})$ *and* $(x, \boldsymbol{\psi}) \mapsto \frac{\partial}{\partial x} g(x, \boldsymbol{\psi})$ *are continuous and, for each* $i = 1, \ldots, K$*, let* $\boldsymbol{\Theta}_i \subset [0,1] \times \boldsymbol{\Psi}$ *be a compact set such that (37) has a unique minimizer* $\boldsymbol{\theta}_{i0} \in \boldsymbol{\Theta}_i$*. Assume further that in the Legendre expansion (9)* $d_m \neq 0$ *for infinitely many* m *and that the Hart criterion (11) is applied by selecting the optimal* M *from a set* $\{2, 3, \ldots, R(N)\}$*, where* $R(N) \to \infty$ *and* $R(N) = o(\sqrt{N})$ *as the sample size* N *tends to infinity. Then for* $i = 1, \ldots, K$ *we have that* $\hat{\boldsymbol{\theta}}_{iN} \xrightarrow{P} \boldsymbol{\theta}_{i0}$ *as* $N \to \infty$*, where* $\hat{\boldsymbol{\theta}}_{iN}$ *minimizes (36) in* $\boldsymbol{\Theta}_i$*.*

We omit the proof because it consists of a straightforward induction on the mixture component index i that at each step essentially repeats the arguments of the previous proof. In practice, the existence of the parameter subspaces $\boldsymbol{\Theta}_i$ and unique local minimizers of (37) in them could be expected to hold at least if the components $f(\cdot|\boldsymbol{\theta}_{0i})$ do not overlap too much and good initial values are available for the true parameters $\boldsymbol{\theta}_{0i}$.

References

Abramowitz, M. and Stegun, I. A., editors (1972). *Handbook of Mathematical Functions with Formulas, Graphs, and Mathematical Tables.* Dover Publications, New York, 10th printing edition.

Berrut, J.-P. and Trefethen, L. N. (2004). Barycentric Lagrange interpolation. *SIAM Review*, 46(3):501–517.

Chan, T. F., Golub, G. H., and LeVeque, R. J. (1979). Updating formulae and a pairwise algorithm for computing sample variances. Technical report, Stanford University, Department of Computer Science. Technical Report STAN-CS-79-773.

Dempster, A. P., Laird, N. M., and Rubin, D. B. (1977). Maximum likelihood from incomplete data via the EM algorithm. *Journal of the Royal Statistical Society. Series B*, 39(1):1–38.

DeVore, R. and Lorentz, G. (1993). *Constructive Approximation.* Die Grundlehren der mathematischen Wissenschaften in Einzeldarstellungen. Springer.

Devroye, L. (1987). *A Course in Density Estimation.* Birkhäuser, Boston.

Devroye, L. and Györfi, L. (1985). *Nonparametric Density Estimation: The* L^1 *View.* John Wiley, New York.

Efromovich, S. (2010). Orthogonal series density estimation. *Wiley Interdisciplinary Reviews: Computational Statistics*, 2:467–476.

Hart, J. D. (1985). On the choice of a truncation point in Fourier series density estimation. *Journal of Statistical Computation and Simulation*, 21:95–116.

Holmström, L. and Klemelä, J. (1992). Asymptotic bounds for the expected L^1 error of a multivariate kernel density estimator. *Journal of Multivariate Analysis*, 42(2):245–266.

Klemelä, J. (2009). *Smoothing of multivariate data: density estimation and visualization.* Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.

Krzyżak, A. and Pawlak, M. (1982). Estimation of a multivariate density by orthogonal series. In Grossmann, W., Pflug, G. C., and Wertz, W., editors, *Probability and Statistical Inference*, pages 211–221. Springer Netherlands.

Levenberg, K. (1944). A method for the solution of certain non-linear problems in least squares. *Quarterly Journal of Applied Mathmatics*, 2:164–168.

Marquardt, D. (1963). An algorithm for least-squares estimation of nonlinear parameters. *SIAM Journal on Applied Mathematics*, 11(2):431–441.

McLachlan, G. and Peel, D. (2000). *Finite Mixture Models.* Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.

Mengersen, K., Robert, C., and Titterington, M. (2011). *Mixtures: Estimation and Applications.* Wiley Series in Probability and Statistics. Wiley, Hoboken, NJ.

Pötscher, B. M. and Prucha, I. R. (1997). *Dynamic Nonlinear Econometric Models.* Springer.

Powell, M. J. D. (1964). An efficient method for finding the minimum of a function of several variables without calculating derivatives. *The Computer Journal*, 7(2):155–162.

Prakasa Rao, B. (1983). *Nonparametric Functional Estimation.* Probability and Mathematical Statistics: A Series of Monographs and Textbooks. Academic Press.

Redner, R. A. and Walker, H. F. (1984). Mixture densities, maximum likelihood and the EM algorithm. *SIAM Review*, 26(2):195–239.

Scott, D. W. (1992). *Multivariate Density Estimation: Theory, Practice, and Visualization.* John Wiley & Sons, Inc., New York.

Scott, D. W. (2001). Parametric statistical modeling by minimum integrated squared error. *Technometrics*, 43(3):274–285.

Silverman, B. (1986). *Density Estimation for Statistics and Data Analysis*. Chapman and Hall, London.

Szegő, G. (1959). *Orthogonal Polynomials*, volume 23 of *Colloquium publications*. American Mathematical Society.

Wand, M. P. and Jones, M. C. (1995). *Kernel smoothing*. Chapman & Hall, London.