Comput Stat (2012) 27:757–777
DOI 10.1007/s00180-011-0289-6

ORIGINAL PAPER

Density estimation and comparison with a penalized mixture approach

Christian Schellhase · Göran Kauermann

Received: 3 March 2011 / Accepted: 18 October 2011 / Published online: 2 November 2011

Abstract The paper presents smooth estimation of densities utilizing penalized splines. The idea is to represent the unknown density by a convex mixture of basis densities, where the weights are estimated in a penalized form. The proposed method extends the work of Komárek and Lesaffre (Comput Stat Data Anal 52(7):3441–3458, 2008) and allows for general density estimation. Simulations show a convincing performance in comparison to existing density estimation routines. The idea is extended to allow the density to depend on some (factorial) covariate. Assuming a binary group indicator, for instance, we can test on equality of the densities in the groups. This provides a smooth alternative to the classical Kolmogorov-Smirnov test or an Analysis of Variance and it shows stable and powerful behaviour.

Keywords Density estimation · Mixture density estimation · Penalized spline smoothing · ANOVA

1 Introduction

Density estimation has a long standing tradition in statistics and the different routines can be roughly categorized in four partly overlapping approaches. (a) First and most prominent there is kernel density estimation which traces back to ideas of Nadaraya (1964) and Watson (1964), see also Nadaraya (1974). The method

C. Schellhase
Department for Business Administration and Economics, Centre for Statistics,
Bielefeld University, Postfach 10 01 31, 33501 Bielefeld, Germany

G. Kauermann (✉)
Department of Statistics, Ludwig-Maximilians-University Munich,
Ludwigstrasse 33, 80539 Munich, Germany
e-mail: goeran.kauermann@stat.uni-muenchen.de

is well established and extensively discussed in e.g. Wand and Jones (1995) or Simonoff (1996). (b) A second approach results by writing the unknown density as

$$\hat{f}(y) = \exp\{\eta(y)\} / \int \exp\{\eta(z)\}\, dz \tag{1}$$

with $\eta(\cdot)$ unknown but smooth function which is estimated using spline technology. This approach traces back to Good and Gaskins (1971), see also Silverman (1982) and the idea has been further developed by Gu (1993) or Dias (1998), see also Gu and Wang (2003). (c) A third approach results by extending and smoothing the classical histogram as originally suggested by Boneva et al. (1971). Following this idea (Lindsey 1974a,b) suggests density estimation by transferring the density estimation problem to a regression estimation scenario, with the number of observations per bin in the histogram as Poisson count, see also Efron and Tibshirani (1996). Eilers and Marx (1996) make use of the idea using penalized spline smoothing, see also Ruppert et al. (2003). The spline approach and the Poisson approach (c) are thereby closely related which results by approximating the integral in (1) with a rectangular method. (d) A fourth line of density estimation has been suggested by using a mixture approach. In this case, the unknown density results by finite mixture of densities components. These mixture components are usually built from known distributions (e.g. normal) with unknown parameters. This yields the classical mixture models discussed extensively in McLachlan and Peel (2000), see also Young et al. (2009), Li and Barron (1999) or Fraley and Raftery (2002). (e) Another approach to estimate the unknown density is the log-spline approach (see Koo et al. 1999), modelling the log-density function by (almost cubic) splines using maximum likelihood estimation and Newton-Raphson method to compute optimal coefficients. (f) A sixth idea to estimate densities is tackled using wavelets, expanding the unknown density in terms of a wavelet expansion (see e.g. Hall and Patil 1995, Nason and Silverman 1999 or Nason 2008). Our approach (g) presented in this paper distinguishes from the classical mixture model in two ways. First, we take completely specified mixture components, that is not only the distribution type, but also the parameters are fixed. Secondly, the number of mixture components is chosen in a lavish way and we impose a penalty to achieve smooth density fits. Ghidey et al. (2004) have proposed to use a finite but penalized mixture of Gaussian densities for the estimation of a random effect distribution in a linear mixed model. The idea has been extended and further developed in a number of papers which include Komárek et al. (2005), Komárek (2006) and Komárek and Lesaffre (2008). The idea of Komárek (2006) shows also similarities to the approach of Babu et al. (2002), who approximate the density with a mixture of Bernstein polynomials. In this paper we generalize the original idea of Komárek and Lesaffre (2008) to univariate density estimation. Extending the mixture to a continuous mixture has recently been proposed by Liu et al. (2009).

In this paper we follow (g) using finite mixture densities for the smooth estimation of an unknown density. The collection of the densities used in the mixture in fact plays the role of a basis and the weights correspond to basis coefficients. The weights itself can be fitted with penalized techniques to obtain a smooth density fit.

In principle, any type of mixture density can be used and there is no requirement for Gaussian mixtures. In this paper we make use of a mixture of B-spline basis functions normed to be densities. This allows to theoretically investigate the properties of the fit and also guarantees stable numerical performance. To achieve smoothness we make use of penalized spline smoothing in the style of Ruppert et al. (2003), see also O'Sullivan (1986) and Eilers and Marx (1996). With the link between penalized spline smoothing and mixed models (see Wand 2003) the method shows its full flexibility and versatility as demonstrated in the commendable survey recently composed by Ruppert et al. (2009).

A general question in penalized spline smoothing concerns the number of splines used for fitting. A rule of thumb has been suggested in Ruppert (2002) who shows that the number of splines does not affect the fit once sufficient splines have been chosen, which is usually a small number compared to the sample size regardless of the form of the function to be fitted. The same conclusion is drawn in Kauermann and Opsomer (2011) who make use of the link between mixed models and penalized spline smoothing. Allowing the spline dimension to depend on the sample size provides an asymptotic framework which has been investigated in Li and Ruppert (2008), Kauermann et al. (2009) and Claeskens et al. (2009). Though these results shed some light on the theoretical properties of penalized spline estimation, there is hardly any practical impact and the rule of thumb for choosing the spline dimension (see Ruppert 2002) is still recommendable.

We also extend the classical density estimation problem by allowing the density to depend on some covariates x, say. That is to say we let the mixture weight depend on exogenous quantities. We restrict this modelling exercise to factorial quantities x, which allows us to compare densities in two (or more) groups. As example we look at the return of stocks of different companies and different years. The idea may be seen as nonparametric Analysis of Variance (ANOVA) and follows closely the testing framework for the Kolmogorov-Smirnov test.

The scientific contributions of the paper are twofold. First, we show how a density can be estimated with a penalized mixture of basis densities. The novel routine is contrasted in simulations to the various competitors described above, that is (a) kernel density estimation, (b) spline based density estimation, (c) Poisson approximated density estimation and (d) classical mixture density estimation, (e) log-spline density estimation and (f) wavelet density estimation. As will be seen, the performance of the available routines is quite diverse and the penalized mixture approach performs promising. The second contribution of the paper is to explore penalized mixture density estimation in testing scenarios when comparing distributions in two (or more) groups.

This paper is organized as follows. In Sect. 2 we introduce the idea of density estimation with penalized splines. Section 3 demonstrates the fitting in simulations and an example. In Sect. 4 we extend the idea by allowing the density to depend on covariate x, which is demonstrated in a simulation and an example in Sect. 5. Section 6 concludes the paper.

2 Penalized density

2.1 Mixture modelling and penalized estimation

We are interested in nonparametric estimation of the density of the univariate random variable y. We therefore approximate the density of y as a mixture of densities

$$f_K(y) = \sum_{k=-K}^{K} c_k \boldsymbol{\phi}_k(y), \tag{2}$$

where $\boldsymbol{\phi}_k(y)$ are subsequently called basis densities. The weights c_k in (2) are parameterized as

$$c_k(\boldsymbol{\beta}) = \frac{\exp(\beta_k)}{\sum_{k=-K}^{K} \exp(\beta_k)} \tag{3}$$

with $\beta_0 \equiv 0$ for identifiability and $\boldsymbol{\beta} = (\beta_{-K}, \ldots, \beta_{-1}, \beta_1, \ldots, \beta_K)$ so that $\int f_K(y)\mathrm{d}y = 1$. The basis densities are thereby known and fixed density functions with specified parameters. We assume that $\boldsymbol{\phi}_k(y)$ is continuous on its support and converges to zero at the boundary of the support. A possible choice for the basis densities is to take $\boldsymbol{\phi}_k(y)$ as Gaussian density with fixed mean μ_k and variance σ_k^2, where the mean values μ_k may be called the knots of the basis. Numerically more stable and theoretically more appealing are B-spline densities which are standard B-splines (see de Boor 1978) normed to be densities. We will subsequently notate the knots at which the basis densities are located as μ_k with k running from $-K$ to K for convenience. We assume, that the knots μ_k cover the range of observed values of y and their location is fixed. A typical and simple setting is to have equidistant knots which will be assumed subsequently. Apparently, the number of knots plays an important role in terms of bias and variance and a small number K will lead to biased estimates while for large values of K the estimates will be wiggled. We will therefore utilize the idea of penalized spline smoothing by choosing the number of knots K in a lavish and generous way and impose a penalty to achieve smoothness. The penalty is put on the basis coefficients β_k by penalizing the variation of c_k over k. Assuming independent observations $y_i, i = 1, ..., n$, the log likelihood takes the form

$$l(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\log \sum_{k=-K}^{K} c_k(\boldsymbol{\beta}) \boldsymbol{\phi}_k(y_i) \right]. \tag{4}$$

The log likelihood is now supplemented by adding a quadratic penalty term to the likelihood which yields the penalized log likelihood

$$l_p(\boldsymbol{\beta}, \lambda) = l(\boldsymbol{\beta}) - \frac{1}{2}\lambda \boldsymbol{\beta}^T D_m \boldsymbol{\beta} \tag{5}$$

where the penalty matrix D_m induces smoothness and λ is the penalty parameter. With respect to the choice of D_m we follow the idea of penalized splines (see Eilers and Marx 1996) and we want the variation of weights c_k to be penalized. This holds if β_k

does not differ abruptly from β_{k-1} or β_{k+1}, respectively. We therefore penalize m-th order differences. Let $\tilde{L}_m$ denote the $(m+1)$-th order difference matrix, where e.g. $\tilde{L}_1$ is

$$\tilde{L_1} = \begin{pmatrix} 1 & -1 & 0 & \cdots & 0 \\ 0 & 1 & -1 & \ddots & \vdots \\ \vdots & \ddots & \ddots & \ddots & 0 \\ 0 & \cdots & 0 & 1 & -1 \end{pmatrix}.$$

Note that $\tilde{L}_m$ is $(\tilde{K} - m) \times \tilde{K}$ dimensional with $\tilde{K} = 2K + 1$. Since $\beta_0 \equiv 0$ by definition, we can omit the linear combination with β_0. Let therefore $L_m = \tilde{L}_m[, \{-K, \ldots, -1, 1, \ldots, K\}]$ denote the matrix by omitting the redundant middle column in L_m corresponding to β_0, where the notation $[, A]$ refers to extracting the columns given by the index set A. The penalty D_m now results as $L_m^T L_m$.

Finally we sketch how to maximize (5) with respect to β using a Newton-Raphson approach. Denote with $\mathcal{C}(\boldsymbol{\beta})$ the $(2K+1) \times (2K)$ matrix with elements

$$\frac{\partial c_k(\boldsymbol{\beta})}{\partial \beta_j}, \quad k = -K, ..., K, \quad j = -K, \ldots, -1, 1, \ldots, K$$

which results as

$$\mathcal{C}(\boldsymbol{\beta}) = \left(\text{diag}(\tilde{\mathbf{c}}) - \tilde{\mathbf{c}}\tilde{\mathbf{c}}^T\right)[, \{-K, ..., -1, 1, ..., K\}] ,$$

where $\tilde{\mathbf{c}} = (c_{-K}(\boldsymbol{\beta}), \ldots, c_0(\boldsymbol{\beta}), \ldots, c_K(\boldsymbol{\beta}))^T$. The derivative of (5) with respect to $\boldsymbol{\beta}$ now equals

$$s_p(\boldsymbol{\beta}; \lambda) = \frac{\partial l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}} - \lambda D_m \boldsymbol{\beta} = \sum_{i=1}^{n} \frac{\mathcal{C}^T(\boldsymbol{\beta})\tilde{\boldsymbol{\phi}}_i}{f(y_i)} - \lambda D_m \boldsymbol{\beta} \tag{6}$$

with $\tilde{\boldsymbol{\phi}}_i = (\phi_{-K}(y_i), \ldots, \phi_0(y_i), \ldots, \phi_K(y_i))^T$ and $f(y)$ as defined in (2). The negative second order derivative of (5) with respect to $\boldsymbol{\beta}$ may be approximated by

$$J_p(\boldsymbol{\beta}; \lambda) = -\frac{\partial^2 l(\boldsymbol{\beta})}{\partial \boldsymbol{\beta}\, \partial \boldsymbol{\beta}} + \lambda D_m \approx \sum_{i=1}^{n} \frac{\mathcal{C}^T(\boldsymbol{\beta})\tilde{\boldsymbol{\phi}}_i \tilde{\boldsymbol{\phi}}_i^T \mathcal{C}(\boldsymbol{\beta})}{f(y_i)^2} + \lambda D_m. \tag{7}$$

Newton-scoring is done for estimating $\boldsymbol{\beta}$, using a fixed λ.

2.2 Selecting the penalty parameter

The penalty parameter λ steers the amount of smoothness of the fitted density and it needs to be selected data driven. A straight forward approach is the Akaike Information Criterion (AIC) (see Akaike 1974) selecting λ by minimizing

$$AIC(\lambda) = -l(\hat{\boldsymbol{\beta}}) + df(\lambda) \tag{8}$$

where

$$df(\lambda) = \operatorname{tr}\left(J_p^{-1}(\hat{\boldsymbol{\beta}}; \lambda)\ J_p(\hat{\boldsymbol{\beta}}; \lambda = 0)\right) \tag{9}$$

approximate the degree of the fit. Note that $df(\lambda = 0) = 2K$ is giving the number of parameters. Alternatively one may apply Generalized Cross Validation (GCV). Apparently, selecting λ by minimizing (8) requires a grid search and fitting the density for a set of λ values, which is usually quite time consuming. Alternatively, in penalized spline smoothing it has been shown useful to make use of the link to mixed models [see Wand (2003), Kauermann (2005) or recent work by Reiss and Ogden (2009) and Wood (2011)]. To do so, we adopt a Bayesian viewpoint and comprehend the penalty as *a priori* distribution in the sense that the coefficient vector is assumed to be random with

$$\boldsymbol{\beta} \sim N(0, \lambda^{-1} D_m^-) \tag{10}$$

where D_m^- denotes the generalized inverse of D_m. The prior (10) is degenerated, which needs to be corrected as follows. We decompose $\boldsymbol{\beta}$ into the two components $\boldsymbol{\beta}^\sim$ and $\boldsymbol{\beta}^\perp$, respectively, such that $\boldsymbol{\beta}^\sim$ is a normally distributed random vector with non degenerated variance and $\boldsymbol{\beta}^\perp$ are the remaining components treated as parameters, see also Wand and Ormerod (2008). In fact based on a singular value decomposition we have

$$D_m = U^\sim \Lambda^\sim U^{\sim T}$$

with $\Lambda^\sim$ as diagonal matrix with positive eigenvalues and $U^\sim \in \mathbb{R}^{p \times h}$ with corresponding eigenvectors where $p = 2K$ is the number of elements in $\boldsymbol{\beta}$ and $h = p - m$ is the rank of D_m with m as degree of the difference matrix $\tilde{L}_m$. Extending $U^\sim$ to an orthogonal basis by $U^\perp$ gives $\boldsymbol{\beta}^\sim = U^{\sim T}\boldsymbol{\beta}$ with the a priori assumption $\boldsymbol{\beta}^\sim \sim N(0, \lambda^{-1}\Lambda^{\sim -1})$ and with $U = (U^\sim, U^\perp)$ as orthogonal basis, we get $\boldsymbol{\beta}^\perp = U^{\perp T}\boldsymbol{\beta}$. Conditioning on $\boldsymbol{\beta}^\sim$, we have y being distributed according to (2) and with (10) we get the mixed model log likelihood

$$l_m(\lambda, \boldsymbol{\beta}^\perp) = \log \int |\lambda\Lambda^\sim|^{\frac{1}{2}} \exp\left\{l_p(\boldsymbol{\beta}, \lambda)\right\} \mathrm{d}\boldsymbol{\beta}^\sim. \tag{11}$$

The integral can be approximated by a Laplace approximation (see also Rue et al. 2009)

$$l_m(\lambda, \hat{\boldsymbol{\beta}}^\perp) \approx \frac{1}{2}\log|\lambda\Lambda^\sim| + l_p(\hat{\boldsymbol{\beta}}, \lambda) - \frac{1}{2}\log|U^{\sim T} J_p(\hat{\boldsymbol{\beta}}; \lambda) U^\sim|. \tag{12}$$

where $\hat{\boldsymbol{\beta}}$ denotes the penalized maximum likelihood estimate. We can now differentiate (12) with respect to λ which gives

$$
\begin{aligned}
\frac{\partial l_m(\lambda, \hat{\beta}^{\perp})}{\partial \lambda} = &-\frac{1}{2}\hat{\boldsymbol{\beta}}^T D_m \hat{\boldsymbol{\beta}} \\
&+\frac{1}{2\lambda}\mathrm{tr}\left\{(U^{\sim T} J_p(\hat{\boldsymbol{\beta}}; \lambda)U^{\sim} + \lambda\Lambda^{\sim})^{-1} U^{\sim T} J_p(\hat{\boldsymbol{\beta}}; \lambda = 0)U^{\sim}\right\}
\end{aligned}
\tag{13}
$$

For practical implementation we approximate the trace component in (13) by $df(\lambda) - (m - 1)$ with $df(\lambda)$ as in (9). In fact with this simplification, we can construct an estimating equation from (13) via

$$
\hat{\lambda}^{-1} = \frac{\hat{\boldsymbol{\beta}}^T D_m \hat{\boldsymbol{\beta}}}{df(\hat{\lambda}) - (m - 1)}. \tag{14}
$$

Apparently, both sides of Eq. (14) depend on λ. An iterative solution is possible by fixing λ on the right hand side in (14), update λ on the left hand side and iterate this step by updating the right hand side of (14). This estimation scheme has been suggested in generalized linear mixed models by Schall (1991), see also Searle et al. (1992). For penalized spline smoothing Wood (2011) shows that the selection of smoothing parameter λ based in the mixed model approach behaves superior compared to AIC selected values, see also Reiss and Ogden (2009).

We can also use the marginal likelihood (12) to check or select the number of knots used in the basis. In fact the maximized $l_m(\lambda, \hat{\boldsymbol{\beta}}^{\perp})$ depends on K which may be denoted as $l_m(\lambda, \hat{\boldsymbol{\beta}}^{\perp}; K)$. Considering K itself as a parameter we can maximize the marginal likelihood. In simulations we well see later that the actual choice of K has little influence on the performance which exactly mirror Ruppert (2002)'s findings in standard smooth regression models.

We show further theoretical properties, (i) that the estimated density has minimal Kullback-Leibler distance to the unknown true density and (ii) the asymptotic normality of the estimated coefficients $\boldsymbol{\beta}$ in the Appendix, Section B. Moreover, we present results about bias and variance of the estimation in the Appendix, Section B.

2.3 Practical settings, numerical implementation and extensions

The fitting requires a number of practical settings which are implemented in the R package `pendensity` (see Schellhase 2010). First, we need to allocate the basis density given a set of observations $y_1, \ldots, y_n$. We suggest to use B-splines allocated at equidistant knots μ_k with the most left knot μ_L, fulfilling $\mu_L \leq \min(y_i)$ and the most right knot $\mu_R \geq \max(y_i)$. The performance of the estimations can be improved using additional equidistant knots beyond $[\mu_L, \mu_R]$. Therefore, the used penalization of neighbouring weights c_k in interaction with additional knots can achieve a better fit of the densities at the boundaries. In our simulations (see Sect. 3) we run estimations with one additional knot placed with the same distance used for the knots in the support at each end of $[\mu_L, \mu_R]$ and observe an improved result for several distributions.

As starting value we found that assuming a uniform distribution is useful, i.e. we set $\beta_k = 0$ to start the Newton procedure. We also experimented with different starting values but observed that the uniform start is preferable in terms of iteration steps to reach the maximum of the penalized likelihood. To avoid terminating the algorithm in a local instead of global maximum, it is advisable to fit the density for a number of different starting values and take the fit with the maximum value of the likelihood. It should be noted, however, that the problem of local maxima occurs if the penalty is not strong enough, since the penalty in (5) works towards the concavity of the penalized likelihood. It is therefore recommendable to start the Newton procedure with a large λ. Finally, the number of knots, i.e. the dimension of the density basis needs to be selected. Generally, we suggest to use a large K, where we have decided upon the default setting $K = 20$, which corresponds to a 41 dimensional basis. This mirrors the rule of thumb suggested in Ruppert (2002). Increasing $K \gg 20$ does not lead to an improved performance of the fit. But K should not be selected too small, due to the appearance of an approximation bias of not ignorable size (see Kauermann et al. 2009). We show the influence of K on the fit in the next section and we confirm the impression of Ruppert (2002) in that the actual choice of K has little influence on the fit.

Conceptually, the approach is easily extended to multivariate density estimation. In this case we replace basis densities $\phi_k(\cdot)$ in (2) by Tensor products of univariate fixed basis densities. The index k is then running over a grid and the penalty should be formulated in each direction of the grid, that is row- and columnwise for two dimensions.

3 Simulations and examples

3.1 Simulations

Univariate Density Estimation

To demonstrate the performance of the penalized density estimate we run a number of simulations. We use (i) a normal distribution $F_0(y) \sim N(0, 1)$, a mixture of normals (ii) $F_0(y) \sim \frac{1}{2}N(-\frac{1}{2}, \frac{1}{4}) + \frac{1}{2}N(\frac{1}{2}, \frac{1}{4})$, two bimodal mixtures (iii) as $F_0(y) \sim \frac{1}{2}N(-\frac{3}{2}, 1) + \frac{1}{2}N(\frac{3}{2}, 1)$ and (iv) with $F_0(y) = \frac{3}{4}N(-\frac{3}{2}, 1) + \frac{1}{4}N(\frac{3}{2}, 1)$, mixture of five normal densities (v) as $F_0(y) \sim \frac{13}{20}N(-1, \frac{1}{2}) + \frac{2}{20}N(-\frac{1}{2}, \frac{1}{2}) + \frac{1}{20}N(0, 1) + \frac{3}{20}N(\frac{1}{2}, \frac{1}{2}) + \frac{1}{20}N(1, \frac{1}{2})$, a normal variance mixture as (vi) with $F_0(y) \sim \frac{1}{2}N(0, 1) + \frac{1}{2}N(0, 10)$, (vii) a gamma distribution $\Gamma(3, 1)$ and (viii) a beta distribution $\text{Beta}(10, 10)$. To compare our results labelled with $\hat{f}_K(\cdot)$ with alternative routines we use, (a) classical kernel density estimates (see Wand and Jones 1995), (b) the density estimation proposal of Gu and Wang (2003), (c) the approach of density estimation of Eilers and Marx (1996), (d) a mixture density approach, (e) the log-spline routine and (f) a wavelet approach, respectively. For the traditional kernel density estimate (a) labelled as $\hat{f}_{kernel}(\cdot)$, we utilize two approaches for selecting the bandwidth. First we use cross validation (bw=ucv) and secondly we choose the bandwidth by the approach of Sheather and Jones (1991) (bw=SJ). Both kernel routines are implemented in the `density()` routine in `R`. For (b) one estimates the unknown density $f(\cdot)$ by the logistic density transform (1) with a roughness penalty imposed on $\eta(y)$ which penalizes integrated squared order derivatives. This routine is implemented in `R` in the gss package (see Gu 2009) and we label the resulting estimated density with $\hat{f}_{spline}(\cdot)$. For the third approach (c) we divide the support of the data points in a large number of bins. Following Ruppert et al. (2003) we use $B = 200$ equidistant subintervals (bins) and notate with b_j the number of observations in the j-th bin, $j = 1, \dots, 200$. With m_j as bin center and d_j as bin width we fit the Poisson model $b_j \sim \text{Poisson}(f(m_j)nd_j)$. One can now fit the density function $f(\cdot)$ using for instance the gam() procedure in R, see Wood (2006). For the fourth approach (d) we make use of the R package mixtools (see Young et al. 2009) and select the number of mixture components using a Bayesian Information Criterion (BIC) and the entropy criterion suggested in Celeux and Soromenho (1996). We thereby increased K successively starting from $K = 1$ until the criterion reaches its optimum. The fifth approach, the log-spline density estimation (e) is implemented in R package logspline (see Kooperberg 2009). Finally, the wavelet density estimation (f) is done with R package wavethresh (see Nason 2010), with finest resolution level equal to one and Daubechies least asymmetric wavelets. For comparison with our penalized density estimate (g) we use $2K + 1$ bins with $K = 20$ and $K = 30$, respectively and label the resulting density estimate with $\hat{f}_{bin,K}(\cdot)$. We also select K data driven to maximize the likelihood derived in Sect. 2.2.

To evaluate the performance of the fit we run $N = 500$ replicates of the simulation for different sample sizes n and different K and calculate the integrated Mean Squared Error. Therefore we first calculate the Mean Squared Error

$$\text{MSE}(\hat{f}(\tilde{y}_k)) = \frac{1}{N} \sum_{j=1}^{N} \left\{ \hat{f}_{(j)}(\tilde{y}_k) - f_0(\tilde{y}_k) \right\}^2,$$

where the calculated estimated densities $\hat{f}_{(j)}$, $j = 1, \dots, N$ and the true densities f_0 are evaluated at fixed and equidistant values $\tilde{y}_k$, $k = 1, \dots, 1000$, say. The IMSE results as follows

$$\widehat{\text{IMSE}}(\hat{f}(\tilde{y})) = \frac{1}{1000} \sum_{k=1}^{1000} \left\{ \text{MSE}(\hat{f}(\tilde{y}_k)) \right\}.$$

Accordingly the results of the competing density estimations $\hat{f}_K(\cdot)$, $\hat{f}_{kernel}(\cdot)$, $\hat{f}_{spline}(\cdot)$, $\hat{f}_{bin,K}(\cdot)$, $\hat{f}_{mixture}(\cdot)$, $\hat{f}_{log}(\cdot)$ and $\hat{f}_{wave}(\cdot)$ are shown in Table 1. Note that for simulation scenario (i) we used for the mixture (d) the true one component normal distribution with fitted parameters which maybe considered as artificial benchmark in this case. In general it appears that the approach with a penalized mixture performs promisingly well in comparison with the four competitors, even though no method is uniformly superior. In general, however, in scenarios where the penalized mixture approach is not optimal its optimal IMSE is not more than 62% larger than the IMSE of the best density estimate, while this number is larger for all other competitors. For small n but even more for large n we observe the well established fact that the quality of the fit remains the same and K does not influence the performance of the fit. We notice an improved performance in some examples, if one adds one additional

Table 1 Relative integrated mean squared error

Density approach rel. IMSE		(g) Penalized mixture			(a) Kernel		(b) Spline	(c) Binning		(d) Mixture		(e) Log-spline	(f) Wavelet	Best absolute IMSE
		$\hat{f}_K\ (y)$ $K = 20$	$\hat{f}_K\ (y)$ $K = 30$	$\hat{f}_K\ (y)$ K_{opt}	$\hat{f}_{kernel}\ (y)$ bw=ucv	$\hat{f}_{kernel}\ (y)$ bw=SJ	$\hat{f}_{spline}\ (y)$ Gu	$\hat{f}_{bin,K}\ (y)$ EM $K = 20$	$\hat{f}_{bin,K}\ (y)$ EM $K = 30$	$\hat{f}_{mixture}\ (y)$ BIC	$\hat{f}_{mixture}\ (y)$ entropy	$\hat{f}_{log}\ (y)$ Koo	$\hat{f}_{wave}\ (y)$ Nason	
(i)	$n = 100$	**1.000**	1.040	1.149	2.485	2.056	1.472	2.124	2.247	3.374	4.515	4.677	6.449	0.396
	$n = 400$	**1.000**	1.042	1.069	3.986	3.208	1.694	2.444	2.514	3.583	3.264	6.181	6.722	0.072
(ii)	$n = 100$	1.186	1.002	**1.000**	1.637	1.292	1.128	1.599	1.615	2.527	2.445	4.027	4.561	1.074
	$n = 400$	1.429	1.397	1.492	1.532	1.169	1.032	1.238	1.241	1.399	1.354	2.799	**1.000**	0.378
(iii)	$n = 100$	**1.000**	1.176	1.136	1.434	1.156	1.358	1.601	1.624	1.526	3.220	2.358	4.965	0.346
	$n = 400$	1.097	1.009	1.035	1.478	1.212	**1.000**	1.124	1.124	**1.000**	1.327	2.159	3.381	0.113
(iv)	$n = 100$	1.052	1.085	1.076	1.291	**1.000**	1.061	1.231	1.247	1.238	3.007	2.110	3.939	0.446
	$n = 400$	1.061	1.111	1.091	1.889	1.495	1.131	1.323	1.333	**1.000**	1.808	2.475	3.818	0.099
(v)	$n = 100$	**1.000**	1.088	1.138	1.433	1.148	1.090	1.227	1.253	1.387	2.440	2.353	4.339	0.980
	$n = 400$	1.025	1.062	1.062	1.864	1.574	1.124	1.326	1.322	**1.000**	1.483	2.657	2.541	0.242
(vi)	$n = 100$	1.145	1.079	1.150	1.053	1.020	**1.000**	1.007	1.015	1.170	1.192	1.167	1.987	0.454
	$n = 400$	**1.000**	1.017	**1.000**	1.030	1.006	**1.000**	1.003	1.003	1.058	1.030	1.141	1.288	0.361
(vii)	$n = 100$	**1.000**	1.371	1.142	1.364	1.056	1.023	1.381	1.427	2.238	2.907	2.358	3.871	0.302
	$n = 400$	1.664	1.626	1.785	1.224	**1.000**	1.159	1.196	1.196	2.841	2.841	1.748	2.542	0.107
(viii)	$n = 100$	1.097	1.039	1.142	1.685	1.360	**1.000**	1.354	1.446	2.632	2.525	3.444	10.612	44.197
	$n = 400$	1.199	1.123	**1.000**	2.676	2.073	1.344	2.197	2.243	2.575	2.255	4.631	51.972	9.012

Optimal performance is set equal to one and in bold. The best absolute IMSE is times 10^3

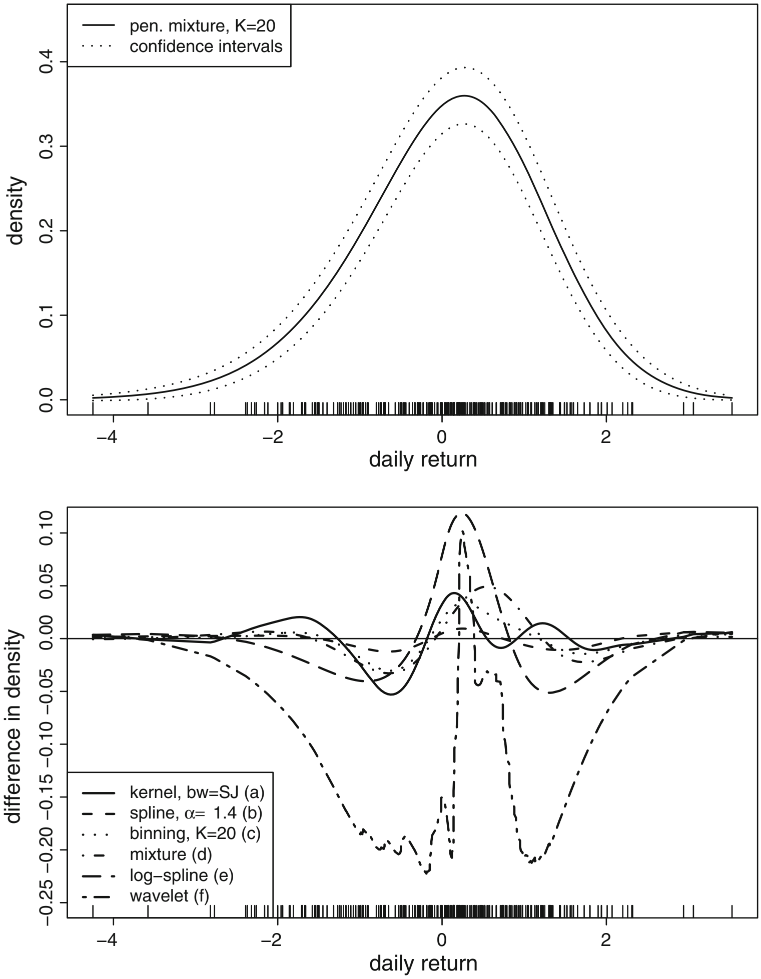

Fig. 1 *Top* penalized mixture density $\hat{f}$ of the return of Deutsche Bank AG in 2006. *Bottom* difference in density estimates of penalized mixture to alternative density estimation routines, **a** kernel density estimation, **b** spline estimation, **c** binning estimation, **d** mixtures, **e** log-spline estimation and **f** wavelet estimation

knot at each end outside of the support. In Table 1, the results of the penalized mixture approach are done with one additional knot at each end. Overall, the density estimation with a penalized mixture appears as reasonable competitor for density estimation.

3.2 Example: daily returns

We give a short example which will be picked up again in the next section. We look at the return of the two Germans stocks *Deutsche Bank AG* and *Allianz AG* in 2006. The corresponding density estimates of the penalized mixture approach are given in

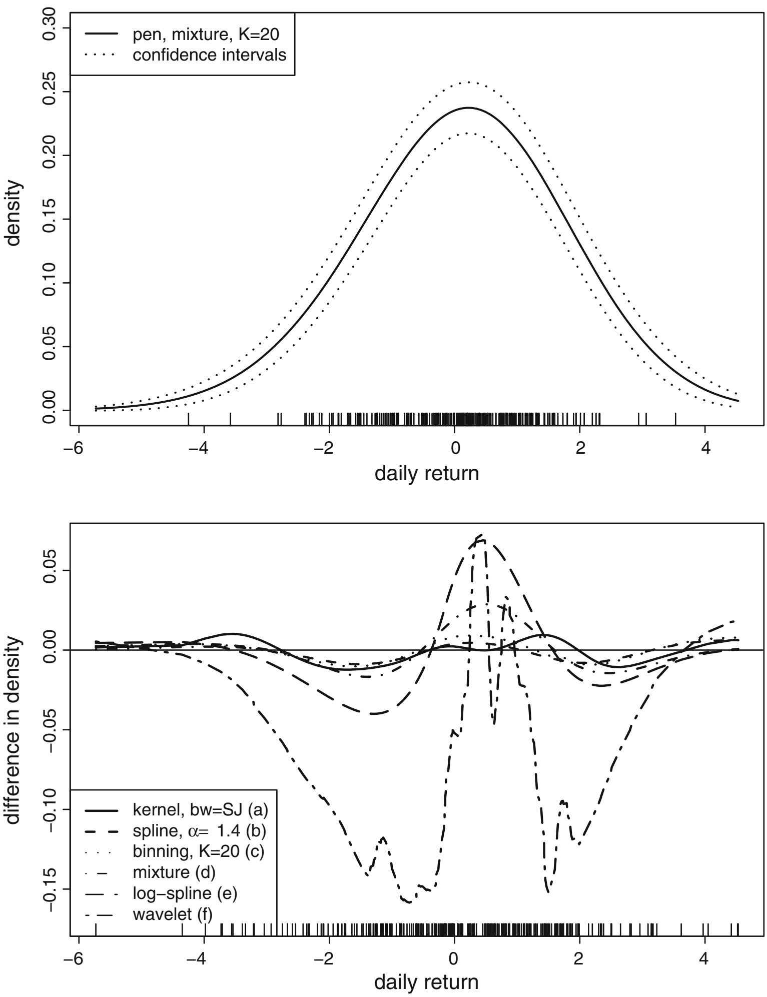

Fig. 2 *Top* penalized mixture density $\hat{f}$ of the return of the Allianz AG in 2006. *Bottom*: difference in density estimates of penalized mixture to alternative density estimation routines, **a** kernel density estimation, **b** spline estimation, **c** binning estimation, **d** mixtures, **e** log-spline estimation and **f** wavelet estimation

Figs. 1 and 2. We show the penalized mixture estimate and the difference in the density estimates to competitors (a) kernel density estimate, (b) spline based approach, (c) the binning based approach, (d) the finite mixture estimation, (e) the log-spline approach and (f) the wavelet estimate. Apparently, the kernel density estimate, the Eilers & Marx estimate and as well as the mixture estimation show for the Deutsche Bank data some peak structure in the center and additional structure for values around −1, while the result of the spline approach is nearly similar to the penalized mixture estimation. Again for the Allianz data, the kernel density estimate and the mixture estimate show some peak structure in the center and additional structure for values around 2 and −2, while the result of the spline approach is nearly similar to the penalized mixture estimation. Clearly, in both scenarios, the true function is unknown, but in the simulations the penalized density estimate performs comparable to the spline approach so that the structure shown by the other five estimates might be spurious.

4 Nonparametric comparison of densities

4.1 Covariate dependent density

We can extend the above density estimation by allowing the density to depend on some covariates x, say. We intend to estimate the conditional density $f(y|x)$. Let $y_i|x_i$ denote a random sample (with x_i either random or fixed) and $x_i = (x_{i1}, \ldots, x_{is})$ is a vector of covariates. We now assume that the weights c_k depend on x which is modelled as

$$c_k(x, \boldsymbol{\beta}) = \frac{\exp(Z(x)\boldsymbol{\beta}_k)}{\sum_{j=-K}^{K} \exp(Z(x)\boldsymbol{\beta}_j)} \tag{15}$$

where $Z(x)$ is a design matrix, e.g. $Z(x_i) = (1, x_{i1}, \ldots, x_{is})$. Let $\boldsymbol{\beta} = (\beta_{-K}^T, \ldots, \beta_{-1}^T, \beta_1^T, \ldots, \beta_K^T)^T$ be the parameter vector and $\beta_0 \equiv 0$ for identifiability reasons. The approach can be compared to finite mixture models with mixture weights depending on covariates, see e.g. Bishop (2006), Chapter 14.5 or Müller et al. (2009). In contrast to the finite mixture, however, we again assume that K is large and will impose penalties on the weights. Let p be the dimension of $Z(x)$, i.e. the number of columns. In principle, we could have a different design for the different knots, but it is convenient and practical to assume that $Z(x)$ does not depend on k and let $\mathcal{Z}(x) = I_{2K} \otimes Z(x)$, where I_{2K} is the $2K$-dimensional unit matrix and $\otimes$ denotes the tensor product. The log likelihood then becomes

$$\boldsymbol{l}(\boldsymbol{\beta}) = \sum_{i=1}^{n} \left[\log\{ \sum_{k=-K}^{K} c_k(x_i, \boldsymbol{\beta}) \boldsymbol{\phi}_k(y_i) \} \right] \tag{16}$$

with $c_k(x, \beta)$ as in (15). Similar to (5) we add a quadratic penalty term to (16) so that the penalized likelihood results as follows. Looking for instance at first order differences, i.e. $m = 1$, we have $\alpha_k(x) - \alpha_{k-1}(x) = Z(x)(\beta_k - \beta_{k-1})$, $k = -K+1, \ldots, K$. Utilizing matrix notation we can write the m-th order difference as $\Delta_m \beta := (1_{\tilde{K}-m} \otimes Z(x))(\tilde{L}_m \otimes I_p)\beta$ with I_p as p dimensional identity matrix. This yields the penalty as squared m-th order difference through $\boldsymbol{\beta}^T \Delta_m^T \Delta_m \boldsymbol{\beta}$. Note that the penalty depends on the particular values of the covariates x. Taking the average over the observed values we obtain the final penalty $\boldsymbol{\beta}^T \boldsymbol{D}_m \boldsymbol{\beta}$ where

$$\boldsymbol{D}_m = (L_m^T \otimes I_p^T) \left(I_{\tilde{K}-m} \otimes \frac{Z^T Z}{n} \right) (L_m \otimes I_p)$$

with $Z = (Z^T(x_1), \ldots, Z^T(x_n))^T \in \mathbb{R}^{n \times p}$. The penalized likelihood results now as $\boldsymbol{l}_p(\boldsymbol{\beta}, \lambda) = \boldsymbol{l}(\boldsymbol{\beta}) - \frac{1}{2}\lambda \boldsymbol{\beta}^T \boldsymbol{D}_m \boldsymbol{\beta}$. Based on (6) the penalized first derivative equals $\boldsymbol{s}_p(\boldsymbol{\beta}; \lambda) = \partial l(\boldsymbol{\beta})/\partial \boldsymbol{\beta} = \sum_{i=1}^{n} \boldsymbol{s}_i(\boldsymbol{\beta}; \lambda)$ where

$$\boldsymbol{s}_i(\boldsymbol{\beta};\lambda) = \mathcal{Z}^T(x_i)\mathcal{C}^T(x_i,\boldsymbol{\beta})\frac{\tilde{\boldsymbol{\phi}}_i}{\hat{f}(y_i|x_i)} - \lambda \boldsymbol{D}_m\boldsymbol{\beta}$$

with obvious definition for $\mathcal{C}(x_i,\beta)$. Analogously to (7) we approximate the negative penalized second order derivative through

$$\boldsymbol{J}_p(\boldsymbol{\beta};\lambda) = -\frac{\partial^2 l_p(\boldsymbol{\beta},\lambda)}{\partial\boldsymbol{\beta}\,\partial\boldsymbol{\beta}^T} \approx \sum_{i=1}^{n}\boldsymbol{s}_i(\boldsymbol{\beta};\lambda)\boldsymbol{s}_i{}^T(\boldsymbol{\beta};\lambda) + \lambda\boldsymbol{D}_m.$$

Estimation can now be carried out in the same way as done in the previous sections. This also applies to the estimation of the penalty parameter λ. Assuming the prior distribution (10) allows with the same arguments used in Sect. (2.2) to calculate the penalty parameter from the mixed model resulting as

$$\hat{\lambda}^{-1} = \frac{\hat{\boldsymbol{\beta}}^T\boldsymbol{D}_m\hat{\boldsymbol{\beta}}}{df(\hat{\lambda}) - p(m-1)}.$$

Moreover, all other results concerning the asymptotic distribution of the estimate extend from the previous section so that we do not explicitly list them here for the sake of space.

4.2 Testing densities on equality

We can employ the idea above now to test the hypotheses on equality of densities. We formulate this by testing

$$H_0 : f(y|x_{(1)}) = f(y|x_{(0)}),\ y \in \mathbb{R} \tag{17}$$

for two specific values of $x_{(1)} = (x_{(1)1},\ldots,x_{(1)s})$ and $x_{(0)} = (x_{(0)1},\ldots,x_{(0)s})$. For instance, if $s = 1$ and $x_{i1} \in \{0,1\}$ indicates two groups, we may test with (17) whether the distribution of y_i is the same in the two groups instead of comparing densities. We look at differences in the distribution functions and define the test statistics

$$T_{max} = \max\{|T(\tau_k)|\,, k = -K,\ldots,K\}$$

with

$$T(y) = \hat{F}(y|x_1) - \hat{F}(y|x_0) = \sum_{k=-K}^{K}(c_k(x_1,\hat{\boldsymbol{\beta}}) - c_k(x_0,\hat{\boldsymbol{\beta}}))\Phi_k(y),$$

and $\tau_{-K},\ldots,\tau_0,\ldots,\tau_K$ are denoting the knots of the basis functions and $\Phi_k(y)$ are distribution functions to basis densities $\phi_k(y)$. Under H_0 we have $E\{T(y)\} =$

0 for all y and based on the asymptotic arguments used before we can show that $\tilde{\boldsymbol{T}} = (T(\tau_{-K}), \ldots, T(\tau_0), \ldots, T(\tau_K))^T$ follows the asymptotic distribution

$$\tilde{\boldsymbol{T}} \overset{a}{\sim} N(0, \boldsymbol{W}) \tag{18}$$

with

$$\boldsymbol{W} = \tilde{\Phi}[\mathcal{C}_1 - \mathcal{C}_0] V(\beta^{(0)}, \lambda)[\mathcal{C}_1 - \mathcal{C}_0]^T \tilde{\Phi}^T$$

where $\mathcal{C}_j = \mathcal{C}(x_j, \hat{\boldsymbol{\beta}})\mathcal{Z}(x_j)$ for $j = 0, 1$ and $\tilde{\Phi} \in \mathbb{R}^{(2K+1)\times(2K+1)}$ as matrix with entries $\Phi_k(\tau_l)$ where (row) index k and (column) index l with $l, k = -K, \ldots, K$. Finally matrix $V(\boldsymbol{\beta}^{(0)}, \lambda)$ is the variance matrix (23) extended to the case of covariate dependent densities. Note that matrix $\boldsymbol{W}$ is easily calculated which allows to simulate the distribution of T_{max} in a straight forward way by sampling $\tilde{T}$ from (18). This can be done relatively fast after some spectral decomposition of $\boldsymbol{W}$ so that any approximate calculation of the distribution of T_{max} is numerically easy.

5 Simulation and example

5.1 Simulation

We run a small simulation to check the performance of the fit, particularly of the testing idea based on T_{max}. To do so we simulate $n = 100$ and $n = 400$ data points from the following distributions. We assume a univariate covariate (group indicator) with $x_i = 0$ for $n/2$ and $x_i = 1$ for the remaining $n/2$ observations. We simulate y given x from the following scenarios. First, (i) we draw y from a standard normal for both $x = 0$ and $x = 1$, i.e. $y|x \sim N(0, 1)$, (ii) we draw $y|x = 0 \sim N(0, 1)$ and $y|x = 1 \sim N(\frac{1}{5}, 1)$ that is we shift the mean by $\frac{1}{5}$ for $x = 1$, and finally (iii) $y|x = 0 \sim N(0, 1)$ and $y|x = 1 \sim \frac{1}{2}N(-\frac{1}{2}, \frac{1}{4}) + \frac{1}{2}N(\frac{1}{2}, \frac{1}{4})$. For all three scenarios we calculate for each simulation the p-value resulting for T_{max}. We repeat the simulation 1000 times and give in Table 2 the number of p-values smaller than a nominal level α. Bear in mind that for scenario (i) the null hypothesis is true so that the p-value should be uniformly distributed on [0, 1]. As reference we also calculate both, the p-value resulting for a Kolmogorov-Smirnov test based on comparing the sample for $x = 0$ against $x = 1$ as well as the p-value resulting from the linear model $y = \beta_0 + x\beta_x$ and a t-test on $H_0 : \beta_x = 0$. As can be seen from the simulated numbers the test on the equalities of densities works convincingly well which supports the idea of density estimation with a penalized mixture.

5.2 Example

As example we look again at the daily returns for the two stocks considered in Sect. 3.2. We look at data from 2006 and 2007, and our focus of interest is to test the hypothesis that the distribution of the returns is the same in the 2 years. The corresponding plot is shown in Fig. 3. Applying the test based on T_{max} to this example yields the p-values

Table 2 Proportion of p-values smaller than α, based on 1,000 simulations

Level	Simulation	Test on T_{max}	Kolmogorov-Smirnov test	Test on $\beta_x = 0$ in linear model
$\alpha = 0.01$	(i) $n = 100$	0.010	0.011	0.009
	$n = 400$	0.009	0.011	0.007
	(ii) $n = 100$	**0.063**	0.042	0.057
	$n = 400$	**0.288**	0.182	**0.288**
	(iii) $n = 100$	**0.031**	0.005	0.003
	$n = 400$	**0.377**	0.080	0.003
$\alpha = 0.05$	(i) $n = 100$	0.058	0.041	0.058
	$n = 400$	0.052	0.049	0.053
	(ii) $n = 100$	**0.163**	0.116	0.155
	$n = 400$	0.504	0.397	**0.526**
	(iii) $n = 100$	**0.134**	0.051	0.030
	$n = 400$	**0.735**	0.313	0.036

Optimal performance is set in *bold*

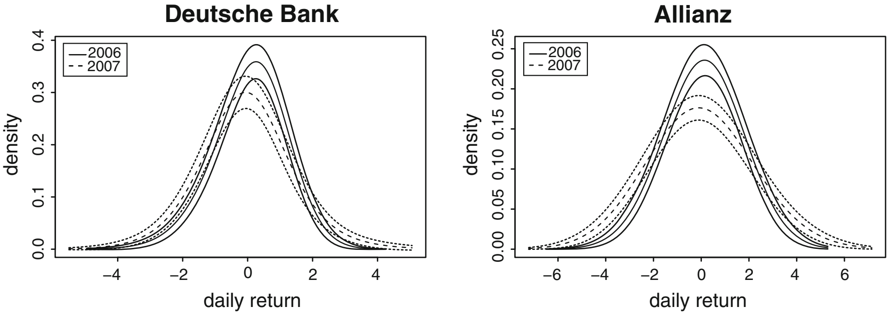

Fig. 3 Density of the return of Deutsche Bank AG and Allianz AG in 2006 and 2007

of 0.048 for *Deutsche Bank AG* and 0.019 for *Allianz AG*. Hence, there is indication that the returns in the 2 years differ in distribution.

6 Conclusion

In this paper we tackled the classical problem of density estimation. Our approach picked up the idea of Komárek and Lesaffre (2008) and extended this to regular as well as covariate dependent density estimation. We examined density estimation scheme based on penalized B-spline bases using the direct link from penalized smoothing splines to mixed models. Simulations showed promising results when comparing our density estimation to competitors. First, in simple density estimation it appears that the penalized mixture approach proposed here behaves better or at least similarly compared to the common alternatives (a) kernel density estimation, (b) spline based density estimation (c) binning based estimation, (d) mixture densities, (e) log-spline density estimation and (f) wavelet density estimation. Moreover, our density estimation approach performed almost as the best, regarding the MSE, while the classical approach (c) binning did not operate optimally in any considered density case. Secondly, extending the procedure towards covariate dependent density estimation allows for testing on the equality of densities in different groups. The approach showed superior behaviour in our simulations when compared to the classical Kolmogorov-Smirnov test. This test on equality of densities in different groups carries some omnibus power, which is seen especially in cases, where the standard tests do not announce inequality of the groups (see Table 2).

The approach is in principle easy to extend to multivariate density estimation. In the multivariate case, though, the numerical requirements of the penalized mixture approach do however exponentially increase due to the increasing number of B-spline basis functions. Because of this curse of dimensionality multivariate density estimation remains a difficult task.

Acknowledgments G. Kauermann & C. Schellhase acknowledge financial support provided by the Deutsche Forschungsgesellschaft (DFG Project-Nr. KA 1188/5-1 and KA 1188/5-2).

A Asymptotic behaviour of B-spline densities

Let $\phi_k(y) = b_{q,k}(y)$ be a normed B-spline basis of order q defined on the support $[\mu_k, \mu_{k+q+1}]$ such that $\int b_{q,k}(y)\mathrm{d}y = 1$. Let $f_{K,q}(y, \boldsymbol{\beta}) = \sum_k c_k(\boldsymbol{\beta}) b_{q,k}(y)$ be the mixture B-spline density and let $r_q(y) = r_q(y, \boldsymbol{\beta}) = f_0(y)/f_{K,q}(y, \boldsymbol{\beta})$ be the ratio of the true and mixture density. Let $\mu_{-K}, \ldots, \mu_0, \ldots, \mu_K, \ldots, \mu_{K+q+1}$ be the knots located equidistantly with order $\mu_k - \mu_{k-1} = O(K^{-1})$. Note that our B-spline basis is q times differentiable within each interval $[\mu_k, \mu_{k+1}]$ and in particular, boundary splines are continuous. With (20) we get

$$\int_{\mu_k}^{\mu_{k+q+1}} b_{q,k}(y) r_q(y)\mathrm{d}y = 1 \tag{19}$$

so that there exists a $\xi_k \in (\mu_k, \mu_{k+q+1})$ with $r_q(\xi_k) = 1$ for $k = -K, \ldots, K$. With the recursive formula for derivatives of B-splines (see Butterfield 1976) we get for $q \geq 2$ with partial integration and making use of (19) for $k = -K, \ldots, K-1$

$$\begin{aligned} \int_{\mu_k}^{\mu_{k+q+2}} b_{q+1,k}(y) r_q'(y)\, dy = b_{q+1,k}(y) r_q(y) \Big|_{\mu_{k+q+2}}^{\mu_k} \\ + K \Bigg\{ \int_{\mu_k}^{\mu_{k+q+1}} b_{q,k}(y) r_q(y)\, dy \\ - \int_{\mu_{k+1}}^{\mu_{k+q+1}} b_{q,k+1}(y) r_q(y)\, dy \Bigg\} = 0 \end{aligned}$$

This in turn shows with the mean value theorem that there exists a $\xi_k^{(1)} \in [\mu_k, \mu_{k+q+2}]$ with $r_q'(\xi_k^{(1)}) = 0$. Considering the derivative of $r_q(y)$ it is easily derived that $f_{K,q}'(\xi_k^{(1)}) = f_0'(\xi_k^{(1)}) + O(K^{-1})$. With the same arguments as above we can show that there exists $\xi_k^{(l)}$ with $1 \leq l \leq q-1$ and $k = -K, \ldots, K-l$ such that $f^{(l)}(\xi_k^{(l)}) = f_{K,q}^{(l)}(\xi_k^{(l)}) + O(K^{-1})$. This allows to conclude with iterative arguments that for $q \geq 1$ and for $l \leq q-1$

$$f_{K,q}^{(l)}(y) = f^{(l)}(y) + O(K^{-q+l})$$

so that for $l = 0$ we get the approximation error

$$f_{K,q}(y) = f(y) + O(K^{-q}).$$

B Properties of the estimate

Looking at theoretical properties of the estimation we focus on two questions. First, how well can the mixture density (2) approximate an unknown true density and secondly, what are the estimation properties of the penalized estimate. Let $f_K(y, \hat{\boldsymbol{\beta}})$ denote the mixture density (2) with weights $c_k(\hat{\boldsymbol{\beta}})$ defined through (3). Moreover, let $f_0(y)$ denote the true continuous unknown density. We define $\boldsymbol{\beta}^{(0)} = (\beta_{-K}^{(0)}, \ldots, \beta_K^{(0)})$ as the true parameter in the sense that $f_K(y, \boldsymbol{\beta}^{(0)})$ and $f_0(y)$ have minimal Kullback-Leibler distance based on the true density. So, we intent to minimize $E_{f_0(y)}\left\{\log\left(\frac{f_K(y,\hat{\boldsymbol{\beta}})}{f_0(y)}\right)\right\}$ with respect to $\hat{\boldsymbol{\beta}}$, which is equivalent to $0 = E_{f_0(y)}(\frac{\partial}{\partial \hat{\boldsymbol{\beta}}} \log f_K(y, \hat{\boldsymbol{\beta}}))$. This means that $\boldsymbol{\beta}^{(0)}$ is implicitly defined through

$$0 = E_{f_0(y)}\left\{\frac{\mathcal{C}(\boldsymbol{\beta}^{(0)})^T \tilde{\boldsymbol{\phi}}(y)}{f_K(y, \boldsymbol{\beta}^{(0)})}\right\} \tag{20}$$

where $\tilde{\boldsymbol{\phi}}(y) = \left(\phi_{-K}(y), \ldots, \phi_0(y), \ldots, \phi_K(y)\right)^T$. Note that $\boldsymbol{\beta}^{(0)}$ depends on K, the number of knots, which is suppressed in our notation for simplification. Let $r(y, \boldsymbol{\beta}) = f_0(y)/f_K(y, \boldsymbol{\beta})$ be the ratio of the true and approximate density and define $H_k = H_k(\boldsymbol{\beta}) = \int \phi_k(y)\, r(y, \boldsymbol{\beta})\, \mathrm{d}y$. Note that $\sum_{k=-K}^{K} c_k(\boldsymbol{\beta}^{(0)}) H_k = 1$. Based on (20) and reflecting the definition of matrix $\mathcal{C}(\boldsymbol{\beta})$ we derive $H_k = 1$ for $k = -K, \ldots, K$. This allows with the well-known mean value theorem for integration to show the existence of $\xi_k \in [\mu_k, \mu_{k+1}]$ with $f_0(\xi_k) = f_K(\xi_k, \boldsymbol{\beta}^{(0)})$ for $k = -K, \ldots, K-1$. It follows with the mean value theorem for integration $\int \phi_k(y) r(y) dy = 1 = \int \phi_k(y) dy\, r(\xi_k)$. So, there exists ξ_k, such that $r(\xi_k) = 1$. Assuming now that the knots are placed densely in the sense $\mu_k - \mu_{k+1} = O(K^{-1})$, $k = -K, \ldots, K-1$ we obtain for $\delta_k(y) = f_0(y) - f_K(y, \boldsymbol{\beta}^{(0)})$ with simple Taylor series expansion the order $\delta_k(y) = O(K^{-1})$ for $\mu_{-K} \leq y \leq \mu_K$. We will call $\delta_k(y)$ subsequently the approximation bias. Using B-splines as basis densities allows us to obtain an even smaller asymptotic order for the approximation bias. In fact, if $f_0(y)$ is q-times differentiable and

$\phi_k(y)$ is a B-spline density of degree q, we obtain for $q \geq 1$ the order $\delta(y) = O\left(K^{-q}\right)$. A proof is given in the Appendix, Section A. It is therefore practically as well as theoretically advisable to set ϕ_k as B-splines. To this end we have derived the approximation bias, so that we have answered the question how well the mixture density (2) can approximate the true unknown density $f_0(y)$. The next step is to investigate the properties of the penalized estimate of parameter $\boldsymbol{\beta}^{(0)}$. In principle this boils down to standard penalized likelihood estimation so that simple and standard expansions yield (see Kauermann et al. 2009) the necessary results. In fact we obtain

$$\hat{\boldsymbol{\beta}} - \boldsymbol{\beta}^{(0)} \approx J_p^{-1}(\boldsymbol{\beta}^{(0)}; \lambda)\, s_p(\boldsymbol{\beta}^{(0)}; \lambda)$$

which allows to formulate the asymptotic normality

$$\hat{\boldsymbol{\beta}} \overset{a}{\sim} N\big(\boldsymbol{\beta}^{(0)} + \text{bias}(\boldsymbol{\beta}^{(0)}, \lambda), V(\boldsymbol{\beta}^{(0)}, \lambda)\big) \tag{21}$$

with

$$\text{bias}(\boldsymbol{\beta}^{(0)}, \lambda) = -\lambda I_p^{-1}(\boldsymbol{\beta}^{(0)}, \lambda) D_m \boldsymbol{\beta}^{(0)} \tag{22}$$

$$V(\boldsymbol{\beta}^{(0)}, \lambda_0) = I_p^{-1}(\boldsymbol{\beta}^{(0)}, \lambda) I_p(\boldsymbol{\beta}^{(0)}, \lambda = 0) I_p^{-1}(\boldsymbol{\beta}^{(0)}, \lambda) \tag{23}$$

where $I_p(\boldsymbol{\beta}^{(0)}, \lambda) = E_{f_0(y)}\big\{J_p(\boldsymbol{\beta}^{(0)}; \lambda)\big\}$. In Sect. 2.3, we will use the above-mentioned well known link between penalized spline smoothing and mixed models. In the context of mixed models (23) is justified by Kass and Steffey (1989) and extended by Searle et al. (1992). The final step is now to transfer (21) to properties of the density estimate $f_K(y, \hat{\beta}) = \sum c_k(\hat{\boldsymbol{\beta}})\, \phi_k(y) = \tilde{\boldsymbol{\phi}}^T(y)\, \tilde{c}(\hat{\boldsymbol{\beta}})$. We get

$$f_0(y) - f_K(y, \hat{\boldsymbol{\beta}}) \overset{a}{\sim} N\big(\text{bias}\big(f_K(y, \hat{\boldsymbol{\beta}})\big), \text{Var}\big(f_K(y, \hat{\boldsymbol{\beta}})\big)\big)$$

with

$$\text{bias}\big(f_K(y, \hat{\boldsymbol{\beta}})\big) = \tilde{\boldsymbol{\phi}}^T(y)\, \mathcal{C}(\boldsymbol{\beta}^{(0)})\, \text{bias}(\boldsymbol{\beta}^{(0)}, \lambda_0)$$

$$\text{Var}\big(f_K(y, \hat{\boldsymbol{\beta}})\big) = \tilde{\boldsymbol{\phi}}^T(y)\, \mathcal{C}(\boldsymbol{\beta}^{(0)})\, V(\boldsymbol{\beta}^{(0)}, \lambda_0)\ \ \mathcal{C}^T(\boldsymbol{\beta}^{(0)}) \tilde{\boldsymbol{\phi}}^T(y)$$

Since the penalized Fisher information $I_p(\boldsymbol{\beta}^{(0)}, \lambda)$ is difficult to calculate we replace it by its observed version $J_p(\boldsymbol{\beta}^{(0)}; \lambda)$ to calculate confidence intervals. Komárek et al. (2005) argue, that there is no guarantee that the middle matrix of (23), $J_p(\boldsymbol{\beta}^{(0)}; \lambda = 0)$ is positive semidefinite. In this case one may use $J_p^{-1}(\boldsymbol{\beta}^{(0)}; \lambda)$ instead of (23) for calculating confidence intervals. The latter can also be justified following the mixed model framework discussed subsequently, as derived in Ruppert et al. (2003, p. 140).

References

Akaike H (1974) A new look at the statistical model identification. IEEE Trans Autom 19(6):716–723

Babu GJ, Canty AJ, Chaubey YP (2002) Application of bernstein polynomials for smooth estimation of a distribution and density function. J Stat Plan Infer 105(2):377–392

Bishop CM (2006) Pattern recognition and machine learning. Springer, New York, NY

Boneva LI, Kendall D, Stefanov I (1971) Spline transformations: three new diagnostic aids for the statistical data- analyst. J R Stat Soc Ser B 33(1):1–71

Butterfield K (1976) The computation of all the derivatives of a b-spline basis. IMA J Appl Math 17(1): 15–25

Celeux G, Soromenho G (1996) An entropy criterion for assessing the number of clusters in a mixture model. J Classif 13: 195–212. doi:10.1007/BF01246098

Claeskens G, Krivobokova T, Opsomer J (2009) Asymptotic properties of penalized spline estimators. Biometrika 96(3):529–544

de Boor C (1978) A practical guide to splines. Springer, Berlin

Dias R (1998) Density estimation via hybrid splines. J Stat Comput Simul 60(4):277–293

Efron B, Tibshirani R (1996) Using specially designed exponential families for density estimation. Ann Stat 24(6):2431–2461

Eilers PHC, Marx BD (1996) Flexible smoothing with B-splines and penalties. Stat Sci 11(2):89–121

Fraley C, Raftery AE (2002) Model-based clustering, discriminant analysis, and density estimation. J Am Stat Assoc 97(458):611–631

Ghidey W, Lesaffre E, Eilers PHC (2004) Smooth random effects distribution in a linear mixed model. Biometrics 60(4):945–953

Good IJ, Gaskins RA (1971) Nonparametric roughness penalties for probability densities. Biometrika 58(2):255–277

Gu C (1993) Smoothing spline density estimation: A dimensionless automatic algorithm. J Am Stat Assoc 88(422):495–504

Gu C (2009) gss: general smoothing splines. R package version 1.0-5

Gu C, Wang J (2003) Penalized likelihood density estimation: direct cross-validation and scalable approximation. Statistica Sinica 13(3):811–826

Hall P, Patil P (1995) Formulae for mean integrated squared error of nonlinear wavelet-based density estimators. Ann Stat 23(3):905–928

Kass RE, Steffey D (1989) Approximate bayesian inference in conditionally independent hierarchical models (parametric empirical bayes models). J Am Stat Assoc 84(407):717–726

Kauermann G (2005) A note on smoothing parameter selection for penalised spline smoothing. J Stat Plan Infer 127(1–2):53–69

Kauermann G, Krivobokova T, Fahrmeir L (2009) Some asymptotic results on generalized penalized spline smoothing. J R Stat Soc Ser B 71(2):487–503

Kauermann G, Opsomer J (2011) Data-driven selection of the spline dimension in penalized spline regression. Biometrika 98(1):225–230

Komárek A (2006) Accelerated failure time models for multivariate doubly-interval-censored data with flexible distributional assumptions. Ph.D. thesis, Leuven: Katholieke Universiteit Leuven, Faculteit Wetenschappen

Komárek A, Lesaffre E (2008) Generalized linear mixed model with a penalized gaussian mixture as a random-effects distribution. Comput Stat Data Anal 52(7):3441–3458

Komárek A, Lesaffre E, Hilton J (2005) Accelerated failure time model for arbitrarily censored data with smoothed error distribution. J Comput Graph Stat 14(3):726–745

Koo JY, Kooperberg C, Park J (1999) Logspline density estimation under censoring and truncation. Scand J Stat 26(1):87–105

Kooperberg C (2009) logspline: Logspline density estimation routines. R package version 2.1.3.

Li JQ, Barron AR (1999) Mixture density estimation. In: Advances in neural information processing systems 12. MIT Press, Cambridge, pp 279–285

Li Y, Ruppert D (2008) On the asymptotics of penalized splines. Biometrika 95(2):415–436

Lindsey JK (1974) Comparison of probability distributions. J R Stat Soc Ser B 36(1):38–47

Lindsey JK (1974) Construction and comparison of statistical models. J R Stat Soc Ser B 36(3):418–425

Liu L, Levine M, Zhu Y (2009) A functional EM algorithm for mixing density estimation via nonparametric penalized likelihood maximization. J Comput Graph Stat 18(2):481–504

McLachlan G, Peel D (2000) Finite mixture models. Wiley, New York

Müller P, Quintana F, Rosner G (2009) Bayesian Clustering with Regression. University of Texas M.D. Anderson Cancer Center, Houston, TX 77030 U.S.A

Nadaraya E (1974) On the integral mean square error of some nonparametric estimates for the density function. Theory Prob Appl 19(1):133–141

Nadaraya EA (1964) On estimating regression. Theory Prob Appl 9(1):141–142

Nason G (2010) wavethresh: Wavelets statistics and transforms. R package version 4.5

Nason GP (2008) Wavelet methods in statistics with R. Springer, Berlin, ISBN 978-0-387-75960-9

Nason GP, Silverman BW (1999) Wavelets for regression and other statistical problems. In: Schimek MG (ed) Smoothing and regression: approaches, computation, and application, series in probability and statistics. Wiley, New York

O'Sullivan F (1986) A statistical perspective on ill-posed inverse problems. Stat Sci 1(4):502–518

Reiss T, Ogden R (2009) Smoothing parameter selection for a class of semiparametric linear models. J R Stat Soc Ser B 71(2):505–523

Rue H, Martino S, Chopin N (2009) Approximate Bayesian inference for latent Gaussian models by using integrated nested Laplace approximations. J R Stat Soc Ser B 71(2):319–392

Ruppert D (2002) Selecting the number of knots for penalized splines. J Comput Graph Stat 11(4):735–757

Ruppert D, Wand M, Carroll R (2003) Semiparametric regression. Cambridge University Press, Cambridge

Ruppert D, Wand MP, Carroll RJ (2009) Semiparametric regression during 2003–2007. Electron J Stat 3:1193–1256

Schall R (1991) Estimation in generalized linear models with random effects. Biometrika 78(4):719–727

Schellhase C (2010) pendensity: density estimation with a penalized mixture approach. R package version 0.2.3

Searle S, Casella G, McCulloch C (1992) Variance components. Wiley, New York

Sheather SJ, Jones MC (1991) A reliable data-based bandwidth selection method for kernel density estimation. J R Stat Soc Ser B 53(3):683–690

Silverman BW (1982) On the estimation of a probability density function by the maximum penalized likelihood method. Ann Stat 10(3):795–810

Simonoff JS (1996) Smoothing methods in statistics. Springer, New York

Wand M (2003) Smoothing and mixed models. Comput Stat 18(2):223–249

Wand M, Jones MC (1995) Kernel smoothing. Chapman and Hall, London

Wand MP, Ormerod JT (2008) On semiparametric regression with O'Sullivan penalised splines. Aust N Z J Stat 50(2):179–198

Watson G (1964) Smooth regression analysis. Sankhya Ser A 26:359–372

Wood S (2011) Fast stable restricted maximum likelihood and marginal likelihood estimation of semiparametric generalized linear models. J R Stat Soc Ser B 73(1):3–36

Wood SN (2006) Generalized additive models. Chapman and Hall/CRC, London

Young D, Hunter D, Chauveau D, Benaglia T (2009) mixtools: an R package for analyzing mixture models. J Stat Softw 32(6):1–29