RESEARCH PAPER

A transformation-based approach to Gaussian mixture density estimation for bounded data

Luca Scrucca (iD)

Department of Economics, Università degli Studi di Perugia, Italy

Correspondence
Luca Scrucca, Department of Economics, Università degli Studi di Perugia, Via A. Pascoli 20, 06123 Perugia, Italy.
Email: luca.scrucca@unipg.it

Abstract

Finite mixture of Gaussian distributions provide a flexible semiparametric methodology for density estimation when the continuous variables under investigation have no boundaries. However, in practical applications, variables may be partially bounded (e.g., taking nonnegative values) or completely bounded (e.g., taking values in the unit interval). In this case, the standard Gaussian finite mixture model assigns nonzero densities to any possible values, even to those outside the ranges where the variables are defined, hence resulting in potentially severe bias. In this paper, we propose a transformation-based approach for Gaussian mixture modeling in case of bounded variables. The basic idea is to carry out density estimation not on the original data but on appropriately transformed data. Then, the density for the original data can be obtained by a change of variables. Both the transformation parameters and the parameters of the Gaussian mixture are jointly estimated by the expectation-maximization (EM) algorithm. The methodology for partially and completely bounded data is illustrated using both simulated data and real data applications.

KEYWORDS

bounded support, density estimation, EM algorithm, Gaussian mixture models, range-power transformation

1 | INTRODUCTION

Density estimation is the problem of inferring a probability density function given a finite number of sample data points drawn from a population described by a probability distribution. Broadly speaking, three alternative approaches to density estimation can be distinguished. In the parametric approach, a parametric distribution is assumed for the density with unknown parameters, which are estimated by fitting the parametric function using the observed data. Conversely, in the nonparametric approach no density function is assumed a priori, but its form is completely determined by the data. Histograms and kernel density estimation (KDE) are two popular methods that belong to this class, and both are characterized by the number of parameters growing with the size of the data set. Furthermore, extension to higher dimensionality is problematic. A third approach is based on finite mixture models, where the unknown density is expressed as a convex combination of one or more probability density functions. In this class, a popular model is the Gaussian mixture model (GMM) that assumes the Gaussian distribution for the underlying component densities. GMMs can approximate any continuous density with arbitrary accuracy provided the model has a sufficient number of components and the parameters of the model are correctly estimated (Escobar & West, 1995; Roeder & Wasserman, 1997).

Bounded data are quite common in biomedical data analyses because of the measurement scale of the data, or the type of variables under study. However, the standard GMM for density estimation does not take into account whether or not a variable has bounded support. Consider the graphs in Figure 1, which show some histograms for random samples drawn
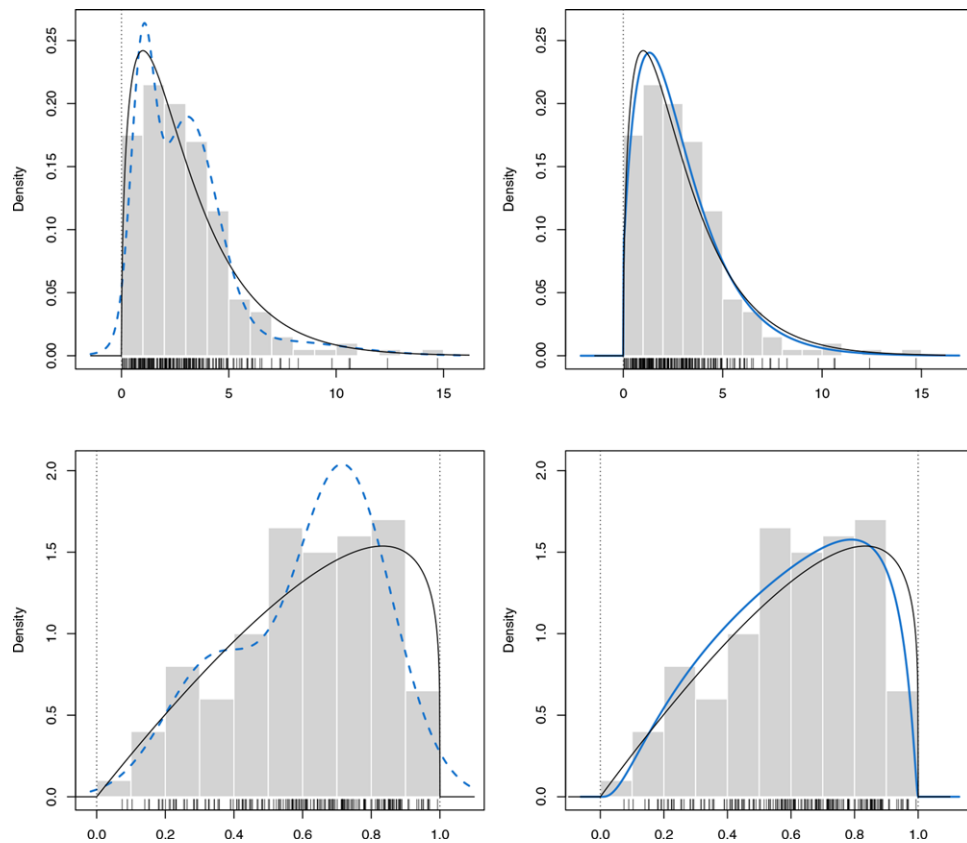
FIGURE 1 Histograms for random samples drawn from a $\chi^2(3)$ (top panels) and a beta(2, 1.2) (bottom panels) distributions with the corresponding density functions (solid lines) and boundaries of the random variable (vertical dotted lines)

Note: Panels on the left show the estimated densities obtained by fitting a GMM on the original scale (blue dashed lines). Panels on the right show the densities estimated by the GMDEB transformation approach (blue solid lines).

from two distributions, one bounded from below (top panels), and one having both lower and upper bounds (bottom panels). In these graphs boundaries are shown as vertical dots, true densities are represented as solid lines, and density estimates based on GMMs as dashed lines (see left panels of Figure 1). In both cases, the estimated densities are unsatisfactory not only at the boundaries but also in the range of admissible values. A possible way to tackle these problems is to abandon the use of GMMs in favor of alternative component distributions. Another option is to remain in the realm of the Gaussian mixtures framework, but analyze the data in a transformed scale. The right panels of Figure 1 show the density estimates obtained with the Gaussian mixture density estimation for bounded (GMDEB) data approach proposed in this paper. In both cases, the true underlying densities appear to be well approximated, and with natural boundaries constraints clearly satisfied.

In this paper, a transformation-based approach to density estimation based on GMMs is proposed and discussed. The basic idea is to use an invertible function to map a bounded variable to an unbounded support, estimate the density of the transformed variable, and then back-transform to the original scale. This approach seems very natural and it has been around for a long time (see Marron & Ruppert, 1994; Wand, Marron & Ruppert, 1991, for KDE), but a simple and efficient implementation of this methodology is not yet available in the context of mixture density estimation. Note that a similar approach based on Manly transformation has been recently proposed by Zhu and Melnykov (2018) for modeling skewed data in model-based clustering. However, our proposal differs in two main respects: first, it has been designed to allow variables with bounded support, second, it aims at a different goal, that is, density estimation as compared to clustering.

In Section 2, the GMMs approach to density estimation is reviewed. Then, Section 3 presents the proposed range–power transformation method for density estimation using GMMs in case of bounded variables. The model is described and the corresponding maximum likelihood estimates are derived through the EM algorithm. Section 4 contains the results of some simulation studies carried out to evaluate the proposed methodology and to compare it with other available methods. In Section 5, some real-world data sets are analyzed. The final section provides some concluding remarks.

2 | FINITE MIXTURE MODELING

2.1 | Finite mixture for density estimation

Consider a vector of random variables $\boldsymbol{x}$ taking values in the sample space $S_{\mathcal{X}} \subseteq \mathbb{R}^p$ with $p \geq 1$, and assume that the probability density function can be written as a finite mixture density of G components of the form

$$f(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g f_g(\boldsymbol{x}; \boldsymbol{\theta}_g),$$

where $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\theta}_1, \ldots, \boldsymbol{\theta}_G)^\top$ is the parameters vector. The mixing weights $(\pi_1, \ldots, \pi_G)$ must satisfy the constraints $\pi_g > 0$ for all $g = 1, \ldots, G$, and $\sum_{g=1}^{G} \pi_g = 1$. The gth component density $f_g(\boldsymbol{x}; \boldsymbol{\theta}_g)$ is usually taken as known except for the associated parameter(s) $\boldsymbol{\theta}_g$. Most applications assume that all component densities arise from the same parametric distribution family, although this need not be the case in general. In particular, a popular model specifies $f_g(\boldsymbol{x}; \boldsymbol{\theta}_g) \equiv \phi(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)$, where $\phi(\cdot)$ is the Gaussian density with mean $\boldsymbol{\mu}_g$ and covariance matrix $\boldsymbol{\Sigma}_g$. Then, the GMM can be written as

$$f(\boldsymbol{x}; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g \phi(\boldsymbol{x}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g), \tag{1}$$

where in this case $\boldsymbol{\Psi} = (\pi_1, \ldots, \pi_{G-1}, \boldsymbol{\mu}_1^\top, \ldots, \boldsymbol{\mu}_G^\top, \text{vech}\{\boldsymbol{\Sigma}_1\}^\top, \ldots, \text{vech}\{\boldsymbol{\Sigma}_G\}^\top)$ represents the entire parameters vector. Note that $\text{vech}\{\cdot\}$ is an operator that forms a vector by extracting unique elements of a symmetric matrix.

In this paper, we refer to (1) as the Gaussian mixture density estimate (GMDE) model. The usual nonparametric KDE can be viewed as a mixture of $G = n$ components with uniform weights, that is, $\pi_g = 1/n$ (Titterington, Smith & Makov, 1985, pp. 28–29). Compared to KDE, finite mixture modeling uses a smaller number of components (i.e., less parameters), so it has smaller variance. Conversely, compared to parametric density estimation, finite mixture modeling has the advantage of (potentially) using more parameters, so introducing less estimation bias. There are also disadvantages related to mixture modeling, such as an increased learning complexity and lack of closed-form solution, so it needs to resort to numerical procedures (e.g., EM algorithm), and in certain cases there can be identifiability issues.

2.2 | Estimation of Gaussian finite mixture model

Consider a random sample $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_n$ of n observations on p variables drawn from the mixture distribution in (1). Then, the log-likelihood is given by

$$\ell(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \log \sum_{g=1}^{G} \pi_g \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g). \tag{2}$$

Direct maximization of the log-likelihood function is not straightforward, so MLEs are usually obtained via the expectation-maximization (EM) algorithm (Dempster, Laird, & Rubin, 1977; McLachlan & Peel, 2000).

An incomplete-data formulation of the mixture problem is introduced by associating to each observation a latent component-label vector $\boldsymbol{z}_i$ ($i = 1, \ldots, n$). This is a G-dimensional vector, with the generic element $z_{ig} = 1$ or 0 according to whether or not $\boldsymbol{x}_i$ arises from the gth component of the mixture. Assuming independence of the complete-data vector $(\boldsymbol{x}_i^\top, \boldsymbol{z}_i^\top)^\top$, and the multinomial distribution for the component-label vectors $\boldsymbol{z}_i$s, the complete-data log-likelihood is given by

$$\ell_C(\boldsymbol{\Psi}) = \sum_{i=1}^{n} \sum_{g=1}^{G} z_{ig} \{\log \pi_g + \log \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\}. \tag{3}$$

The log-likelihood (2) is maximized using the EM algorithm, an iterative algorithm that alternates two steps, called E-step and M-step, which guarantees, under fairly general conditions, the convergence to at least a local maximizer. The objective function at iteration $(m + 1)$ of the EM algorithm is the conditional expectation of the complete-data log-likelihood (3), the so-called Q-function:

$$Q(\boldsymbol{\Psi}; \boldsymbol{\Psi}^{(m)}) = \sum_{i=1}^{n} \sum_{g=1}^{G} \hat{z}_{ig}^{(m)} \{\log \pi_g + \log \phi(\boldsymbol{x}_i; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g)\},$$

where $\widehat{z}_{ig}^{(m)} = \mathrm{E}(I(z_i = g)|\boldsymbol{x}_i, \boldsymbol{\Psi}^{(m)})$, that is, the estimated posterior probability at iteration m of the EM algorithm, with $I(\cdot)$ the indicator function, which equals 1 if the condition is fulfilled and 0 otherwise.

In the E-step the Q-function is evaluated, using the parameter values $\pi_g, \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g$ obtained at the previous step, to get the updated posterior probabilities

$$\widehat{z}_{ig}^{(m+1)} = \frac{\widehat{\pi}_g^{(m)} \phi(\boldsymbol{x}_i; \widehat{\boldsymbol{\mu}}_g^{(m)}, \widehat{\boldsymbol{\Sigma}}_g^{(m)})}{\sum_{k=1}^{G} \widehat{\pi}_k^{(m)} \phi_k(\boldsymbol{x}_i; \widehat{\boldsymbol{\mu}}_k^{(m)}, \widehat{\boldsymbol{\Sigma}}_k^{(m)})}.$$

Then, in the M-step the parameters vector $\boldsymbol{\Psi}$ is updated by maximizing the Q-function given the previous values $\widehat{\boldsymbol{\Psi}}^{(m)}$ and the updated posterior probabilities $\widehat{z}_{ig}^{(m+1)}$, that is,

$$\widehat{\boldsymbol{\Psi}}^{(m+1)} = \arg \max_{\boldsymbol{\Psi}} Q(\boldsymbol{\Psi}; \widehat{\boldsymbol{\Psi}}^{(m)}).$$

In the case of a multivariate Gaussian mixture the M-step yields

$$\widehat{\pi}_g^{(m+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)}}{n} \qquad \text{and} \qquad \widehat{\boldsymbol{\mu}}_g^{(m+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)} \boldsymbol{x}_i}{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)}}.$$

The update formula for the covariance matrix depends on the structure of the within-component covariance matrices. Parsimonious parameterization of the covariance matrices can be expressed through the eigendecomposition $\boldsymbol{\Sigma}_g = v_g \boldsymbol{O}_g \boldsymbol{A}_g \boldsymbol{O}_g^{\top}$, where v_g is a scalar controlling the volume of the corresponding ellipsoid, $\boldsymbol{A}_g$ is a diagonal matrix specifying the shape of the density contours, and $\boldsymbol{O}_g$ is an orthogonal matrix, which determines the orientation of the ellipsoid (Banfield & Raftery, 1993; Celeux & Govaert, 1995). For instance, assuming an unconstrained covariance matrix, the updating formula is

$$\widehat{\boldsymbol{\Sigma}}_g^{(m+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)} \left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_g^{(m+1)} \right) \left(\boldsymbol{x}_i - \widehat{\boldsymbol{\mu}}_g^{(m+1)} \right)^{\top}}{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)}}.$$

Scrucca, Fop, Murphy, and Raftery (2016, table 3) summarize some parameterizations of within-component covariance matrices, and the corresponding geometric characteristics, currently available in the `mclust` software. The previous unconstrained covariance matrix is indicated as VVV model. Note, however, that for some models no closed-formula is available, so numerical optimization is required.

The EM algorithm requires the specification of initial values for the parameters, say $\boldsymbol{\Psi}^{(0)}$. Alternatively, an initial assignment of observations to the components of the mixture can be made, basically starting the EM algorithm from the M-step. The initialization of the EM algorithm is often crucial because the likelihood surface tends to have multiple modes, and in certain cases it can be even unbounded. Nevertheless, the EM algorithm usually produces sensible results when started from reasonable starting values (Wu, 1983, p. 150). For a further discussion on this point and a recent proposal, see Scrucca and Raftery (2015).

Information criteria based on penalized forms of the log-likelihood are routinely used in finite mixture modeling for model selection, that is, to decide not only how many components should be included in the mixture but also which covariance parameterizations to adopt in the Gaussian case. Two popular criteria are the Bayesian information criterion (BIC; Fraley & Raftery, 1998; Schwartz, 1978) and the integrated complete-data likelihood (ICL) criterion (Biernacki, Celeux & Govaert, 2000). When the goal is density estimation, Roeder and Wasserman (1997) showed that the GMDE model selected using BIC is a consistent estimator of the true density. If only the order of the mixture is needed, formal hypothesis testing can also be pursued by the likelihood ratio test (LRT). However, standard regularity conditions do not hold for the null distribution of the LRT statistic to have its usual chi-squared distribution (McLachlan & Peel, 2000, Chap. 6), and significance must be assessed by resampling approaches. For a recent review, see McLachlan and Rathnayake (2014), and for an implementation in the `mclust` software see Scrucca et al. (2016).

3 | METHODOLOGY

3.1 | GMDE for variables with bounded support

Let $\boldsymbol{x}$ be a p-variate random vector from a distribution with density f having bounded support $S_{\mathcal{X}} \subset \mathbb{R}^p$, and $\{t(\boldsymbol{x}; \lambda); \lambda \in \boldsymbol{\Lambda}\}$ be some family of continuous monotonic transformations that map $S_{\mathcal{X}}$ to an unbounded p-dimensional support. Then, we can write $\boldsymbol{y} = t(\boldsymbol{x}; \lambda)$ as the transformed set of variables with density h having unbounded support $S_{\mathcal{Y}}$.

Suppose that the density of the transformed data can be expressed through a Gaussian finite mixture density of the form

$$h(\boldsymbol{y}; \boldsymbol{\Psi}) = \sum_{g=1}^{G} \pi_g \phi(\boldsymbol{y}; \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g). \tag{4}$$

Then, by the continuous change of variable theorem, the density of the untransformed data can be expressed as

$$f(\boldsymbol{x}; \boldsymbol{\Psi}, \lambda) = h(t(\boldsymbol{x}; \lambda)) \cdot |\boldsymbol{J}(t(\boldsymbol{x}; \lambda))|,$$

where $\boldsymbol{J}(t(\boldsymbol{x}; \lambda))$ is the Jacobian of the transformation, that is, the determinant of the matrix of partial derivatives.

3.2 | Range–power transformation for variables with bounded support

3.2.1 | Lower bound case

Suppose x is a univariate random variable with lower bounded support $S_{\mathcal{X}} \equiv (l, \infty)$, where $l > -\infty$ and density $f(x)$. Consider a preliminary range transformation defined as $x \mapsto (x - l)$, which maps $S_{\mathcal{X}} \to \mathbb{R}^+$. Let $\{t(x; \lambda \in \Lambda)\}$ be a continuous monotonic transformation. Based on the well-known Box–Cox transformation (Box & Cox, 1964), we consider the following *range–power* transformation

$$t(x; \lambda) = \begin{cases} \dfrac{(x - l)^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log(x - l) & \text{if } \lambda = 0, \end{cases} \tag{5}$$

which has continuous first derivative equal to $t'(x; \lambda) = (x - l)^{\lambda - 1}$ for any $\lambda \in \Lambda$.

The original Box–Cox power transformation method is restricted to the univariate case, but it can be extended also to the multivariate case as described in Velilla (1993). However, further development of the multivariate case $\boldsymbol{x} = (x_1, \dots, x_p)^{\top}$ can be greatly simplified by working in a coordinate-wise fashion. Thus, in this paper, we propose the use of the range-power transformation in (5) for each dimension separately.

3.2.2 | Lower and upper bound case

Suppose now that x is a univariate random variable with bounded support $S_X \equiv (l, u)$, where $-\infty < l < u < +\infty$. Consider the preliminary range transformation $x \mapsto (x - l)/(u - x)$, which maps $S_X \to \mathbb{R}^+$. As in the previous case, adopting a *range–power* transformation we can write

$$t(x; \lambda) = \begin{cases} \dfrac{\left(\dfrac{x - l}{u - x}\right)^{\lambda} - 1}{\lambda} & \text{if } \lambda \neq 0 \\ \log\left(\dfrac{x - l}{u - x}\right) & \text{if } \lambda = 0, \end{cases} \tag{6}$$

with continuous first derivative given by

$$t'(x; \lambda) = \begin{cases} \left(\dfrac{x - l}{u - x}\right)^{\lambda - 1} \dfrac{u - l}{(u - x)^2} & \text{if } \lambda \neq 0 \\ \dfrac{1}{x - l} + \dfrac{1}{u - x} & \text{if } \lambda = 0. \end{cases}$$

Following the approach discussed in Section 3.2.1, the multivariate case can be tackled by working in a coordinate-wise fashion, hence applying the range–power transformation in (6) to each variable separately.

3.3 | Estimation

Maximum likelihood estimation can be pursued via the EM algorithm under the assumption that the density on the transformed scale can be expressed as in (4), with $\mathbf{y} = t(\mathbf{x}; \lambda)$ the vector of range–power transformed variables according to (5) or (6). If the previous assumption holds, then the density function on the original scale is given by

$$f(\mathbf{x}; \mathbf{\Psi}, \lambda) = \sum_{g=1}^{G} \pi_g \phi(t(\mathbf{x}; \lambda); \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) \cdot |\mathbf{J}(t(\mathbf{x}; \lambda)|, \tag{7}$$

where $t(\mathbf{x}; \lambda) = (t(x_1; \lambda_1), \dots, t(x_p; \lambda_p))^\top$ and $\mathbf{J}(t(\mathbf{x}; \lambda))$ is the Jacobian of the transformation. Note that as a consequence of the coordinate independent approach to multivariate range–power transformation, the matrix of first derivatives is diagonal, so the Jacobian reduces to

$$\mathbf{J}(t(\mathbf{x}; \lambda)) = \det\left[\frac{\partial t(\mathbf{x}; \lambda)}{\partial \mathbf{x}}\right] = \prod_{j=1}^{p} \frac{\partial t(x_j; \lambda_j)}{\partial x_j}.$$

The conditional expectation of the complete-data log-likelihood given the observed data can be expressed as

$$Q(\mathbf{\Psi}; \mathbf{\Psi}^{(m)}) = \sum_{i=1}^{n} \sum_{g=1}^{G} \widehat{z}_{ig}^{(m)} \left\{\log \pi_g + \log \phi(t(\mathbf{x}_i; \lambda); \boldsymbol{\mu}_g, \boldsymbol{\Sigma}_g) + \log |\mathbf{J}(t(\mathbf{x}_i; \lambda))|\right\},$$

where $\widehat{z}_{ig}^{(m)} = \mathrm{E}(I(z_i = g)|\mathbf{x}_i, \mathbf{\Psi}^{(m)})$. Therefore, in the E-step the posterior probabilities are updated using

$$\widehat{z}_{ig}^{(m+1)} = \frac{\widehat{\pi}_g^{(m)} \phi\left(t(\mathbf{x}_i; \widehat{\lambda}^{(m)}); \widehat{\boldsymbol{\mu}}_g^{(m)}, \widehat{\boldsymbol{\Sigma}}_g^{(m)}\right)}{\sum_{k=1}^{G} \widehat{\pi}_k^{(m)} \phi\left(t(\mathbf{x}_i; \widehat{\lambda}^{(m)}); \widehat{\boldsymbol{\mu}}_k^{(m)}, \widehat{\boldsymbol{\Sigma}}_k^{(m)}\right)}.$$

In the M-step, the parameters $(\mathbf{\Psi}, \lambda)$ are updated by maximizing the Q-function given the previous values of the parameters and the updated posterior probabilities. According to the expectation-conditional-maximization (ECM) algorithm proposed by Meng and Rubin (1993), this maximization can be pursued in two steps. In the first step, an updated value $\widehat{\lambda}^{(m+1)}$ is computed by numerically maximizing the Q-function with respect to λ because no closed-form expression is available. To this goal, a Newton-type numerical optimization algorithm can be used. Although, in theory, λ parameters are unrestricted, in practice reasonable values are typically found in a narrower range, for example, between -3 and 3. For this reason, in our implementation we used the L-BFGS-B method of Byrd, Lu, Nocedal, and Zhu (1995) available in the optim() function for the R statistical software. The remaining parameters are then obtained as in standard EM algorithm but accounting for the updated transformation parameters $\widehat{\lambda}^{(m+1)}$, that is,

$$\widehat{\pi}_g^{(m+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)}}{n} \quad \text{and} \quad \widehat{\boldsymbol{\mu}}_g^{(m+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)} t(\mathbf{x}_i; \widehat{\lambda}^{(m+1)})}{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)}}.$$

Again, the update formula for the covariance matrix depends on the assumed eigendecomposition model. In the most general case of an unconstrained covariance matrix, that is, the VVV model, we have

$$\widehat{\boldsymbol{\Sigma}}_g^{(m+1)} = \frac{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)} \left(t(\mathbf{x}_i; \widehat{\lambda}^{(m+1)}) - \widehat{\boldsymbol{\mu}}_g^{(m+1)}\right)\left(t(\mathbf{x}_i; \widehat{\lambda}^{(m+1)}) - \widehat{\boldsymbol{\mu}}_g^{(m+1)}\right)^\top}{\sum_{i=1}^{n} \widehat{z}_{ig}^{(m+1)}}.$$

Initialization of the above EM algorithm is obtained by first estimating the optimal marginal transformations, then using the final classification from a k-means algorithm on the range–power transformed variables. This initial partition of data points is used to start the algorithm from the M-step. Finally, the EM algorithm is stopped when the log-likelihood improvement falls below a specified tolerance value or a maximum number of iterations is reached.
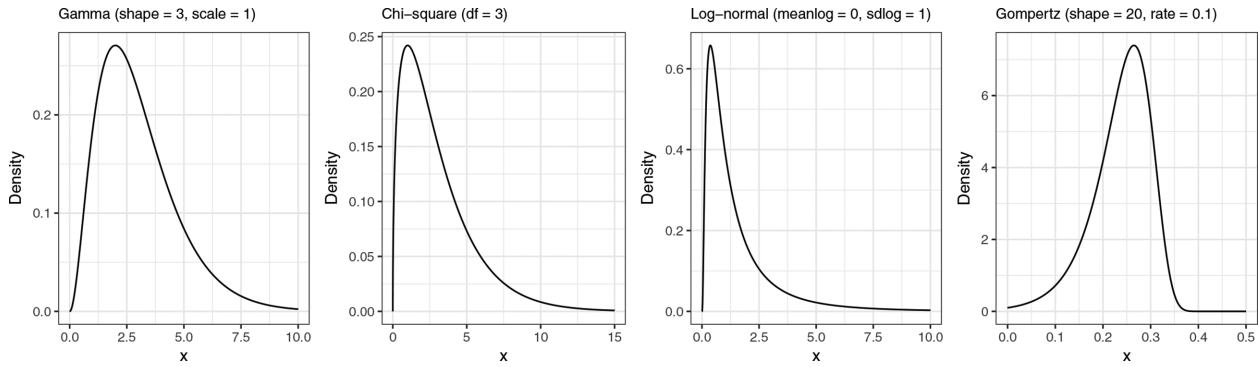
FIGURE 2   Univariate densities with lower bound considered in the simulation study

4 | SIMULATION STUDIES

In this section, we present some simulation studies designed to compare the proposed GMDEB approach to some density estimators for bounded variables discussed in the literature. The comparison is based on the integrate squared error (ISE):

$$\text{ISE}(\widehat{f}) = \int \left[\widehat{f}(x) - f(x) \right]^2 dx,$$

where f is the unknown true density and $\widehat{f}$ is its estimate based on a random sample of n observations. Thus, the ISE is a measure of discrepancy between the true and the estimated density based on a squared loss criterion. It is equal to 0 when the estimated density perfectly coincides with the true density, and increases as the differences between the two densities get larger. For more details, see Scott (2009, section 2.3). In the following sections, the ISE is computed via numerical integration.

4.1 | Distributions with lower bound

The univariate densities with lower bounded support considered in this study are shown in Figure 2. They are all bounded at zero with different degrees of skewness, positive for the first three densities and negative for the last one.

The proposed method (GMDEB) is compared with the following density estimators:

- simpleKern, which refers to the simple boundary correction method proposed by Jones (1993), which is equivalent to a kernel weighted local linear fitting near the boundary;

- reflectKern, which indicates the reflection method of Schuster (1985), which amounts to reflect the observed data points at the origin, then a density estimate is obtained using this augmented data set with a simple correction to ensure that integrate to one;

- cutnormKern, which is the cut and normalization method of Gasser and Müller (1979), where the kernel is truncated at the boundary and renormalized to unity;

- logtransKern, which is the method proposed by Marron and Ruppert (1994), which fits a KDE on the log-scale and then back-transforms the result with an explicit normalization step;

- logSpline, which estimates a density using cubic splines to approximate the log-density using knots located as described in Stone, Hansen, Kooperberg, and Truong (1997);

- GaMixDE, which estimates a density by fitting a mixture of Gamma densities;

- GMDE, which is the standard density estimate from GMMs with no boundary correction.

The first four methods mentioned above are implemented in the evmix R package (Hu & Scarrott, 2018; Scarrott, Hu, Akbar & of Canterbury, 2018), whereas the logSpline estimator is available in the logspline R package (Kooperberg, 2016). For GaMixDE, the code is available in the R package mixtools (Benaglia, Chauveau, Hunter & Young, 2009; Young et al., 2017), whereas GMDE is obtained from the mclust R package (Fraley, Raftery, & Scrucca, 2017). In the last two cases, the number of mixture components is selected using the BIC criterion.

Figure 3 graphically summarizes the simulation results obtained on 1,000 replications. Overall, the GMDEB estimator appears to be able to approximate the true density better than the other KDE methods with boundary correction, in particular when the
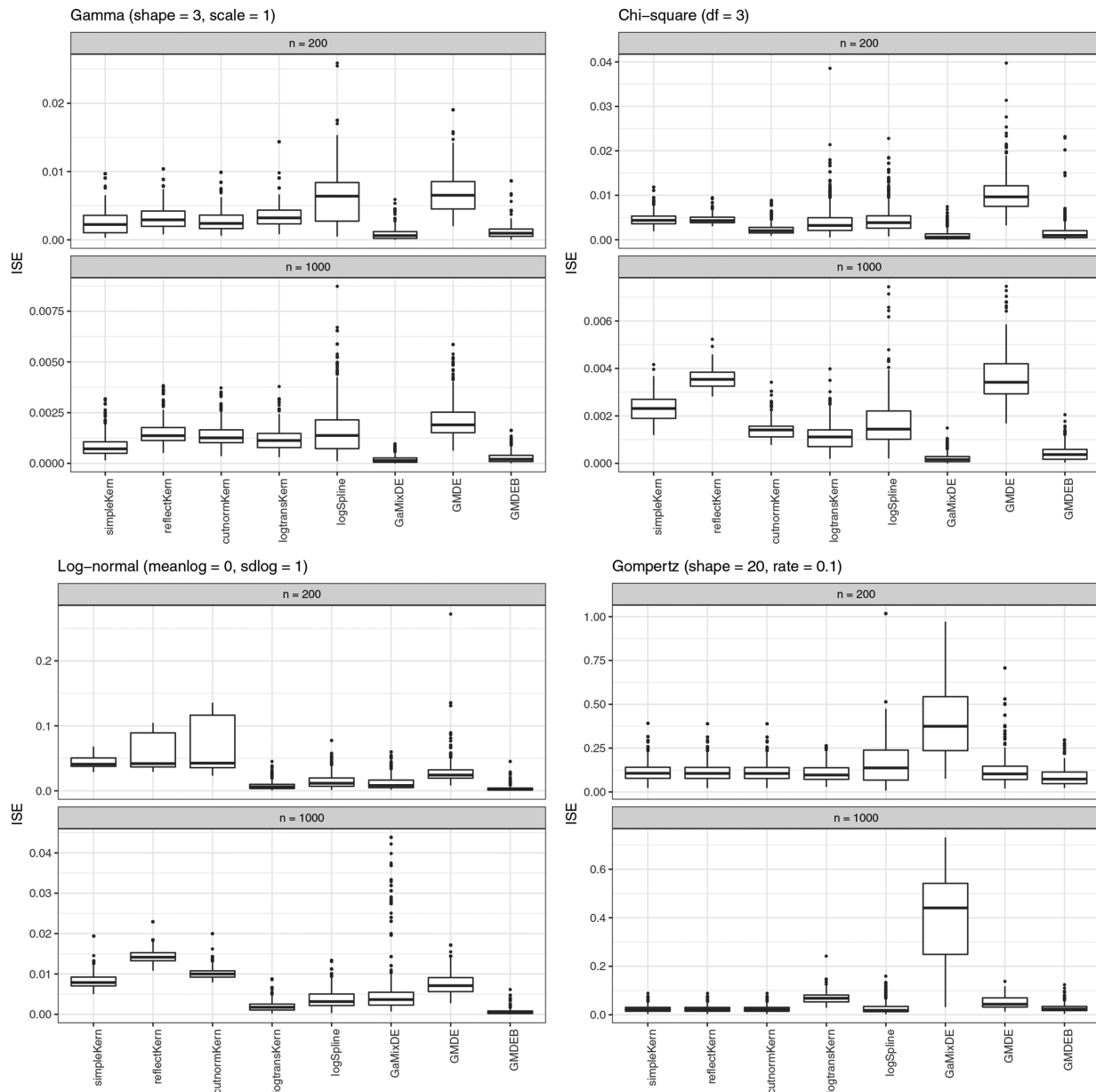
FIGURE 3 Boxplots of ISE distribution from 1,000 replications of the simulation study for the selected univariate densities with lower bound

sample size is small. The proposed approach also shows less variability, which decreases as the sample size increases. Clearly, the GMDEB approach appears to be inferior to the Gamma mixture density estimator, although not by much, when the true density belongs to the Gamma family of distributions (i.e., in the first two cases), but it is better in the last two cases, in particular for the left-skewed Gompertz distribution.

4.2 | Distributions with lower and upper bounds

For the case of univariate densities with both lower and upper bounded support, a list of distributions considered in the simulation study is shown in Figure 4. The first two settings involve the beta distribution with parameters selected to produce a symmetric case and a skewed case. The last two settings consider two further asymmetric cases, the Kumaraswamy distribution with density $f(x) = \alpha \beta x^{\alpha-1}(1 - x^{\alpha})^{\beta-1}$ on [0,1] and the logarithmic peak distribution with density $f(x) = -\log(x)$ on $(0, 1)$.

The cases described above should provide a broad spectrum of examples for comparing different estimators. To this goal, the proposed method (GMDEB) is compared with the following density estimators:
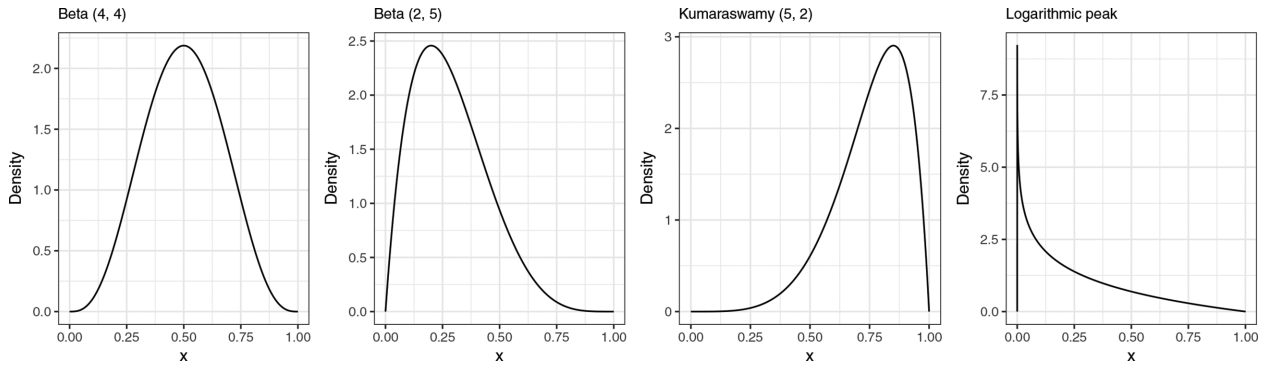
FIGURE 4   Univariate densities with lower and upper bounds considered in the simulation study

- `beta1Kern`, `beta2Kern`, which use the beta and modified beta kernels proposed by Chen (1999) followed by a renormalization to ensure a proper density;
- `copulaKern`, which uses the bivariate Gaussian copula–based kernels of Jones and Henderson (2007);
- `logSpline`, which fits a density using cubic splines to approximate the log-density using knots located as described in Stone et al. (1997);
- `BeMixDE`, which estimates the density by fitting a mixture of beta distributions.

The first two estimators are available in the R package `evmix` (Hu & Scarrott, 2018; Scarrott et al., 2018). The `logSpline` estimator is available in the `logspline` R package (Kooperberg, 2016). For the `BeMixDE` the `betareg` R package (Grün, Kosmidis & Zeileis, 2012) is used with the number of mixture components selected using BIC.

Figure 5 reports the simulation results obtained on 1,000 replications. By looking at the boxplots, the GMDEB approach appears to be more accurate than the other nonparametric density estimators. Furthermore, its accuracy is slightly lower than the beta mixture density estimator in the first three cases (which, however, are all cases related to the beta distribution), but is better in the last case. Overall, GMDEB seems to provide robust reliable density estimates when both lower and upper bounds are present.

4.3 │ A note on computing time

A major concern might be the computational effort required by the estimation of the λ parameter through numerical optimization within the EM algorithm. To investigate the runtime of the proposed procedure, we designed a small simulation study where we generated data from a chi-square distribution with 3 degrees of freedom for sample sizes $n = \{200, 1000, 10000\}$. A multivariate case was also investigated by considering a 10-dimensional variable with independent marginals drawn as in the univariate case. We estimated the density of GMDEB by both estimating the lambda parameter, and by fixing it at the corresponding MLE value. Furthermore, we executed the algorithm both sequentially and in parallel (over the mixture components and covariance parameterizations). Experiments were carried out on an iMac with 4 cores i5 Intel CPU running at 2.8 GHz and with 16GB of RAM.

Figure 6 shows the results averaged over 100 replications of the experiments. Clearly, in the univariate case the effect of estimating the transformation parameter is negligible. On the contrary, the effect is visible in the multivariate case, but a considerable speedup can be achieved by parallelization. The worst case, that is, transformation parameter to be estimated sequentially with a sample of size 10,000 on 10 dimensions, required on average just over 1 minute.

5 │ REAL DATA ANALYSES

5.1 │ Acidity data

This data set provides the values of an acidity index (acid-neutralizing capacity, `ANC`) measured in a sample of 155 lakes in North-Central Wisconsin. Several authors have previously analyzed the data using a mixture of Gaussian distributions on the log-scale (Crawford, 1994; Crawford, DeGroot, Kadane, & Small, 1992; McLachlan & Peel, 2000; Richardson & Green, 1997).
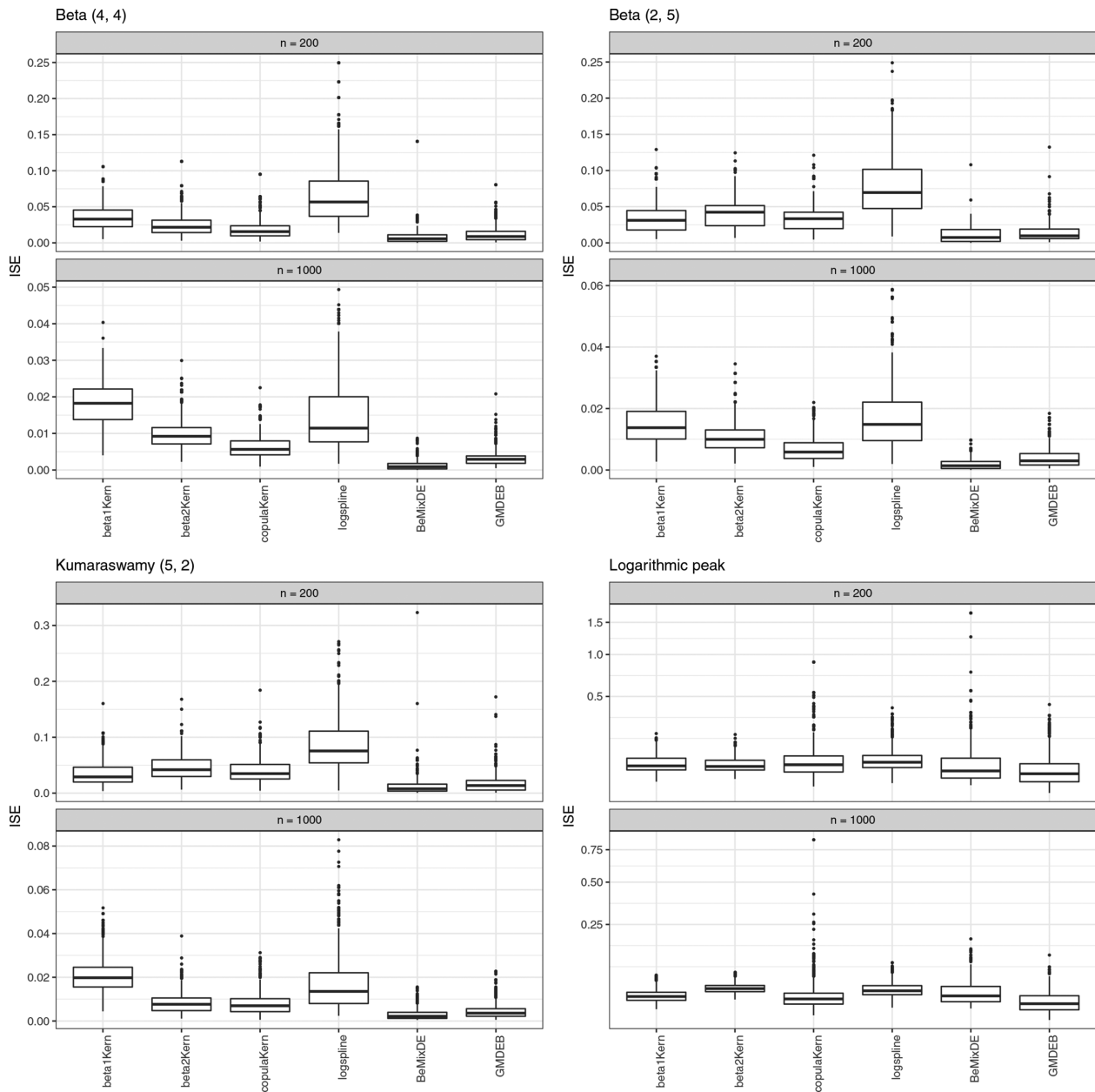
FIGURE 5 Boxplots of ISE distribution from 1,000 replications of the simulation study for the selected univariate densities with lower and upper bounds

On the contrary, we analyze the data in the original scale because the proposed method automatically selects the "optimal" transformation and takes into account the implicit lower bound of the index that cannot assume negative values.

From the left panel of Figure 7 we can see that according to BIC the best model is the one with two mixture components having different variances (V,2), closely followed by models (E,2) and (V,3). The right panel of Figure 7 shows the histogram of the data and the density estimated with model (V,2) using the GMDEB approach with transformation parameter $\hat{\lambda} = -0.293$ (blue thick line). For comparison, we also draw the density estimated by GMM without any boundary correction (black dashed line). The density estimated by GMDEB appears to accurately follow the distribution of the data, indicating the presence of two separated skewed distributions having different dispersions, smaller for the component close to the origin and larger for higher values of ANC. This is also confirmed by the graphs in Figure 8, which show the component densities scaled by the estimated prior probabilities $\hat{\pi} = (0.6322, 0.3678)$ (left panel) and the estimated posterior probabilities $\hat{z}_{ij}$. Using the standard cutoff value of 0.5, lakes with ANC smaller than about 232 are assigned to the first group, otherwise to the second group. This is in agreement with previous findings.

FIGURE 6 Average run times for different sample sizes (expressed in $\log_{10}$-scale) obtained by considering the transformation parameter either fixed or to be estimated, by running the GMDEB algorithm sequentially or in parallel, and for number of variables 1 and 10
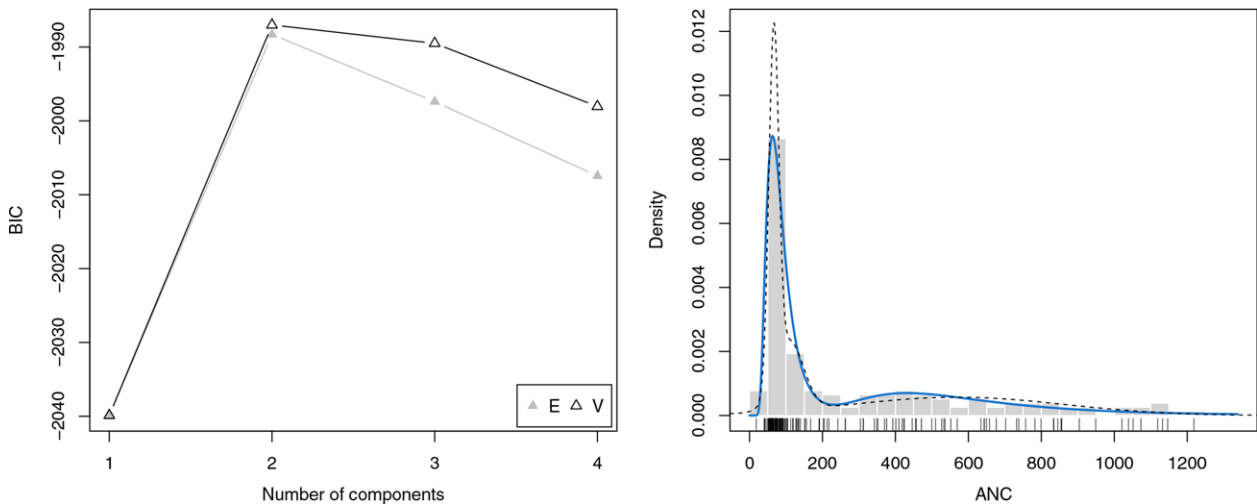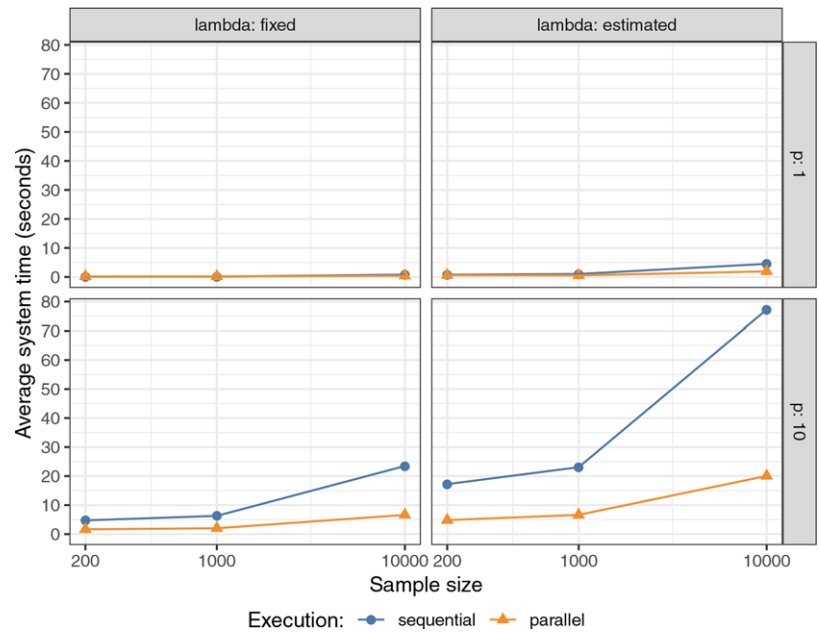

FIGURE 7 Plot of BIC values for different number of mixture components and within-component variances (left panel); histogram of acidity data with GMDEB (blue thick line) and GMM (black dashed line) density estimates (right panel)
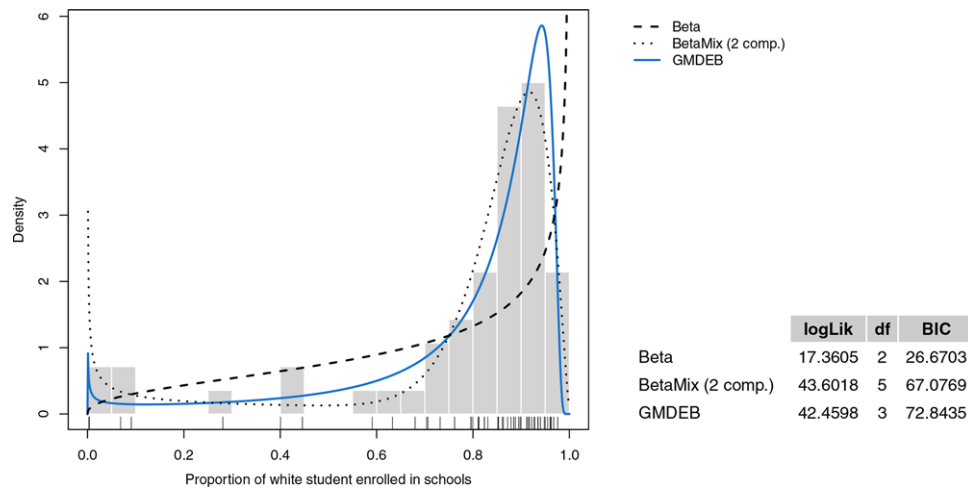
5.2 │ Racial data

Geenens (2013) presented an analysis on data giving the proportion of white students enrolled in 56 school districts in Nassau County (Long Island, New York), for the 1992–1993 school year. The density estimate for this data set should only be supported on the [0,1] range. See also Simonoff (1996, section 3.2).

The selected model on the transformed scale is (E,1) with $\hat{\lambda} = 0.387$, and the corresponding density estimate on the original scale is shown in Figure 9. This can be compared graphically with the beta density and the beta mixture using two components. Both models were estimated by maximum likelihood using the `betareg` R package (Grün et al., 2012), with a single component in the first case, and the optimal number of mixture components selected using BIC in the second case. The beta density seems to put too much emphasis close to the upper boundary and in the middle values of the distribution, while completely missing the bulk of the data between 70 and 90% of white students. On the contrary, the proposed GMDEB approach provides an estimate of the density that correctly identify the majority of the data with at least 70% of white students, but also the small peak near the lower boundary containing schools with almost 0% white students. The two-component beta mixture density is quite close to that provided by GMDEB, but the latter should be preferred according to BIC (see table in Figure 9). These findings largely agree with those reported in Geenens (2013, figure 3).
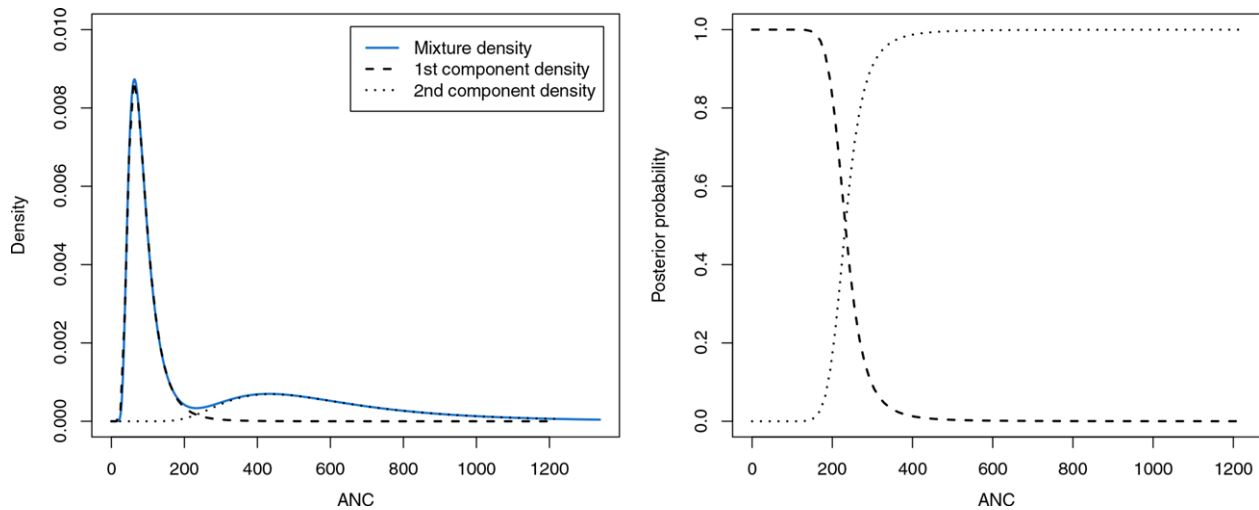
FIGURE 8 Plot of estimated mixture density and rescaled component densities (left panel), and corresponding estimated posterior probabilities (right panel) for the acidity data

FIGURE 9 Density estimates for the racial data obtained using the GMDEB method, the standard beta distribution, and a two-component beta mixture

Note: The included table reports the log-likelihood, the number of estimated parameters, and the BIC (larger values are preferred).

5.3 | Plasma data

Consider the data from a study on the association between the low plasma concentrations of retinol, beta-carotene, or other carotenoids on the increased risk of developing certain types of cancer (Nierenberg et al., 1989). The joint distribution of plasma retinol (ng/mL) and plasma beta-carotene (ng/mL) is bounded below at zero for both variables. The left panel of Figure 10 shows the scatterplot of data points observed on 314 patients.

If the bivariate density is estimated using the standard GMM, the model with the largest BIC = −8101.773 has three components and unconstrained covariance matrices (VVV). This relatively large number of components is related to the presence of a strong skewness in the data distribution. Furthermore, a nonnegligible mass of density is assigned to negative values of plasma beta-carotene.

Both issues can be solved using the proposed range-power transformation approach. The selected model for the bivariate density estimation has BIC = −8044.852, with a diagonal equal variance structure and a single component (EII,1). The transformation parameters are estimated as $\hat{\lambda} = (0.155, 0.0295)$. The right panel of Figure 10 shows the highest density regions (HDRs; Hyndman, 1996) corresponding to proportions $(0.25, 0.5, 0.75, 0.9)$. In this case, the joint data distribution appears to be well approximated by the estimated density.
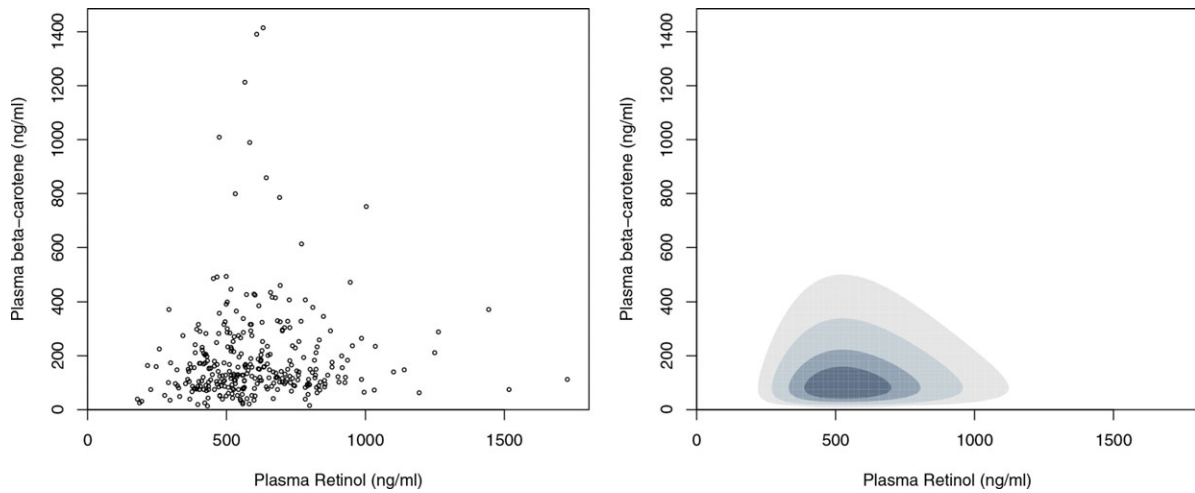
FIGURE 10 Plot of data points for the plasma data set (left panel) and the corresponding bivariate density estimated using the GMDEB approach (right panel)
Note: In the latter case, the graph shows the highest density regions corresponding to proportions $(0.25, 0.5, 0.75, 0.9)$.
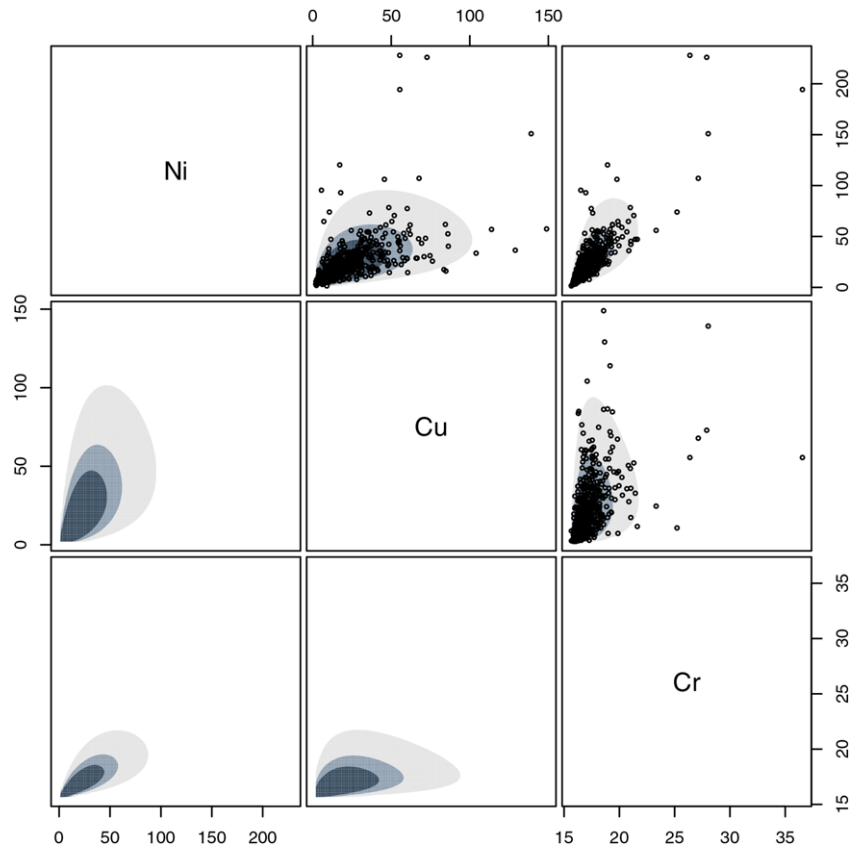

FIGURE 11 Scatterplot matrix of chromium (Cr), copper (Cu), and nickel (Ni) concentrations on the C-horizon layer
Note: The density estimated using the GMDEB approach is shown using HDRs corresponding to 25%, 50%, and 75% probability regions.

5.4 | C-horizon layer of the Kola data

The Kola Ecogeochemistry Project (1993–1998) collected data on more than 50 chemical elements on four different primary sample materials: terrestrial moss, and the O-, B-, and C-horizon of podzolic soils located in parts of northern Finland, Norway, and Russia. The main aim of the project was the documentation of the impact of the Russian nickel industry on the Arctic environment. The data are available on Reimann et al. (1998), see also Reimann, Filzmoser, Garrett, and Dutter (2011). Here, we analyze the distribution of nickel (Ni), copper (Cu), and chromium (Cr) on the C-horizon layer. For each of the 605 sites, the detected concentrations of the abovementioned heavy metals are provided. Clearly, concentrations are bounded below at zero, and a preliminary data exploration suggests that the joint distribution is highly skewed.
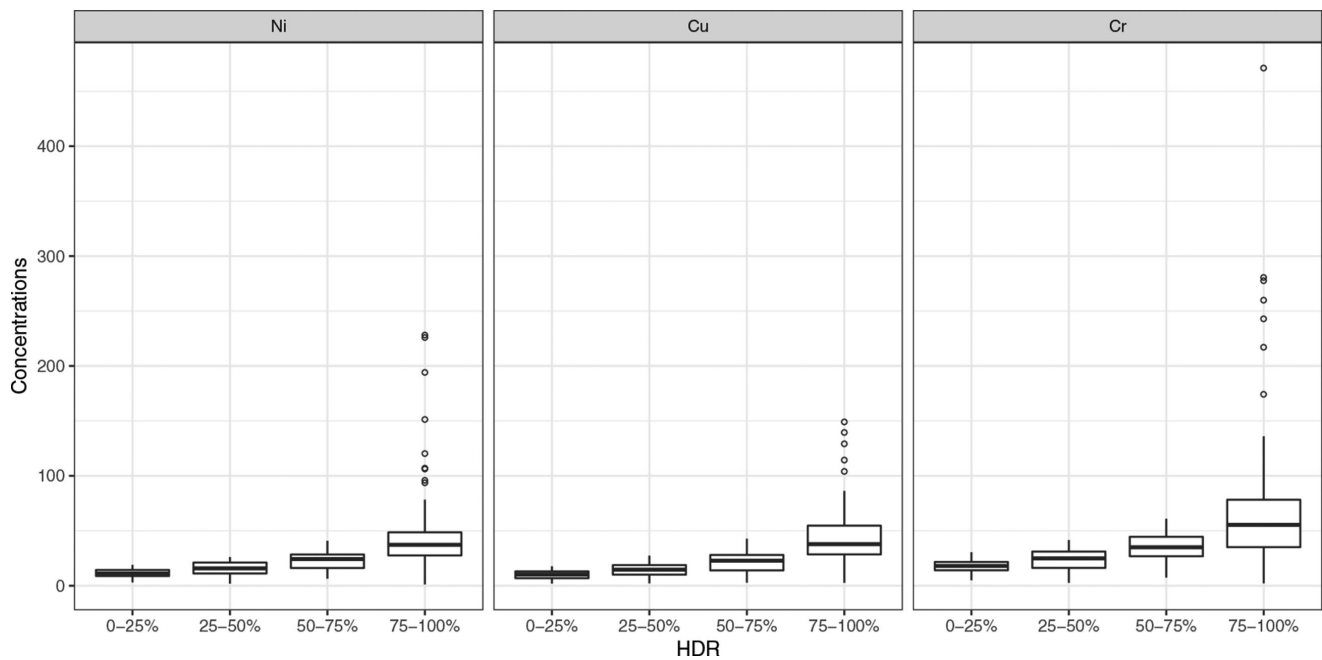
FIGURE 12 Boxplots for the distributions of heavy metals conditional on HDR regions from the GMDEB density estimate

The selected GMDEB model according to the BIC criterion is a two-component mixture model with variable volume and equal shape and orientation, that is, VEE in `mclust` nomenclature. The estimated vector of transformation parameters is $\hat{\lambda} = (-0.0384, 0.0010, -0.0975)$. Figure 11 contains the scatterplot matrix of heavy metal concentrations with the estimated density projected onto the marginal bivariate subspaces. The latter are shown as HDRs corresponding to 25%, 50%, and 75% probability regions. The distribution of nickel, copper, and chromium on the C-horizon layer is clearly skewed, with most sites having concentrations close to the origin. However, there are also a number of sites with relatively high concentrations. Further insights can be obtained by examining the distribution of metal concentrations conditional on the HDR to which the observed sites belong, as shown in Figure 12. Looking at the boxplots for the conditional distributions, we can see that the central part of the distribution, that is, that corresponding to 0–25% HDR, is characterized by the lowest concentration levels, whereas higher concentrations of heavy metals can be found as we move to regions of lower density.

6 | DISCUSSION

This paper addressed the problem of density estimation using GMMs when variables are partially or completely bounded. By introducing a range–power transformation of the data, it is possible to obtain a GMM for density estimation on the transformed data, and then to derive an accurate estimate of the density on the original scale, which takes into account the natural bounds of the variables. The proposed model is estimated by maximum likelihood using the EM algorithm. We showed that this transformation-based approach is able to deal with variables having either lower bounds or both lower and upper bounds, and the results obtained are often better than those provided by other methods usually based on modified versions of KDE.

The transformation-based approach seems to be very promising and, in principle, it could be applied to other types of non-Gaussian variables, for example, skewed variables, and for other purposes outside density estimation, for instance, in clustering. A straightforward extension to investigate is the use of other families of transformations, such as those proposed by Manly (1976) and Yeo and Johnson (2000). Furthermore, although the paper deals with the problem of density estimation, the proposed methodology has implications also on model-based clustering for bounded data. These very important issues are deferred to future works.

CONFLICT OF INTEREST

The author has declared no conflict of interest.

ORCID

Luca Scrucca (iD) https://orcid.org/0000-0003-3826-0484

REFERENCES

Banfield, J., & Raftery, A. E. (1993). Model-based Gaussian and non-Gaussian clustering. *Biometrics*, *49*, 803–821.

Benaglia, T., Chauveau, D., Hunter, D. R., & Young, D. (2009). mixtools: An R package for analyzing finite mixture models. *Journal of Statistical Software*, *32*(6), 1–29.

Biernacki, C., Celeux, G., & Govaert, G. (2000). Assessing a mixture model for clustering with the integrated completed likelihood. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *22*(7), 719–725.

Box, G., & Cox, D. (1964). An analysis of transformations. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *26*(2), 211–252.

Byrd, R. H., Lu, P., Nocedal, J., & Zhu C. (1995). A limited memory algorithm for bound constrained optimization. *SIAM Journal on Scientific Computing*, *16*(5), 1190–1208.

Celeux, G., & Govaert, G. (1995). Gaussian parsimonious clustering models. *Pattern Recognition*, *28*, 781–793.

Chen, S. X. (1999). Beta kernel estimators for density functions. *Computational Statistics & Data Analysis*, *31*(2), 131–145.

Crawford, S. L. (1994). An application of the Laplace method to finite mixture distributions. *Journal of the American Statistical Association*, *89*(425), 259–267.

Crawford, S. L., DeGroot, M. H., Kadane, J. B., & Small, M. J. (1992). Modeling lake-chemistry distributions: Approximate Bayesian methods for estimating a finite-mixture model. *Technometrics*, *34*(4), 441–453.

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

Escobar, M. D., & West, M. (1995). Bayesian density estimation and inference using mixtures. *Journal of the American Statistical Association*, *90*(430), 577–588.

Fraley, C., & Raftery, A. E. (1998). How many clusters? Which clustering method? Answers via model-based cluster analysis. *The Computer Journal*, *41*, 578–588.

Fraley, C., Raftery, A. E., & Scrucca, L. (2017). *mclust: Gaussian mixture modelling for model-based clustering, classification, and density estimation*. R package version 5.4.

Gasser, T., & Müller, H.-G. (1979). *Kernel estimation of regression functions*. Heidelberg: Springer.

Geenens, G. (2013). Probit transformation for kernel density estimation on the unit interval. *Journal of the American Statistical Association*, *109*(505), 346–358.

Grün, B., Kosmidis, I., & Zeileis, A. (2012). Extended beta regression in R: Shaken, stirred, mixed, and partitioned. *Journal of Statistical Software*, *48*(11), 1–25.

Hu, Y., & Scarrott, C. (2018). evmix: An R package for extreme value mixture modeling, threshold estimation and boundary corrected kernel density estimation. *Journal of Statistical Software*, *84*(5), 1–27.

Hyndman, R. J. (1996). Computing and graphing highest density regions. *The American Statistician*, *50*(2), 120–126.

Jones, M. C. (1993). Simple boundary correction for kernel density estimation. *Statistics and Computing*, *3*(3, 135–146.

Jones, M. C., & Henderson, D. A. (2007). Miscellanea kernel-type density estimation on the unit interval. *Biometrika*, *94*(4), 977–984.

Kooperberg, C. (2016). *logspline: Logspline Density Estimation Routines*. R package version 2.1.9.

Manly, B. F. J. (1976). Exponential data transformations. *The Statistician*, *25*(1), 37–42.

Marron, J. S., & Ruppert, D. (1994). Transformations to reduce boundary bias in kernel density estimation. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *56*, 653–671.

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

McLachlan, G. J., & Rathnayake, S. (2014). On the number of components in a Gaussian mixture model. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery*, *4*(5), 341–355.

Meng, X.-L., & Rubin, D. B. (1993). Maximum likelihood estimation via the ECM algorithm: A general framework. *Biometrika*, *80*(2), 267–278.

Nierenberg, D. W., Stukel, T. A., Baron, J. A., Dain, B. J., Greenberg, E. R., & Group, S. C. P. S. (1989). Determinants of plasma levels of beta-carotene and retinol. *American Journal of Epidemiology*, *130*(3), 511–521.

Reimann, C., Äyräs, M., Chekushin, V., Bogatyrev, I., Boyd, R., de Caritat, P., Volden, T., et al. (1998). *Environmental geochemical atlas of the Central Barents Region*. Trondheim, Norway: Geological Survey of Norway.

Reimann, C., Filzmoser, P., Garrett, R., & Dutter, R. (2011). *Statistical data analysis explained: Applied environmental statistics with R*. Chichester, UK: John Wiley & Sons.

Richardson, S., & Green, P. J. (1997). On Bayesian analysis of mixtures with an unknown number of components (with discussion). *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, *59*(4), 731–792.

Roeder, K., & Wasserman, L. (1997). Practical Bayesian density estimation using mixtures of normals. *Journal of the American Statistical Association*, *92*(439), 894–902.

Scarrott, C., Hu, Y., Akbar, A., & Canterbury, U. (2018). *evmix: Extreme value mixture modelling, threshold estimation and boundary corrected kernel density estimation*. R package version 2.10.

Schuster, E. F. (1985). Incorporating support constraints into nonparametric estimators of densities. *Communications in Statistics—Theory and Methods*, *14*(5), 1123–1136.

Schwartz, G. (1978). Estimating the dimension of a model. *Annals of Statistics*, *6*, 31–38.

Scott, D. W. (2009). *Multivariate density estimation: Theory, practice, and visualization* (2nd ed.). New Jersey: John Wiley & Sons.

Scrucca, L., Fop, M., Murphy, T. B., & Raftery, A. E. (2016). mclust 5: Clustering, classification and density estimation using Gaussian finite mixture models. *The R Journal*, *8*(1), 205–233.

Scrucca, L., & Raftery, A. E. (2015). Improved initialisation of model-based clustering using Gaussian hierarchical partitions. *Advances in Data Analysis and Classification*, *4*(9), 447–460.

Simonoff, J. (1996). *Smoothing methods in statistics*. New York: Springer.

Stone, C. J., Hansen, M. H., Kooperberg, C., & Truong, Y. K. (1997). Polynomial splines and their tensor products in extended linear modeling. *Annals of Statistics*, *25*(4, 1371–1470.

Titterington, D. M., Smith, A. F., & Makov, U. E. (1985). *Statistical analysis of finite mixture distributions*. Chichester, New York: Wiley.

Velilla, S. (1993). A note on the multivariate Box-Cox transformations to normality. *Statistics and Probability Letters*, *17*, 441–451.

Wand, M. P., Marron, J. S., & Ruppert, D. (1991). Transformations in density estimation. *Journal of the American Statistical Association*, *86*(414), 343–353.

Wu, C. J. (1983). On the convergence properties of the EM algorithm. *The Annals of Statistics*, *11*(1), 95–103.

Yeo, I.-K., & Johnson, R. A. (2000). A new family of power transformations to improve normality or symmetry. *Biometrika*, *87*(4, 954–959.

Young, D., Benaglia, T., Chauveau, D., & Hunter, D. (2017). *mixtools: Tools for Analyzing Finite Mixture Models*. R package version 1.1.0.

Zhu, X., & Melnykov, V. (2018). Manly transformation in finite mixture modeling. *Computational Statistics & Data Analysis*, *121*, 190–208.

SUPPORTING INFORMATION

Additional supporting information including source code to reproduce the results may be found online in the Supporting Information section at the end of the article.