APPROXIMATION OF DENSITY FUNCTIONS USING SIMPLICIAL SPLINES

J. Machalová, R. Talská, K. Hron
Department of Mathematical Analysis and Applications of Mathematics,
Faculty of Science, Palacký University Olomouc
Olomouc, CZECH REPUBLIC
e-mail: jitka.machalova@upol.cz

Abstract

Probability density functions result in practice frequently from aggregation of massive data and their further statistical processing is thus of increasing importance. However, specific properties of density functions prevent from analyzing a sample of densities directly using tools of functional data analysis. Moreover, it is not only about the unit integral constraint, which results from representation of densities within the equivalence class of proportional positive-valued functions, but also about their relative scale which emphasizes the effect of small relative contributions of Borel subsets to the overall measure of the support. For practical data processing, it is popular to approximate first the input (discrete) data with a proper spline representation. Aim of the contribution is to introduce new class of B-splines within the Bayes space methodology which is suitable for representation of density functions. Accordingly, the original densities are expressed as real functions using the centred logratio transformation and optimal smoothing splines with B-spline basis honoring the resulting zero-integral constraint are developed.

Keywords: data science, simplicial spline, density function

1 Introduction

Probability density functions are non-negative functions, popularly represented with a unit integral constraint. However, in some fields, e.g., in Bayesian statistics density functions are considered in a more general setting, where any representation within the equivalence class of proportional functions can be taken. This reflects better the basic property of densities - their scale invariance. Accordingly, the sample space of densities is formed by a set of equivalence classes of proportional positive functions. In this paper a bounded support $I = [a, b] \subset \mathbb{R}$ of densities is considered which occurs frequently in practice. Specific properties of density functions are captured by the Bayes space $\mathcal{B}^2(I)$ of functions with square-integrable logarithm [2, 5]; in a default setting the Lebesgue reference measure is taken. The Bayes space $\mathcal{B}^2(I)$ has structure of separable Hilbert space which enables to construct an isometric isomorphism between $\mathcal{B}^2(I)$ and $L^2(I)$, the L^2 space restricted to I. An isometric isomorphism between $\mathcal{B}^2(I)$ and $L^2(I)$ is represented by the *centred log-ratio* (clr) transformation [2]. It is defined for a density $f \in \mathcal{B}^2(I)$ as

$$\mathrm{clr}(f)(x) = f_c(x) = \ln f(x) - \frac{1}{\eta} \int_I \ln f(y)\, \mathrm{d}y,$$

with $\eta = b - a$. The clr transformation induces an additional zero-integral constraint that needs to be taken into account for computation and analysis on clr-transformed density functions. As the clr space is clearly a subspace of $L^2(I)$, hereafter it is denoted as $L_0^2(I)$. Although the clr transformed densities are standard real functions, their constrained character calls for modification of methods for their approximation and further statistical processing using methods of functional data analysis. This is also the case of approximation using splines, described in a detail in the next section.

2 Optimal smoothing splines in $L^2(I)$

Firstly we recall the basic knowledge about B-spline representation of splines, see [3, 4, 12]. Let $\mathcal{S}_k^{\Delta\lambda}[a,b]$ denote the vector space of polynomial splines of degree $k > 0$, defined on a finite interval $I = [a,b]$ with the sequence of knots $\Delta\lambda$, where

$$\Delta\lambda := \lambda_0 = a < \lambda_1 < \ldots < \lambda_g < b = \lambda_{g+1}.$$

It is known that $dim\left(\mathcal{S}_k^{\Delta\lambda}[a,b]\right) = g + k + 1$. Then every spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ in $L^2(I)$ has a unique representation

$$s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x).$$

For this representation it is necessary to add some additional knots, e.g. such that

$$\lambda_{-k} = \cdots = \lambda_{-1} = \lambda_0, \quad \lambda_{g+1} = \lambda_{g+2} = \cdots = \lambda_{g+k+1}. \tag{1}$$

Vector $\mathbf{b} = (b_{-k}, \ldots, b_g)^\top$ is called *the vector of B-spline coefficients* of $s_k(x)$, functions $B_i^{k+1}(x)$, $i = -k, \ldots, g$ are *B-splines of degree k* and form basis in $\mathcal{S}_k^{\Delta\lambda}[a,b]$. In matrix notation it can be written as

$$s_k(x) = \mathbf{C}_{k+1}(x)\mathbf{b},$$

where $\mathbf{C}_{k+1}(x) = \left(B_i^{k+1}(x)\right)_{i=-k}^{g}$ is so called *collocation matrix*. It is known that derivative of order l, $l \in \{1, \ldots, k-1\}$, of the spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ is a spline $s_{k-l}(x) \in \mathcal{S}_{k-l}^{\Delta\lambda}[a,b]$ with the same knots. Using properties of B-splines the spline derivatives can be written in matrix notation as

$$s_k^{(l)}(x) = \mathbf{C}_{k+1-l}(x)\mathbf{b}^{(l)},$$

where $\mathbf{b}^{(l)} \in \mathbb{R}^{g+k+1-l}$ is given by $\mathbf{b}^{(l)} = \mathbf{D}_l\mathbf{L}_l\mathbf{b}^{(l-1)} = \mathbf{D}_l\mathbf{L}_l \ldots \mathbf{D}_1\mathbf{L}_1\mathbf{b} = \mathbf{S}_l\mathbf{b}$ and $\mathbf{b}^{(0)} = \mathbf{b}$. Upper triangular matrix $\mathbf{S}_l = \mathbf{D}_l\mathbf{L}_l \ldots \mathbf{D}_1\mathbf{L}_1 \in \mathbb{R}^{g+k+1-l,g+k+1}$ has full row rank. Matrix $\mathbf{D}_j \in \mathbb{R}^{g+k+1-j,g+k+1-j}$ is diagonal such that

$$\mathbf{D}_j = (k+1-j)\, diag\left(d_{-k+j}, \ldots, d_g\right), \qquad d_i = \frac{1}{\lambda_{i+k+1-j} - \lambda_i} \quad \forall i = -k+j, \ldots, g$$

and

$$\mathbf{L}_j := \begin{pmatrix} -1 & 1 & & \\ & \ddots & \ddots & \\ & & -1 & 1 \end{pmatrix} \in \mathbb{R}^{g+k+1-j,g+k+2-j}.$$

Now we assume that data (x_i, y_i), $a \le x_i \le b$, weights $w_i \ge 0$, $i = 1, \ldots, n$, sequence of knots $\Delta\lambda$, $n \ge g + 1$, and a parameter $\alpha \in (0, 1)$ are given. The optimal smoothing problem, [9, 10], which is in fact generalization of smoothing problem [3, 4], is defined as a task to find a spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$, which minimizes the functional

$$J_l(s_k) \;=\; \int_a^b \left[s_k^{(l)}(x)\right]^2 \mathrm{d}x + \alpha \sum_{i=1}^{n} w_i \left[y_i - s_k(x_i)\right]^2. \tag{2}$$

The choice of parameter l will affect smoothness of the resulting spline. Let us denote $\mathbf{x} = (x_1, \ldots, x_n)^\top$, $\mathbf{y} = (y_1, \ldots, y_n)^\top$, $\mathbf{w} = (w_1, \ldots, w_n)^\top$ and $\mathbf{W} = diag\,(\mathbf{w})$. The functional $J_l(s_k)$ can be written in a matrix form as

$$J_l(\mathbf{b}) \;=\; \mathbf{b}^\top \mathbf{N}_{kl} \mathbf{b} + \alpha \left[\mathbf{y} - \mathbf{C}_{k+1}(\mathbf{x})\mathbf{b}\right]^\top \mathbf{W} \left[\mathbf{y} - \mathbf{C}_{k+1}(\mathbf{x})\mathbf{b}\right],$$

see [9, 10] for details. The matrix $\mathbf{N}_{kl} = \mathbf{S}_l^\top \mathbf{M}_{kl} \mathbf{S}_l$ is positive semidefinite, where

$$\mathbf{M}_{kl} = \begin{pmatrix} \left(B_{-k+l}^{k+1-l}, B_{-k+l}^{k+1-l}\right) & \ldots & \left(B_{g}^{k+1-l}, B_{-k+l}^{k+1-l}\right) \\ \vdots & & \vdots \\ \left(B_{-k+l}^{k+1-l}, B_{g}^{k+1-l}\right) & \ldots & \left(B_{g}^{k+1-l}, B_{g}^{k+1-l}\right) \end{pmatrix} \in \mathbb{R}^{g+k+1-l, g+k+1-l}$$

and

$$\left(B_i^{k+1-l}, B_j^{k+1-l}\right) = \int_a^b B_i^{k+1-l}(x) B_j^{k+1-l}(x)\,\mathrm{d}x$$

stands for scalar product of B-splines in $L^2(I)$ space. Matrix $\mathbf{M}_{kl}$ is positive definite, because $B_i^{k+1-l}(x) \ge 0$, $i = -k + l, \ldots, g$ are basis functions. Now the task is to find a minimum of function $J_l(\mathbf{b})$. It is obvious that this minimum fulfils the condition

$$\frac{\partial J_l(\mathbf{b})}{\partial \mathbf{b}^\top} = 0,$$

which can be written as a system of linear equations $\mathbf{G}\mathbf{b} = \mathbf{g}$ with

$$\mathbf{G} \;=\; \alpha^{-1}\mathbf{N}_{kl} + \mathbf{C}_{k+1}^\top(\mathbf{x})\mathbf{W}\mathbf{C}_{k+1}(\mathbf{x}), \qquad \mathbf{g} \;=\; \mathbf{C}_{k+1}^\top(\mathbf{x})\mathbf{W}\mathbf{y}.$$

If this system is consistent, then there exists a solution which is given by $\mathbf{b}^* = \mathbf{G}^-\mathbf{g}$, see [10]. So finally $s_k^*(x) = \mathbf{C}_{k+1}(x)\mathbf{b}^*$ is resulting optimal smoothing spline, i.e. spline which minimizes functional (2).

3 Optimal smoothing splines in $L_0^2(I)$

In this section the case of smoothing clr-transformed density functions is considered. The task is find spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a, b]$ which minimizes functional (2) and which satisfies an additional condition

$$\int_a^b s_k(x)\,\mathrm{d}x \;=\; 0. \tag{3}$$

There are two possibilities how to deal with this problem. The first approach, which was published in [8], is based on using expression between coefficients of spline and its derivative. The second possibility uses new B-spline basis functions which satisfy the condition (3). This process is described in [7].

Now the first approach will be described in more details. Note that the spline

$$s_k(x) = \sum_{i=-k}^{g} b_i B_i^{k+1}(x)$$

is a derivative of spline

$$s_{k+1}(x) = \sum_{i=-k-1}^{g} c_i B_i^{k+2}(x), \tag{4}$$

if

$$b_i = (k+1)\,\frac{c_i - c_{i-1}}{\lambda_{i+k+1} - \lambda_i} \quad \forall i = -k, \ldots, g. \tag{5}$$

For each spline $s_k(x) \in \mathcal{S}_k^{\Delta\lambda}[a,b]$ satisfying the condition (3) we have

$$0 = \int_a^b s_k(x)\,\mathrm{d}x = \left[s_{k+1}\,(x)\right]_a^b = s_{k+1}(\lambda_{g+1}) - s_{k+1}(\lambda_0),$$

because $a = \lambda_0$, $b = \lambda_{g+1}$. With respect to the definition and properties of B-splines, the additional knots (1) and notation (4) we get

$$0 = s_{k+1}(\lambda_{g+1}) - s_{k+1}(\lambda_0) = c_g - c_{-k-1},$$

so that $c_{-k-1} = c_g$. The relationship (5) between the vector $\mathbf{b} = (b_{-k}, \ldots, b_g)^\top$ of B-spline coefficients of $s_k(x)$ and the vector $\mathbf{c} = (c_{-k-1}, \ldots, c_g)^\top$ of $s_{k+1}(x)$, $\mathbf{c} \in \mathbb{R}^{g+k+2}$ such that $c_{-k-1} = c_g$, can be written as

$$\mathbf{b} = \mathbf{D}\mathbf{K}\bar{\mathbf{c}},$$

where $\bar{\mathbf{c}} = (c_{-k}, \ldots, c_g)^\top \in \mathbb{R}^{g+k+1}$. Matrices $\mathbf{D}$ and $\mathbf{K}$ are known, see [8]. So with this relationship we are able to rewrite function $J_l(\mathbf{b})$ as a function $J_l(\bar{\mathbf{c}})$. Then we find its minimum $\bar{\mathbf{c}}^*$ and finally the vector of B-spline coefficients $\mathbf{b}^*$ for optimal smoothing spline which has zero integral is obtained by

$$\mathbf{b}^* = \mathbf{D}\mathbf{K}\bar{\mathbf{c}}^*.$$

The corresponding spline is given by $s_k^*(x) = \mathbf{C}_{k+1}(x)\mathbf{b}^*$.

The second approach for finding optimal smoothing spline with zero integral, which is presented in [7], uses new functions $Z_i^{k+1}(x)$ for $k \geq 0$. They are defined by formula

$$Z_i^{k+1}(x) := \frac{\mathrm{d}}{\mathrm{d}x} B_i^{k+2}(x).$$

More precisely for $k = 0$

$$Z_i^1(x) = \begin{cases} 1 & \text{if } x \in [\lambda_i, \lambda_{i+1}) \\ -1 & \text{if } x \in (\lambda_{i+1}, \lambda_{i+2}] \end{cases} \tag{6}$$

and for $k \geq 1$

$$Z_i^{k+1}(x) = (k+1)\left(\frac{B_i^{k+1}(x)}{\lambda_{i+k+1} - \lambda_i} - \frac{B_{i+1}^{k+1}(x)}{\lambda_{i+k+2} - \lambda_{i+1}}\right). \tag{7}$$

Noteworthy, functions $Z_i^{k+1}(x)$ have similar properties as B-splines $B_i^{k+1}(x)$, we called them ZB-splines, for more details see [7]. Example of quadratic ZB-splines $Z_i^3(x)$ is displayed in Figure 1.

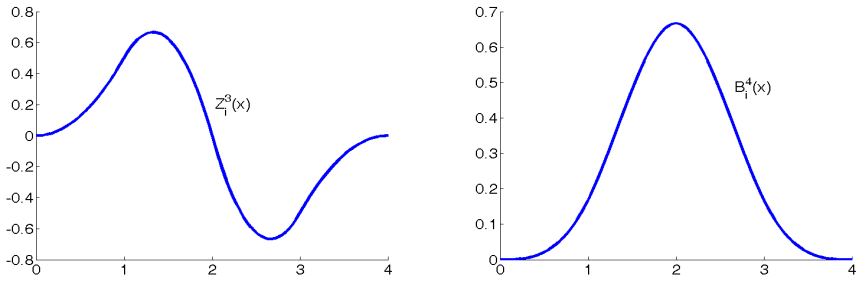

Figure 1: Quadratic ZB-spline $Z_i^3(x) = \dfrac{\mathrm{d}}{\mathrm{d}x} B_i^4(x)$ with equidistant knots $0, 1, 2, 3, 4$.

From the perspective of $L_0^2(I)$ a crucial point is that integral of $Z_i^{k+1}(x)$ equals to zero:

$$\int_{-\infty}^{+\infty} Z_i^{k+1}(x)\,\mathrm{d}x = \int_{\lambda_i}^{\lambda_{i+k+2}} Z_i^{k+1}(x)\,\mathrm{d}x = \int_{\lambda_i}^{\lambda_{i+k+2}} \frac{\mathrm{d}}{\mathrm{d}x} B_i^{k+2}(x)\,\mathrm{d}x =$$
$$= \left[B_i^{k+2}(x)\right]_{\lambda_i}^{\lambda_{i+k+2}} = 0.$$

In the following, $\mathcal{Z}_k^{\Delta\lambda}[a, b]$ denotes the vector space of polynomial splines of degree $k > 0$, defined on a finite interval $[a, b]$ with the sequence of knots $\Delta\lambda$ and having zero integral on $[a, b]$. With respect to the condition of the zero integral it is clear that $dim\left(\mathcal{Z}_k^{\Delta\lambda}[a, b]\right) = g + k$, for more details see [7]. With the additional knots (1) we can construct $g + k$ functions $Z_{-k}^{k+1}(x), \cdots, Z_{g-1}^{k+1}(x)$, which are basis functions of the space $\mathcal{Z}_k^{\Delta\lambda}[a, b]$. Then every spline $s_k(x) \in \mathcal{Z}_k^{\Delta\lambda}[a, b]$ has a unique representation

$$s_k(x) = \sum_{i=-k}^{g-1} z_i Z_i^{k+1}(x).$$

In matrix notation it can be expressed as

$$s_k(x) = \mathbf{Z}_{k+1}(x)\,\mathbf{z},$$

where $\mathbf{Z}_{k+1} = \left(Z_i^{k+1}(x)\right)_{i=-k}^{g-1}$, $\mathbf{z} = (z_{-k}, \ldots, z_{g-1})^\top$. Next steps are similar as we used in the first approach, the functional (2) is expressed as a function of variable $\mathbf{z}$. Then we find its minimum $\mathbf{z}^*$ and finally the optimal smoothing spline with zero integral is given by formula $s_k^*(x) = \mathbf{Z}_{k+1}(x)\mathbf{z}^*$.

Reduction of dimension for splines in $L_0^2(I)$ by one is a very natural consequence of clr transformation of density functions. Note that this feature is present also for clr coefficients of compositional data [1].

4 Simplicial splines in the Bayes space

Construction of splines directly in $L_0^2(I)$ has important practical consequences, however, it is important also from theoretical perspective. Expressing B-splines as functions in $L_0^2(I)$ enables to back-transform them to the original Bayes space $\mathcal{B}^2(I)$ using inverse clr transformation [2]. It results in *simplicial B-splines (SB-splines)*, obtained from (6), (7) as

$$\zeta_i^{k+1}(x) = \frac{\exp[Z_i^{k+1}(x)]}{\int_I \exp[Z_i^{k+1}(y)]\,\mathrm{d}y}, \quad i = -k, \ldots, g-1,\ k \geq 0.$$

Note that SB-splines $\zeta_i^{k+1}(x)$ fulfill the unit integral constraint. As a consequence, it is immediate to define vector space $\mathcal{C}_k^{\Delta\lambda}[a,b]$ of simplicial polynomial splines of degree $k > 0$, defined on a finite interval $[a,b]$ with the sequence of knots $\Delta\lambda$. From isomorphism between $\mathcal{C}_k^{\Delta\lambda}[a,b]$ and $\mathcal{Z}_k^{\Delta\lambda}[a,b]$ it holds that $dim\left(\mathcal{C}_k^{\Delta\lambda}[a,b]\right) = g + k$. Moreover, from isometric properties of clr transformation it follows that every simplicial spline $\xi_k(x) \in \mathcal{C}_k^{\Delta\lambda}[a,b]$ in $B^2(I)$ can be uniquely represented as

$$\xi_k(x) = \bigoplus_{i=-k}^{g-1} c_i \odot \zeta_i^{k+1}(x),$$

where $\odot$ stands for powering operation in $\mathcal{B}^2(I)$ [2]

The resulting simplicial splines can be used for representation of densities directly in $\mathcal{B}^2(I)$. This is an important step in construction of methods of functional data analysis involving density functions, like for ANOVA modeling or for the SFPCA method.

5 Outlook

Once the sample of probability density functions is approximated using optimal smoothing splines in $L_0(I)$, any from popular methods of functional data analysis [11] can be applied by considering the zero integral constraint of the clr transformed densities. These methods usually strongly rely just on a proper spline representation of densities. Accordingly, simplicial functional principal component analysis [6] or compositional regression with functional response [13] were developed; they show a clear way how also other methods could be adapted for this important class of functions.

Acknowledgments

The authors gratefully acknowledge both the support by the grant IGA PrF IGA_PrF_2019_006, Mathematical Models of the Internal Grant Agency of the Palacký University in Olomouc.

References

[1] Aitchison J. (1986). *The statistical analysis of compositional data*. Chapman and Hall, London.

[2] Van den Boogaart K.G., Egozcue J.J., Pawlowsky-Glahn V. (2014). Bayes Hilbert spaces *Australian & New Zealand Journal of Statistics*. Vol. **56**, Num. **2**, pp. 171-194.

[3] De Boor C. (1978). *A practical guide to splines*. Springer-Verlag, New York.

[4] Dierckx P. (1993). *Curve and surface fitting with splines*. Oxford University Press, New York.

[5] Egozcue J.J., Díaz-Barrero J.L., Pawlowsky-Glahn V. (2006). Hilbert space of probability density functions based on Aitchison geometry, *Acta Mathematica Sinica*. Vol. **22**, Num. **4**, pp. 1175-1182.

[6] Hron K., Menafoglio A., Templ M., Hrůzová K., Filzmoser P. (2016). Simplicial principal component analysis for density functions in Bayes spaces. . *Computational Statistics and Data Analysis*. Vol. **94**, pp. 330-350.

[7] Machalová J., Talská R., Hron K., Gába A. (2019). Simplicial splines for representation of density functions. *arXiv preprint arXiv:1905.06858*.

[8] Machalová J., Hron K., Monti G. S. (2016). Preprocessing of centred logratio transformed density functions using smoothing splines. *Journal of Applied Statistics*. Vol. **43**, Num. **8**, pp. 1419-1435.

[9] Machalová J. (2002). Optimal interpolatory splines using B-spline representation. *Acta Universitatis Palackianae Olomucensis. Facultas Rerum Naturalium. Mathematica*. Palacký University Olomouc. Vol. **41**, Num. **1**, pp. 105-118.

[10] Machalová J. (2002). Optimal interpolating and optimal smoothing spline. *Journal of Electrical Engineering*. Vol. **53**, pp. 79-82.

[11] Ramsay J., Silverman B.W. (2005). *Functional Data Analysis*, second ed. Springer, New York.

[12] Schumaker L. (2007). *Spline functions: basic theory*. Cambridge University Press.

[13] Talská R., Menafoglio A., Machalová J., Hron K., Fišerová E. (2018). Compositional regression with functional response. *Computational Statistics & Data Analysis*. Vol. **123**, pp. 66-85.