Updating on the Kernel Density Estimation for Compositional Data

Martín-Fernández, J. A.[1], Chacón-Durán, J. E.[2], and Mateu-Figueras, G.[1]

[1] Dpt. Informàtica i Matemàtica Aplicada, Universitat de Girona, Campus Montilivi, Edifici P-IV. E-17071, Girona, Spain, josepantoni.martin@udg.es
[2] Dpto. Matemáticas, Universidad de Extremadura, Escuela Politécnica, E-10071, Cáceres, Spain

Summary. Existing kernels for compositional data cannot apply the common simplifications of the bandwidth matrix. In this work new kernel density estimation methods are proposed. These methods incorporate recent advances from log-ratio methodology and bandwidth matrix selection theory. We present examples where the behaviour of the proposed approach is illustrated.

Key words: Dirichlet distribution, isometric log-ratio transformation, simplex

1 Introduction

The sample space of compositional data ([Ait86]) is the simplex S^D defined as $S^D = \{x = [x_1, ..., x_D] : x_i > 0, x_1 + ... + x_D = c\}$, where c can be 1, 100, 10^6 or any other constant depending on the units of measurement. As it was stated during the last Compositional Data Workshop ([MB05]) this kind of data appears in many disciplines: Archaeometry, Geology, Economy, Biomedical Research, and Space Research. Compositional observations are proportions of some whole and therefore positive and of constant sum. After [Ait86] and with general agreement, one accepts that compositional data reflect only relative magnitude, and thus interest lies in relative - and not absolute - changes. The *key* question here is which metric could be appropriate for this kind of data. The usual Euclidean metric measures absolute changes, whereas relative changes can be measured using some logarithmic scale ([ABMP00]).

In [AL85] two multivariate kernel methods of density estimation for compositional data are introduced: Dirichlet and additive logistic-normal (aln). The authors recommend selecting the aln kernel rather than the Dirichlet except "if there is the least suspicion of sparseness in the data". In addition, when the data set has observations with zero values they recommend replace in advance the rounding zeros by a small value. As is stated by the authors, since the aln kernel is based on the additive log-ratio (alr) transformation we cannot use the common simplifications of the *bandwidth matrix* ([WJ95], p. 91). However, nowadays we can take advantage of

new, just developed, full bandwidth matrix selection methods ([DH03],[DH05]) that were not available when the paper of [AL85] was published.

In this work new kernel density estimation methods for compositional data are proposed. These methods incorporate recent advances from log-ratio methodology and bandwidth matrix selection theory. We first introduce basic concepts related to the vector space structure of the simplex and the ilr transformation. Proceed afterwards with the new proposals for the kernel density estimation methods. Finally, we present examples where the behaviour of the proposed approach is illustrated.

2 Recent developments on compositional data analysis

The *closure operator* $\mathcal{C}$ is defined by $\mathcal{C}(w) = [w_1/\sum w_j, \ldots, w_D/\sum w_j]$, where $w \in R_+^D$. The *perturbation operation*, $x \oplus x^* = \mathcal{C}\,[x_1 x_1^*, \ldots, x_D x_D^*]$, defined on $\mathcal{S}^D \times \mathcal{S}^D$, and the *power transformation*, $\alpha \otimes x = \mathcal{C}[x_1^\alpha, \ldots, x_D^\alpha]$, defined on $\mathrm{R} \times \mathcal{S}^\mathrm{D}$ induce a vector space structure in the simplex. Then, the *perturbation difference*

$$x \oplus x^{*-1} = \mathcal{C}\,[x_1/x_1^*, \ldots, x_D/x_D^*]$$

allows to introduce the centring operation. To better understand the perturbation difference operation it is useful to establish a parallelism with the vector subtraction operation of the real space. An interesting property is that the neutral element is the composition $e = [1/D, ..., 1/D]$ which is the geometric centre of the simplex and which has same role than the origin of coordinate axes in real space.

In [Ait86] the additive log-ratio (alr) and the centered log-ratio (clr) transformations are introduced:

$$\mathrm{alr}(x) = \big[\ln(x_1/x_D), \ldots, \ln(x_{D-1}/x_D)\big],$$

$$\mathrm{clr}(x) = \big[\ln(x_1/g(x)), \ldots, \ln(x_D/g(x))\big],$$

where $g(x) = (x_1 \cdots x_D)^{1/D}$ stands for the geometric mean of the composition x. One must proceed with caution when the alr transformation is applied because is asymmetric in the components. We must verify that our statistical technique is invariant under permutations of the components. On the other hand the clr transformation has the weakness that the covariance matrix of the transformed data set is singular. In the literature the alr transformation is mainly applied in parametric contexts and the clr transformation is mostly used in nonparametric studies. To avoid above difficulties the isometric log-ratio transformation (ilr) is introduced in [EPMB03]:

$$\mathrm{ilr}(x) = y = [y_1, \ldots, y_{D-1}] \in R^{D-1}, \text{ where } y_i = \frac{1}{\sqrt{i(i+1)}} \ln\left(\frac{\prod_{j=1}^{i} x_j}{(x_{i+1})^i}\right).$$

The alr and clr transformations are more interpretable than the ilr. Nevertheless, the ilr transformation is useful when some technique as the *transformation method* ([BA97], p. 14) is applied. This method is recommended for a data with bounded support, for example the simplex. In essence, the transformation method consists of estimating the density of the transformed data and then transform back to the original space.

Compositional variables frequently take null values and dealing with log-ratios excludes dealing with zeros. Therefore, a strategy is needed on how to deal with zeros in a given data set. In [AL85] (p. 133) a general procedure for replacing the zero values in a composition is recommended. Recently, a new replacement method has overcome this procedure ([MBP03]): consider a composition $x = [x_1, \ldots, x_D] \in S^D$ containing rounded zeros. Then, x can be replaced by a new composition $r = [r_1, \ldots, r_D] \in S^D$ without zeros according to the following replacement rule:

$$r_j = \begin{cases} \delta_j & \text{if } x_j = 0, \\ \left(1 - \frac{\sum_{\{x_k=0\}} \delta_k}{c}\right) x_j & \text{if } x_j > 0, \end{cases}$$

where δ_j is a small value less than a given threshold for the component x_j.

3 New strategies for compositional kernels

Following [AL85] the Dirichlet class $\Delta^{D-1}(\alpha)$ with density function

$$\Delta(x|\alpha) = \frac{\Gamma(\alpha_1 + \ldots + \alpha_D)}{\Gamma(\alpha_1) \ldots \Gamma(\alpha_D)} x_1^{\alpha_1 - 1} \ldots x_D^{\alpha_D - 1},$$

can be considered for defining a kernel $K(x|X, h)$ for $x \in S^D$ *centered* at $X \in S^D$, where h is the smoothing factor. The *key* question here is the centering operation. In [AL85] the authors suggest taking the kernel $K(x|X, h) = \Delta(x|j + (1/h)X)$ where j is the D-vector of units. It is easy to proof that for $\alpha = j + (1/h)X$ the mode of the Dirichlet distribution is at X. Nevertheless, it is also clear that for this α the concentration of the distribution about the mode depends on the values of X. In essence, this fact causes the different behaviour between the Dirichlet and the aln kernels in relation to the sparseness in the data. This different behaviour was stated in [AL85] from an empirical point of view but without corresponding theoretical support. In order to avoid this effect we propose a different strategy. In real space it is natural to adopt a kernel centred on the origin which evaluates the vector difference $x - X$. Analogue strategy for compositional data consists of adopting a kernel centred on the geometric centre of the simplex $e = [1/D, \ldots, 1/D]$ and such that evaluates the perturbation difference $x \oplus X^{-1}$. To achieve this we propose taking the kernel

$$K(x|X, h) = \Delta(x \oplus X^{-1}|\frac{1}{hD}j),$$

where j is the D-vector of units. Here the smoothing factor h is directly related with the concentration of the distribution because the larger h is, the less concentrated the distribution about the mode.

Actually, most of the researchers deal with kernels based on the standard multivariate normal density in real space. In addition, the most relevant results modelling compositional data analysis have been achieved assuming normal distribution for the transformed data. In this work we focus our attention to the transformation method. Future research will analyze the behaviour of the Dirichlet kernel, and the modified kernel method suggested in [BA97] (p. 15).

Following analogue strategy as for the additive logistic-normal (aln) class of distributions ([AL85], p. 130) the isometric logistic-normal (iln) class $\mathrm{I}^{D-1}(\xi, \Psi)$ ([Mat03]) can be defined with density function

$$\mathrm{I}^{D-1}(\xi, \Psi) = \frac{1}{\sqrt{D}\ x_1 \dots x_D}\ \phi(y|\xi, \Psi),$$

where $\phi(y|\xi, \Psi)$ is the density of the normal $\mathrm{N}^{D-1}(\xi, \Psi)$ distribution evaluated at $y = \mathrm{ilr(x)}$, the isometric log-ratio transformed vector. The transformation method ([BA97], p. 14) is considered in combination with the iln class and we propose the *iln kernel* on S^D defined by

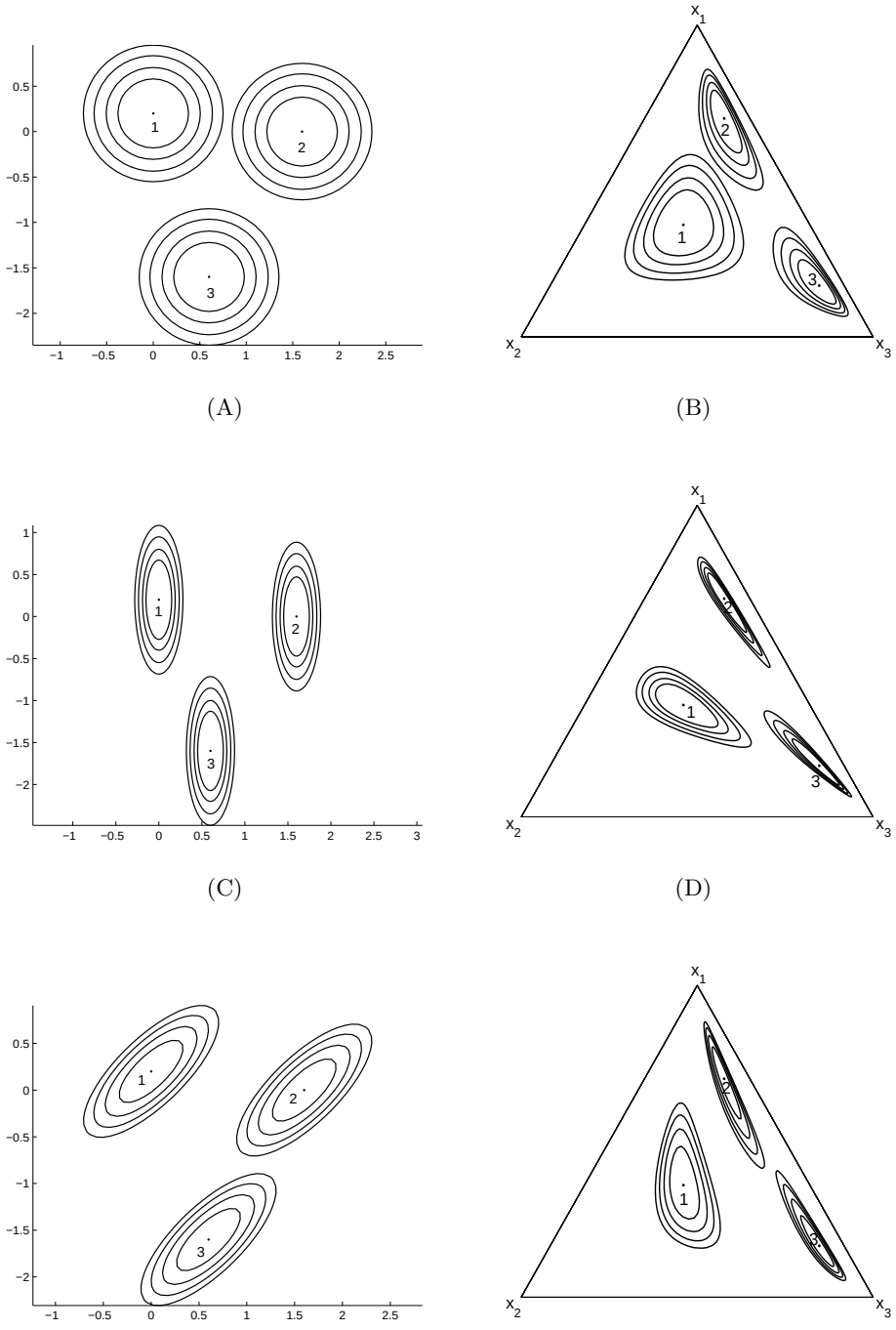
$$K(x|X, H) = \frac{1}{\sqrt{D}\ x_1 \dots x_D}\ \phi(y|Y, H),$$

where $Y = \mathrm{ilr(X)}$ and H is the bandwidth matrix. This iln kernel is equivalent to ilr-transform the compositional data set; then, obtain density estimates in R^{D-1} using a multivariate kernel $\phi(y|Y, H)$ with some suitable bandwidth matrix H; and finally, transform back using the ilr-inverse transformation to the simplex. In relation to the centring operation, it is easy to state that the iln kernel verifies $K(x|X, H) = K(x \oplus X^{-1}|e, H)$. Note that we have an analogue property in real space in relation to the vector subtraction operation.

In [AL85] the authors conclude that the common simplifications of the bandwidth matrix H are impossible when the aln kernel is used. The problem is that the results are not invariant under permutations of the components. One only can work with a bandwidth matrix proportional to the sample covariance matrix of the alr-transformed data set. In real space, [WJ95] (p. 106) state that this bandwidth matrix is appropriate in the case of multivariate normal data and not for general density shapes. Consequently, the aln kernel is appropriate in the case of aln class and not for other density shapes. With our strategy, using the iln kernel, all parameterisations of the bandwidth matrix H are feasible.

4 Examples

In [WJ95] (p. 92) the authors consider $\mathcal{F}$ the class of symmetric, positive definite $(D-1)\times(D-1)$ matrices; $\mathcal{D}$ the subclass of diagonal positive definite matrices; and $\mathcal{S}$ the subclass $\{h^2 Id : h > 0\}$, where Id is the identity matrix. In our strategy, if $H \in \mathcal{S}$ then $H = h^2 Id$ and we are working with circles in the ilr-transformed space. If we take $H \in \mathcal{D}$ then $H = diag(h_1^2, \dots, h_{D-1}^2)$ and our kernels are ellipses such that their axes are parallel to the coordinate directions in the ilr-transformed space. For the full bandwidth matrix class $\mathcal{F}$ the axes of the ellipses in the ilr-transformed space are not parallel to the coordinate axes. In order to illustrate the behaviour of the iln kernel some simple cases for the three different parameterisations of the bandwidth are plotted. Figure 1 respectively shows the contour plots for $H \in \mathcal{S}$, $H \in \mathcal{D}$ and $H \in \mathcal{F}$. Figure 1A corresponds to the circles in the ilr-transformed space and Figure 1B shows these contours plots transformed back in the simplex S^3. Figures 1C and 1D show the contour plots for $H \in \mathcal{D}$. Observe that the axes of the ellipses are parallel to the coordinate axes. Figures 1E and 1F show the contour plots for $H \in \mathcal{F}$. In this case - full bandwidth matrix - the ellipses have their axes non parallel to the coordinate axes.



(A) (B)

(C) (D)

(E) (F)

Fig. 1. Contours plots of iln kernels for $D = 3$: (Left) in the ilr-space; (Right) in the Ternary diagram.

Having proposed the iln kernel we analyse its application to the three-composition AFM of 23 aphyric Skye lavas ([AL85]; [Ait86], p. 360). The variables A, F, M respectively stand for the relative proportions of alkali or Na_2+K_2O, Fe_2O_3, and MgO. Figure 2A shows the data set Skye Lavas in ternary diagram. In Figure 2B the corresponding ilr-transformed data set is plotted. Observe that the points (Fig. 2B) show a linear pattern in the ilr-space and this pattern is not parallel to the coordinate axes. This pattern appears in the ternary (Fig. 2A) as a curved line. Considering different parameterisations of the bandwidth matrix, two different methods for selecting the bandwidth are applied: 2-stage Plug-In (PI) ([DH03]) and Least Squares Cross Validation (LSCV) ([DH05]). For the sake of an easier readability we do not reproduce here all the results. We present here (Figures 2C, 2D and 3) only the results produced by the LSCV method for a diagonal bandwidth matrix $H \in \mathcal{D}$.

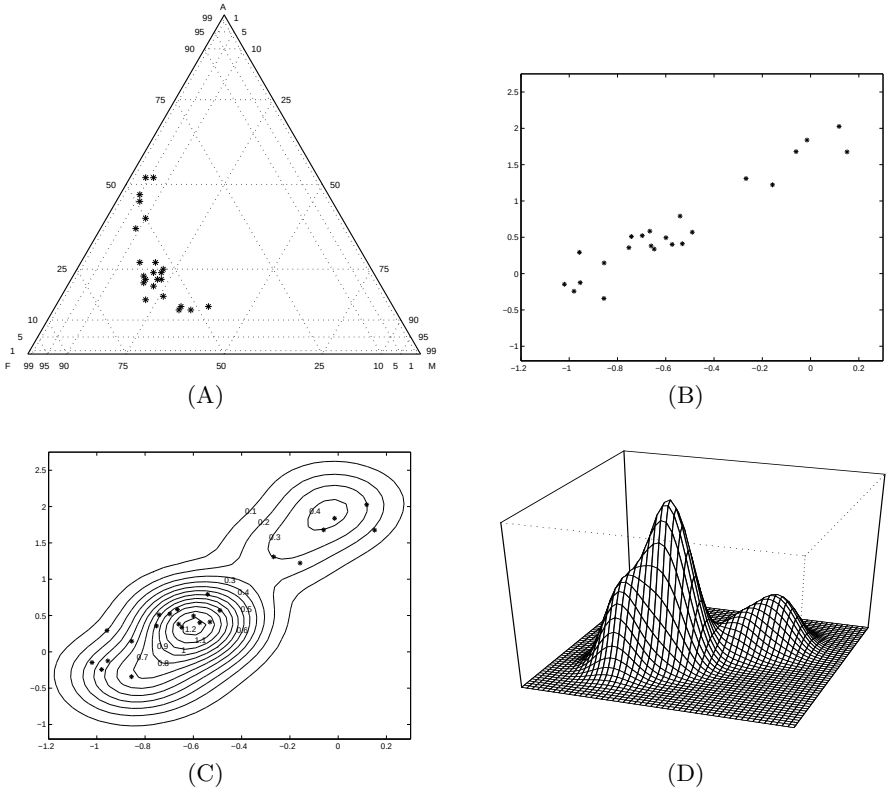


Fig. 2. Kernel Density Estimation for Skye lavas data set: (A) Data in the ternary diagram (B) transformed data in the ilr-space; (C) Contour plots in the ilr-space; (D) Density estimates in the ilr-space.

As was suggested by [WJ95] (p. 108), although the $\mathcal{F}$ parameterisation is recommended for a data not oriented parallel to the coordinate axes, in this example we state (Figure 3) that using $H \in \mathcal{D}$ we can obtain reasonable results.

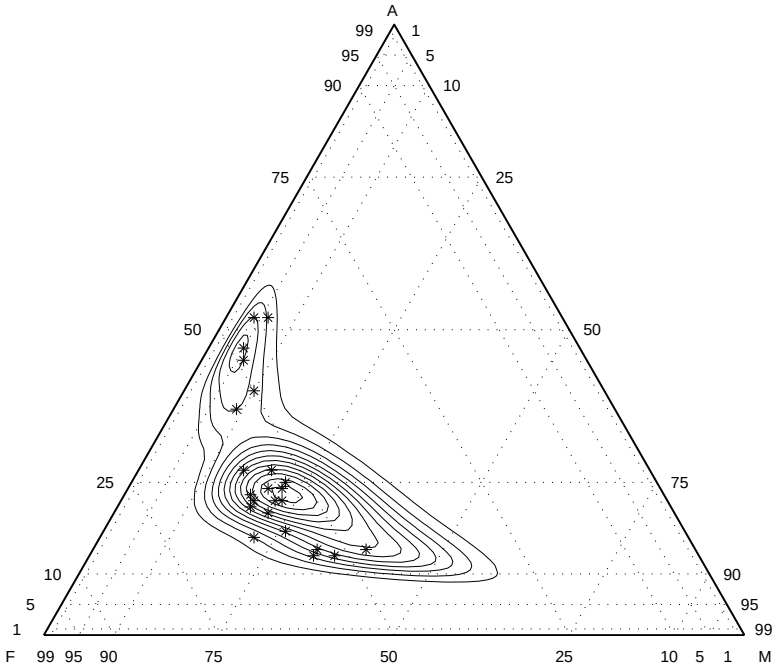


Fig. 3. Kernel Density Estimation for Skye lavas data set: contour plots in the ternary diagram.

5 Acknowledgments

This work has received financial support from the Dirección General de Investigación of the Spanish Ministry for Science and Technology through the projects BFM2003-05640/MATE and MTM2005-06348.

References

[ABMP00] Aitchison, J., Barceló-Vidal, C., Martín-Fernández, J.A., Pawlowsky-Glahn, V.: Logratio analysis and compositional distance. Math. Geol., textbf32(3), 271–275 (2000)

[Ait86] Aitchison, J.: The statistical analysis of compositional data. Chapman & Hall, London (1986). Reprinted in 2003 by Blackburn Press.

[AL85] Aitchison. J., Lauder, I.J.: Kernel Density Estimation for Compositional Data. Applied Statistics, **34(2)**, 129–137 (1985)

[BA97] Bowman, A.W., Azzalini, A.: Applied smoothing techniques for data analysis: the Kernel approach with S-Plus illustrations. Clarendon Press, Oxford (1997)

[DH03] Duong, T., Hazelton, M.L.: Plug-in bandwidth matrices for bivariate kernel density estimation. Journal of Nonparametric Statistics, **15**, 17–30 (2003)

[DH05] Duong, T., Hazelton, M.L.: Cross-validation bandwidth matrices for multivariate kernel density estimation. Scandinavian Journal of Statistics, **32**, 485–506 (2005)

[EPMB03] Egozcue, J.J., Pawlowsky-Glahn, V., Mateu-Figueras, G., Barceló-Vidal, C.: Isometric log-ratio transformations for compositional data analysis. Math. Geol. **35 (3)**, 279–300 (2003)

[MBP03] Martín-Fernández, J.A., Barceló-Vidal, C., Pawlowsky-Glahn, V.: Dealing with Zeros and Missing Values in Compositional Data Sets. Math. Geol., **35(3)**, 253–278 (2003)

[MB05] Mateu-Figueras, G., Barceló-Vidal, C.: Eds., Proceedings of the Second Com-positional Data Analysis Workshop (CODAWORK'05), October 19-21, University of Girona (Spain), CD-ROM, ISBN: 84-8458-222-1 [available in http://ima.udg.es/Activitats/CoDaWork05/] (2005)

[Mat03] Mateu-Figueras, G.: Distribution Models on the Simplex. Ph.D. thesis, Universitat Politècnica Catalunya, Barcelona, [available in www.tdcat.cesca.es/index_tdx_an.html], ISBN: 84-688-6734-9, 202p. (2003)

[WJ95] Wand, M.P., Jones, M.C.: Kernel Smoothing. Chapman & Hall, London, 212p. (1995)