Gaussian kernels for density estimation with compositional data

J.E. Chacón a,*, G. Mateu-Figueras b, J.A. Martín-Fernández b

a Dept. de Matemáticas, U. de Extremadura, 06006 Badajoz, Spain
b Dept. de Informática y Matemática Aplicada, U. de Girona, 17071 Girona, Spain

A R T I C L E   I N F O
Article history:
Received 9 June 2009
Received in revised form 27 November 2009
Accepted 11 December 2009
Available online 30 October 2010

Keywords:
Bandwidth
Isometric log-ratio
Normal distribution
Simplex

A B S T R A C T
Common simplifications of the bandwidth matrix cannot be applied to existing kernels for density estimation with compositional data. In this paper, kernel density estimation methods are modified on the basis of recent developments in compositional data analysis and bandwidth matrix selection theory. The isometric log-ratio normal kernel is used to define a new estimator in which the smoothing parameter is chosen from the most general class of bandwidth matrices on the basis of a recently proposed plug-in algorithm. Both simulated and real examples are presented in which the behaviour of our approach is illustrated, which shows the advantage of the new estimator over existing proposed methods.

© 2010 Elsevier Ltd. All rights reserved.

1. Introduction

Compositions are defined as vectors in which the components represent a relative contribution of different parts of a whole; therefore, their sum is a constant \( c \) depending on the units of measurement (typically, \( c = 1 \), 100 or \( 10^6 \)). The sample space of compositional data (Aitchison, 1986) is the simplex \( S^p \), which is defined as

\[
S^p = \{ \mathbf{x} = [x_1, \ldots, x_p] : x_1 > 0, x_1 + \cdots + x_p = c \}.
\]

As stated during a recent Compositional Data Workshop (Daunis-i-Estelada and Martín-Fernández, 2008), this type of data appear frequently in geosciences and in many other disciplines such as archaeometry, economics, biomedical research or space research. After Aitchison (1986), it was generally accepted that compositional data reflect only relative magnitude, and thus the main interest lies in relative rather than absolute change. A statistical analysis using log-ratios is based on the simplex exhibiting a natural geometry that is consistent with the intuitive concept of relative difference. During recent years, some progress has been made, and the specific algebraic-geometric structure of \( S^p \) is now widely used (Billheimer et al., 2001; Pawlowsky-Glahn and Egozcue, 2001) and applied in practical studies (e.g. Kolb et al., 2006; Pierotti et al., 2009).

Furthermore, kernel density estimation is a useful tool for analytical purposes (Silverman, 1986); it is widely used by statisticians and is included with most statistical software packages. Although this method was primarily developed for real data, both univariate and multivariate (Wand and Jones, 1995), some other applications are available, including Hall et al. (1987) in which density estimation with spherical data was explored, or Bowman and Azzalini (1997, p. 14) where a transformation method was proposed to handle non-standard data.

Another application was provided in the paper by Aitchison and Lauder (1985) in which two multivariate kernel methods of density estimation for compositional data were introduced, and the Dirichlet and the additive logistic normal (aln) densities on the simplex were proposed as two alternatives for the kernel in the density estimator. The authors recommended the selection of the aln kernel rather than the Dirichlet except “if there is the least suspicion of sparseness in the data”. In this paper, this recommendation is strongly related to the selected kernel function. In addition, as stated by the authors, since the aln kernel is based on the additive log-ratio (alr) transformation, only the typical parameterisations of the bandwidth matrix (Wand and Jones, 1995, p. 91) involving the covariance matrix can be used. Nevertheless, for real bivariate data Wand and Jones (1993) demonstrate that this parameterisation of the bandwidth matrix is sometimes inadequate, and thus its use is not recommended in practice. Fortunately, we can currently take advantage of new, full (i.e. unconstrained) bandwidth matrix selection methods (Duong and Hazelton, 2003, 2005; Chacón and Duong, 2010) that were not available when the paper Aitchison and Lauder (1985) was published.

In this paper, kernel density estimation methods for compositional data are updated. The proposed approach incorporates recent developments related to compositional data analysis and bandwidth matrix selection theory. First, basic concepts are introduced that are related to the vector space structure of the
simplex and the definition of probability density functions on \( S^p \).

Then, new methods for kernel density estimation are proposed, and examples of the behaviour of the proposed approach are given. Finally, some important issues related to the techniques in this paper are given, and subjects of future research are discussed.

### 2. Recent developments on compositional data analysis

As previously discussed, special treatment is required for compositional data. The methodology for working with log-ratios was introduced by Aitchison (1986); in particular, a strategy based on transformations was proposed to handle compositional data. Both the additive log-ratio (alr) and the centred log-ratio (clr) transformations were introduced as

\[
alr(x) = [\ln(x_1/x_0), \ldots, \ln(x_{p-1}/x_0)],
\]

\[
clr(x) = [\ln(x_1/g(x)), \ldots, \ln(x_p/g(x))],
\]

where \( g(x) = (x_1, \ldots, x_p)^{1/p} \) is the geometric mean of the composition \( x \).

Essentially, the transformation method proposed in Aitchison and Lauder (1985) regarding kernel density estimation consists of estimating the density of the alr-transformed data followed by a transformation back to the simplex. Some caution is advised if the alr transformation is applied because it is asymmetric in its components. In addition, the alr transformation is not isometric, and thus it does not preserve distances. Although the clr transformation is isometric, its weakness is that the covariance matrix of the transformed dataset is singular. To avoid these complications, the isometric log-ratio transformation (ilr) was introduced in Egozcue et al. (2003):

\[
ilr(x) = y = [y_1, \ldots, y_{p-1}] \in \mathbb{R}^{D-1},
\]

where

\[
y_l = \frac{1}{\sqrt{l(l+1)}} \ln \left( \frac{\prod_{j=1}^l y_j}{x_{l+1}} \right)
\]

The simplex \( S^p \) has a \((D-1)\)-dimensional Euclidean space structure (Buccianti et al., 2006) with specific operations. Let \( C(\cdot) \) denote the closure operation, which normalises any vector \( x \) to a constant sum (Aitchison, 1986), and let \( x, x' \in S^p \) and \( b \in R \). The internal operation, perturbation, is defined as \( x \oplus x' = C[x_1 x_2 \ldots x_p x'_p] \). The corresponding inverse operation, which will be applied later, is defined as \( x \ominus x' = C[x_1/x_2 \ldots x_p/x'_p] \). Note that \( x \ominus x' \) is the neutral element \( e = [1/1, \ldots, 1/D] \), which is the centre of the sample space. Therefore, with perturbation, any composition can be moved to the centre of the simplex or to any other location in the same manner that real data are moved in the real space. The external operation, power transformation, is defined as \( b \odot x = C[x_1^b \ldots x_p^b] \). Finally, the inner product (the Aitchison inner product (Aitchison, 2002)) is defined as \( \langle x, x' \rangle_{S^p} = (1/D) \sum_{j=1}^p \ln(x_j/x'_j) \ln(x'_j/x_j) \).

The Aitchison inner product and its associated norm ensure the existence of orthonormal bases on the simplex. Consequently, the coordinates of each composition can be used with respect to an orthonormal basis. The orthonormal basis is not unique, and the determination of the most appropriate basis for a given problem is not trivial. In this paper, the specific basis given in Egozcue et al. (2003) is used. For this basis, the orthonormal coordinates equal the isometric log-ratio transformed vector in Eq. (2). Therefore, the notation ilr(\( x \)) is used to emphasise this similarity. In addition, any vector of coordinates consists of log-ratios, which is in agreement with Aitchison’s (1986) theory. It is important to note that all of the standard real analysis techniques can be applied to the coordinates because they behave as real vectors (Eaton, 1983; Pawlowsky-Glahn, 2003). Nevertheless, if a statistical method is applied to the coordinates, it is interesting to note whether the results depend on the selected orthonormal basis. In this paper, a kernel based on the ilr coordinates is proposed, and an analysis of its dependence on the basis is also provided.

Furthermore, a natural measure on the simplex, \( \lambda_a \), that is compatible with its space structure can also be defined using orthonormal coordinates (Pawlowsky-Glahn, 2003). This measure is continuous with respect to the Lebesgue measure, \( \lambda \), on the real space, and the relationship between them is given as

\[
|d\lambda_a/d\lambda| = (\sqrt{D} x_1 \ldots x_p)^{-1}.
\]

To define probability density functions on \( S^p \), a reference measure is required. Given the special algebraic-geometric structure of \( S^p \), it is convenient to work with probability density functions with respect to the natural measure on \( S^p \), \( \lambda_a \) (Mateu-Figueras and Pawlowsky-Glahn, 2005, 2007, 2008). In practice, the density functions with respect to \( \lambda_a \) are equivalent to the density functions with orthonormal coordinates. However, since orthonormal coordinates behave as real random vectors, any density in \( S^p \) with respect to \( \lambda_a \) can be transformed to a density function with respect to the Lebesgue measure in the coordinate space (Mateu-Figueras and Pawlowsky-Glahn, 2008).

Therefore, there are different methods to express the density functions on the simplex; they can be expressed on the simplex with respect to \( \lambda_a \) or as density functions for orthonormal coordinates. It is not the objective of this paper to discuss or to compare various equivalent possibilities; rather, the purpose is to define kernel density estimators using density functions on \( S^p \) and to take into account the special geometric structure of the simplex. Both equivalent strategies can be considered, but the simplest one will be used in each case. More information on those strategies can be found in the papers by Mateu-Figueras and Pawlowsky-Glahn (2005, 2008) in which different possibilities are presented and a comparison with the classical methodology (classical Dirichlet and the aln distributions) is provided.

Additionally, compositional variables frequently take null values due to rounding, and log-ratios exclude the process of dealing with zeros. In Aitchison and Lauder (1985, p. 133), a general procedure for replacing rounding zero values in a composition was recommended, but the non-parametric multiplicative replacement method introduced by Martín-Fernández et al. (2003) overcame this procedure. Another recent technique based on an EM algorithm improved the results from the multiplicative replacement as the number of zeros increases (Palarea-Albaladejo and Martín-Fernández, 2008).

### 3. New strategies for compositional kernels

Suppose that we have compositional data \( \mathbf{X}_1, \ldots, \mathbf{X}_n \) having a continuous distribution on the simplex \( S^p \), with density \( f: S^p \rightarrow \mathbb{R} \) with respect to \( \lambda_a \). Then a preliminary kernel estimator of \( f \) can be defined as

\[
f_{n,k,h}(x) = \frac{1}{n} \sum_{i=1}^n k(x_i; h), \quad x \in S^p,
\]

where the bandwidth \( h \) is a positive real number, and the kernel \( k(\cdot; h) : S^p \rightarrow \mathbb{R} \) is a density function on \( S^p \) centred on the data point \( x \), and spread out depending on the relative size of the smoothing factor \( h \).

#### 3.1. Kernels based on the Dirichlet density

The well-known Dirichlet class \( J(x|z) \) of distributions on the simplex can be used to define a kernel for compositional data. The expression of its density function with respect to the Lebesgue measure \( \lambda \) is widely known and appears frequently in the literature.
Using Eq. (3), the Dirichlet density function is obtained with respect to $\mathcal{X}$:

$$
\sqrt{D}f(z_1 + \cdots + z_D) = \frac{1}{I(z_1) \cdots I(z_D)} x_1^{\gamma_1 - 1} \cdots x_D^{\gamma_D - 1},
$$

(4)

where $x = [x_1, \ldots, x_D] \in S^D$, and $z = [z_1, \ldots, z_D] \in R^D$. The expression of the classical Dirichlet and Eq. (4) is quite similar, but they present some important differences, especially in the moments (Mateu-Figueras and Pawlowsky-Glahn, 2005). In both cases, the Dirichlet density $A(x(z))$ is centred at $e$ if $x$ is a multiple of $[1, \ldots, 1] \in R^D$, which is the $D$-vector of units. The key factor is the type of centring operation that is appropriate for the kernel estimator.

Aitchison and Lauder (1985) use the classical Dirichlet and suggest taking

$$
k(x(X, h) = A\left(\frac{x + h X}{1 + h X}\right).
$$

(5)

In this case, it is easy to show that for $x = \frac{1}{1 + (1/h)X}$, the mode of the classical Dirichlet distribution is at $X$. Additionally, it is clear that for this value of $x$, the concentration of the distribution around the mode depends on the values of $X$. Essentially, this fact causes different behaviours between the Dirichlet and the aln kernels with respect to the sparseness in the data. This behaviour was discussed by Aitchison and Lauder (1985) from an empirical point of view. To avoid this effect, a different strategy (Martín-Fernández et al., 2006) is used: first, density in Eq. (4) is used, and the kernel is then given as

$$
k(x(X, h) = A\left(\frac{x \otimes X}{1 + h X}\right).
$$

Then, the mode of the distribution $A(x(1/h X))$ is $e$, but due to the centring operation $x \otimes X$, it follows that $A(x \otimes X(1/h X))$ is centred on $X$. Additionally, the smoothing factor $h$ is exclusively related to the concentration of the distribution. Typically, if $h$ becomes larger, then the distribution about the mode becomes less concentrated. Although this strategy improves the Dirichlet kernel, this family of kernels has additional difficulties in practice (see Section 5 for additional discussion). For this reason, this paper focuses on the use of Gaussian kernels.

3.2. Kernels based on the normal density

Currently, one of the most popular choices for the kernel in a multivariate density estimator is the standard multivariate normal density. In fact, the smoothing parameter $h$ can also be generalised to a symmetric positive definite matrix $H$, which is known as the bandwidth matrix (Duong and Hazelton, 2005). In Mateu-Figueras and Pawlowsky-Glahn (2008), the normal distribution on $S^D$ is defined through the density function of the orthonormal coordinates. The idea is simple: a random composition is considered to have a normal distribution on $S^D$ if the density function of the corresponding orthonormal coordinates, which is denoted as $\text{ilr}(x)$, is the standard normal density on the real space $R^{D-1}$, which is denoted as $\phi$. In this paper, the Gaussian kernel on $S^D$ is proposed, which is called the iln kernel and is defined as

$$
k_{\text{iln}}(x(X, H)) = \phi_{\text{d}}(\text{ilr}(x) - \text{ilr}(X)), \quad x \in S^D,
$$

where $\text{ilr}(x)$ and $\text{ilr}(X)$ are the corresponding orthonormal coordinates of $x$ and $X$, and $\phi_{\text{d}}(y) = (2\pi)^{-(D-1)/2} |H|^{-1/2} \exp\left(-\frac{1}{2} y^T H^{-1} y\right)$ for $y \in R^{D-1}$. The bandwidth matrix $H$ is a positive definite symmetric matrix of order $(D - 1) \times (D - 1)$. This general smoothing parameter is called a full bandwidth matrix.

Fig. 1 shows the contour plots of iln kernels for $D=3$ in the coordinate space and in the simplex $S^2$. This figure shows that in the coordinate space, the typical circles and ellipses of the multivariate normal density are obtained.

Regarding the centring operation, the iln kernel verifies $k_{\text{iln}}(x(X, H)) = k_{\text{iln}}(x \otimes X(e, H))$. This property is similar to the procedure in the real space by means of the vector subtraction operation. Regarding the variance structure, with the iln kernel, a more general method of spreading out the probability mass around the point $X$ is used. Whereas the use of a single bandwidth parameter $h$ implies that the same amount of smoothing is applied along all of the coordinate directions, the use of a full bandwidth matrix $H$ allows for various amounts of smoothing in different directions, even in directions other than the coordinate directions.

With respect to full bandwidth matrices, Wand and Jones (1995, p. 92) consider $F$ to be the class of symmetric positive definite $(D - 1) \times (D - 1)$ matrices, $D$ to be the subclass of diagonal positive definite matrices, and $S$ to be the subclass $(h^2 I_d: h > 0)$, where $I_d$ is the identity matrix. For our methodology, if $H \in S$, then the isodensity curves of the considered kernels are circles in the ilr-transformed space. If $H \in D$, then $H = \text{diag}(h_1^2, \ldots, h_D^2)$, and the isodensity curves of the kernels are ellipses such that their axes are parallel to the coordinate directions in the ilr-transformed space. For the full bandwidth matrix class $F$, the axes of the ellipses in the ilr-transformed space are not required to be parallel to the coordinate axes. To illustrate the behaviour of the iln kernel, simple cases for the three different parameterisations of the bandwidth are shown. Fig. 1 shows the contour plots for $H \in S$, $H \in D$ and $H \in F$. Fig. 1A corresponds to the circles in the ilr-transformed space, and Fig. 1B shows the contours plots transformed back to the simplex $S^2$. Figs. 1C and D show the contour plots for $H \in D$; the axes of the ellipses are parallel to the coordinate axes. Figs. 1E and F show the contour plots for $H \in F$. In the case of the full bandwidth matrix, the axes of the ellipses are not parallel to the coordinate axes.

Our proposal is more general than the one in Aitchison and Lauder (1985). In fact, Aitchison and Lauder (1985) observed that the usual parameterisations of the bandwidth matrix $H$ are not possible if the aln kernel is used because the alr transformation is used, and consequently, the results are not invariant for permutations of the components. More specifically, in Eq. (1), the component $x_D$ was chosen as the common divisor, but any other component could have been chosen. Consequently, for an analysis involving alr vectors, it is important to check the invariance with respect to the group of permutations of the components (Aitchison, 1986; Aitchison and Lauder, 1985) or the invariance with respect to the choice of the divisor in the alr transformation. As Aitchison and Lauder (1985) advised, the kernel density estimation procedure using the alr transformation is not invariant under permutation of its components. To overcome this difficulty, they proposed a bandwidth matrix proportional to the sample covariance matrix of the alr-transformed dataset, which is denoted as $S$, i.e. $H = SH$ (see the Appendix for details). For data in real space, Wand and Jones (1993) later demonstrated that this particular kind of bandwidth matrix is only appropriate in the case of multivariate normal data and not for general density shapes. Consequently, the aln kernel is only appropriate for the aln class. Using our strategy with the iln kernel, all parameterisations of the bandwidth matrix $H$ are feasible, and all of the densities shapes can be estimated.

On the other hand, it is important to mention the invariance under changes of the orthonormal basis for the iln kernel. Note that we are working with ilr vectors, which can be considered to be vectors of coordinates with respect to an orthonormal basis in the simplex. The ilr vector in Eq. (2) was obtained using the specific orthonormal basis given in Egozcue et al. (2003), but any other orthonormal basis could have been chosen as well. Thus, in any compositional analysis involving ilr vectors, it is important to check the invariance under the change of orthonormal bases. This issue is not specific to compositional data, it is a general problem in any real space. For example, Céleux and Govaert (1995) analysed a similar situation related to the dependence of the basis in model-based
cluster analysis. For our case, the choice of the iln kernel guarantees that the results are invariant with respect to changes of the orthonormal basis. See the Appendix for further details.

4. Numerical examples

In this section, the behaviour of our proposed method, which is based on the iln kernel, is illustrated. First, we provide an ample simulation study in which the performance of the proposed method is compared to that of the kernel based on the alr transformation (Aitchison and Lauder, 1985). Then a real dataset is used to discuss and illustrate the difficulties related to the change of orthonormal bases or the permutation of the components using the iln or aln kernel methods, respectively.

4.1. Simulation study

An extensive simulation was designed to demonstrate the advantages of the proposed methods in practice. Particularly, the performance of three methods was compared over a certain set of test densities (described below). The considered methods are as follows:

- The kernel estimator with the iln kernel and the full bandwidth matrix chosen by cross-validation (Duong and Hazelton, 2005). This method is denoted as CV.
- The kernel estimator with the iln kernel and the full bandwidth matrix chosen by the unconstrained plug-in method (Chacón and Duong, 2010). This method is denoted as CD.
- The kernel estimator suggested by Aitchison and Lauder, using the aln kernel. In this case, as discussed above, we were forced to use the bandwidth matrix of the form $H = hS$, where $S$ stands for the sample covariance matrix, and $h$ was chosen to maximise the pseudo-likelihood (Aitchison and Lauder, 1985). This method is denoted as AL.

Chacón (2009) proposed a set of 12 densities to compare the performance of various bandwidth selection methods in a real space. A similar set of densities was used in Celeux and Govaert.
In this paper, a similar strategy is used on the simplex, and a set of test densities consisting of 12 normal mixture densities on $S^3$ was chosen (see Mateu-Figueras and Pawlowsky-Glahn, 2008 for the definition of the normal density on $S^3$). Their contour plots are shown in Fig. 2.

For every test density in the simulation study, $B = 500$ random samples of size $n = 100$ were generated. Then the kernel estimator was computed using the three methods described above. Finally, to evaluate the performance of each kernel estimator $f_{n,k,H}(x)$, the integrated square error (ISE) was used. The ISE is defined as the squared $L_2$-distance between $f_{n,k,H}$ and the test density $f$, namely

$$\text{ISE}(H) = \int_{S^3} (f_{n,k,H}(x) - f(x))^2 d\lambda_d(x).$$

(6)

Therefore, the performance of the three methods was compared on the basis of the distribution of their corresponding ISEs.
4.1.1. Some previous computational details

To compute the ISE for densities on the simplex, some difficulties arose from the fact that the integral in Eq. (6) is not an ordinary integral.

If the iln kernel is used, then the easiest solution is to work on the coordinate space with respect to an orthonormal basis. Because the densities were expressed in terms of coordinates and because the orthonormal coordinates behave as real random vectors,
the resulting densities could be expressed with respect to the Lebesgue measure in the coordinate space (Mateu-Figueras and Pawlowsky-Glahn, 2008). Consequently, if the iln kernel is considered, then Eq. (6) reduces to

\[
\text{ISE}(\mathbf{H}) = \int_{\mathbb{R}^2} \left( \tilde{f}_{\mathbf{n},\mathbf{H}}(\mathbf{y}) - \tilde{f}(\mathbf{y}) \right)^2 d\lambda(\mathbf{y}), \tag{7}
\]

where \(\tilde{f}\) is the density function of the ilr coordinates, which corresponds to the density \(f\) on \(S^D\). In this case, \(f\) is a normal mixture density on \(\mathbb{R}^2\). Additionally, the function \(\tilde{f}_{\mathbf{n},\mathbf{H}}(\mathbf{y}) = f_{\mathbf{n},\mathbf{H}}(\mathbf{y}; \mathbf{Y}_1, \ldots, \mathbf{Y}_n)\) is the usual kernel estimator of \(f\) in \(\mathbb{R}^2\) using a Gaussian kernel and based on the data \(\mathbf{Y}_1 = \text{ilr}(\mathbf{x}_1), \ldots, \mathbf{Y}_n = \text{ilr}(\mathbf{x}_n)\); i.e., \(\tilde{f}_{\mathbf{n},\mathbf{H}}(\mathbf{y}) = (1/n) \sum_{i=1}^{n} \phi_H(\mathbf{y} - \mathbf{Y}_i)\). This results in significant computational gain because there are exact formulas to compute Eq. (7) in the case of normal mixtures, which avoids the process of numerical integration (refer to Appendix C in Wand and Jones, 1995).

Using the aln kernel, we computed the ISE in a different manner. Aitchison and Lauder (1985) proposed the application of the alr transformation and a multivariate normal kernel. Consequently, the corresponding ISE could also be computed in the alr-transformed space. The primary goal was to express the set of 12 test densities \(f\) in terms of alr coordinates. The main difficulty was that the densities \(f\) are normal mixture densities on \(S^D\). From the definition of the normal in the \(S^D\) model (Mateu-Figueras and Pawlowsky-Glahn, 2008), the expression of our densities were in terms of ilr coordinates. To overcome this difficulty, the matrix relationship between the alr and ilr coordinates (Egozcue et al., 2003) and the transformation property of the normal distribution on the real space was used. Following this strategy, the

---

**Fig. 4.** The Skye lavas dataset: (A) ternary diagram; (B) in the alr space using A as the divisor; (C) in the alr space using F as the divisor; (D) in the alr space using M as the divisor; (E) in the ilr space using a typical orthonormal basis; (F) in the ilr space using a rotated orthonormal basis.
4.1.2. Simulation results

For each test density in the simulation study, Fig. 3 shows boxplots of the distribution of the ISE for the three estimation methods in the study for $B=500$ simulation runs.

In all of the cases, two facts were clearly noticeable:

- The CV method was much more variable than the other two methods. This behaviour has also been noted in the case of real data (Cao et al., 1994). Its median performance was generally acceptable, and it was better than the CD method for the asymmetric fountain density.
- The AL and CD methods typically presented a similar amount of variability, within a reasonable range. However, the CD method always outperformed the AL method in median terms (significantly in some cases). The CD method also outperformed the CV method in all of the cases except for the asymmetric fountain density.

Overall, the CD method retained the good properties of both methods. The CD method performed better than the CV method by exhibiting less variability, and it performed better than the AL method by obtaining a lower median ISE. Therefore, the CD method is recommended in practice.

As suggested by an anonymous referee, we repeated the simulation study with a small sample size ($n=30$) to explore a situation similar to the real data example below and a very large sample size ($n=1000$) to gain insight into the asymptotic properties of the estimators. The boxplots of the results were omitted to save space because they were very similar to the case for $n=100$; however, the full details are available from the authors.

4.2. Real data example

The application of both the aln kernel and the iln kernel were analysed using the three-composition AFM of 23 aphyric Skye lavas (Aitchison, 1986; Aitchison and Lauder, 1985). The variables A, F and M represent the relative proportions of Na$_2$O+K$_2$O, Fe$_2$O$_3$ and MgO, respectively. This dataset was also used to demonstrate the dependence of the choice of the denominator in the aln kernel and the invariance under the change of orthonormal bases in the iln kernel.

Fig. 4A shows the Skye Lavas dataset in a ternary diagram. Larger variability appeared on components A and M. In Figs. 4E and F, two ilr-transformed datasets are plotted using two different orthonormal bases. In Figs. 4B–D, the corresponding alr-transformed datasets, which used three different components as a divisor, are shown. In all of the cases, the points exhibit a linear pattern that is not parallel to the coordinate axes. However, for the ilr case, the effect of changing the basis is only a rotation ($90^\circ$ in this case), whereas in the alr case, a change in the denominator also changes the variability structure.

The results of the iln kernel and the CD method with a full bandwidth matrix are shown in Fig. 5. In this case, for any parameterisation of the bandwidth matrix, there is invariance with respect to the choice of the orthonormal basis; thus, the kernel density estimator on the simplex (Fig. 5B) produced exactly the same results for the two orthonormal bases.

Using the aln kernel, the method proposed in Aitchison and Lauder (1985) was applied, i.e. a bandwidth matrix proportional to the alr sample covariance matrix was used. Fig. 6 shows the results of this method. Different results were obtained compared to the iln kernel approach. It is clear that the AL estimator uses a bandwidth that is too much concentrated around the data points, which results in an artificially spiky density estimate. This is consistent with the results of the simulation study because it is well known that high ISE values such as those obtained by the AL method are typically...
obtained. For instance, in Figs. 7A and B, the final estimated
meterisations of the bandwidth matrix were used (for example,
transformation.

invariant with respect to the choice of the divisor in the alr
function.

by defining

\[ K_{h}(x \odot X_i) = \frac{1}{n^{D-1}} K \left( \frac{1}{h} \odot (x \odot X_i) \right), \quad x \in S^D. \]  

(8)

In this manner, \( k(x;X_i,h) \) is a straightforward extension of the real
multivariate case because \( K_{h}(x \odot X_i) \) represents a probability mass
centred at \( X_i \), which is scattered depending on the value of \( h \).

Using the previous notation in Eq. (8), the kernel \( K \) can be
recovered by setting \( h = 1 \) and \( X_i = e \); i.e., \( K(x) = K_{1}(x \odot e) = k(x;e,1) \).
This can be used to show that both methods proposed in Section 3.1
are not constructed in this manner. For instance, if the proposed
method in Aitchison and Lauder (1985) were based on some
function \( K \), then necessarily \( K(x) = k(x;e,1) = A(xj + e) \) is required,
which is certainly a density function on the simplex centred on \( e \).
However, it is not difficult to check that

\[ A \left( xj + \frac{1}{h} X_i \right) = \frac{1}{n^{D-1}} A \left( \frac{1}{h} \odot (x \odot X_i) \mid j + e \right) = K_{h}(x \odot X_i), \]

and this shows that Eq. (5) does not consist of scaling a fixed kernel
function, as it is usual in the real multivariate case. Although for the
definition of the kernel density estimator it is not essential that the
kernel satisfies the condition \( k(x;X_i,h) = K_{h}(x \odot X_i) \), not fulfilling it
may represent an important drawback when studying the asymptotic
properties of the kernel estimator.

Therefore, this offers the further possibility of using the
Dirichlet class of densities in kernel density estimation for composi-
tional data (which will be reported elsewhere). Particularly, we
can set \( K(x) = A(xj) \) and use \( k(x;X_i,h) = K_{h}(x \odot X_i) \) as the kernel in the
density estimator.

Also, some uncertainty remains regarding the possibility of
using a more general form of the variance structure rather than a
single smoothing parameter with the Dirichlet kernel. In other
words, we are not certain whether it is possible to use both the
Dirichlet class (as a kernel) and a full bandwidth matrix. If the iln
kernel is used, then this drawback can be overcome.

6. Conclusions

By combining recent advances from compositional data analysis
and multivariate kernel density estimation, a new strategy for
density estimation with compositional data were presented, which
outperforms the previous approaches in practice. The new meth-
ology is based on the combination of the use of the isometric
log-ratio normal kernel, which allows the handling of general
bandwidth matrices and plug-in methods for the automatic selec-
tion of such a general bandwidth matrix from the data. In addition,
we demonstrated that the new estimator is invariant with respect
to the choice of the basis on the simplex, and we identified some
directions for future research in the field.

Acknowledgments

This research has been supported by the Direcció General de Inves-
tigación of the Spanish Ministry of Science and Innovation
under projects MTM2006-06172 (first author) and MTM2009-
13272 (second and third authors), and by the Agència de Gestió
d’Ajuts Universitaris i de Recerca of the Generalitat de Catalunya
under project Ref: 2009SGR424 (second and third authors).

Appendix

If the density is estimated with the method proposed by
Aitchison and Lauder (1985), then the property of invariance under
permutation of components holds. However, if we use the common parameterisations of the bandwidth matrix, $\mathbf{H} \in F, D, \text{ or } S$, then the result depends on the divisor used in the air transformation. In fact, given a $D$-part compositional dataset $X_1, \ldots, X_n$, the aln kernel density estimate proposed by Aitchison and LAuder (1985) can be obtained with the corresponding alr-transformed sample $Y_1, \ldots, Y_n$, where $Y_i = \text{alr}(X_i)$, using a multivariate normal kernel with a bandwidth matrix that is proportional to the sample covariance matrix, i.e., $\mathbf{H} = \mathbf{hS}$, where

$$\mathbf{S} = \frac{1}{n-1} \sum_{j=1}^{n} (Y_j - \bar{Y})(Y_j - \bar{Y})^T, \quad \bar{Y} = \frac{1}{n} \sum_{j=1}^{n} Y_j,$$

and $h$ is chosen to maximise the pseudo-likelihood function. If another component is chosen as the common divisor in the air transformation in Eq. (1), then the transformed sample and the corresponding covariance matrix will change. Nonetheless, the pseudo-likelihood function used to compute $h$ depends only on the determinant ($\mathbf{S}$) and on the quadratic form:

$$\mathbf{y} - \mathbf{y}_i / \mathbf{S}^{-1} (\mathbf{y} - \mathbf{y}_i).$$

Aitchison (1986) demonstrated that the determinant ($\mathbf{S}$) and the quadratic form in Eq. (9) are invariant under the group of permutations of parts. Therefore, although the bandwidth matrix $\mathbf{H}$ depends on the component selected as the divisor, the value of $h$ remains invariant. In this case, it is clear that the kernel density estimator depends on $\mathbf{H}$ through the determinant $|\mathbf{H}|$ and the quadratic form in Eq. (9). Because $|\mathbf{H}| = |h\mathbf{S}| = h^{1/2} |\mathbf{S}|$, the aln kernel proposed by Aitchison and LAuder (1985) is invariant under permutation of components. On the other hand, if common parameterisations of the bandwidth matrix $\mathbf{H}$ are used, $\mathbf{H} \in F, D, \text{ or } S$, then it is trivial to check that the quadratic form $(\mathbf{y} - \mathbf{y}_i) / \mathbf{H}^{-1} (\mathbf{y} - \mathbf{y}_i)$ and the determinant $|\mathbf{H}|$ depends on the common divisor used in the air transformation, and consequently, that the kernel density estimator is not invariant.

If the iln kernel is used, then the invariance with respect to the choice of orthonormal bases holds. In fact, let $x \in S^D$ be a $D$-part composition and $\text{ilr}(x)$ and $\text{ilr'}(x)$ be the ilr coordinates with respect to two different orthonormal bases. Denote the bandwidth matrices as $\mathbf{H}$ and $\mathbf{H}'$ from $F, D, \text{ or } S$ that were obtained using the corresponding ilr vectors. From lineal algebra, there is a change-of-basis matrix $\mathbf{A}$ such that

$$\text{ilr'}(x) = \mathbf{A} \text{ilr}(x) \quad \text{and} \quad \mathbf{H}' = \mathbf{AHA}'.
$$

Because the matrix $\mathbf{A}$ is orthogonal, i.e. its inverse coincides with its transpose, the invariance of the determinant and the quadratic form involved in the iln kernel density estimate is proved as follows:

$$|\mathbf{H}'| = |\mathbf{A}| |\mathbf{H}| |\mathbf{A}|^{-1} = |\mathbf{A}| |\mathbf{H}| |\mathbf{A}|^{-1} = |\mathbf{H}|,$$

$$(\text{ilr'}(x) - \text{ilr'}(x_i))/(\mathbf{H}'^{-1}(\text{ilr'}(x) - \text{ilr'}(x_i)))$$

$$= (\text{ilr}(x) - \text{ilr}(x_i))/|\mathbf{H}^{-1}| (\text{ilr}(x) - \text{ilr}(x_i))$$

$$= (\text{ilr}(x) - \text{ilr}(x_i))/|\mathbf{A}| |\mathbf{A}|^{-1} |\mathbf{H}| |\mathbf{A}|^{-1} (\text{ilr}(x) - \text{ilr}(x_i))$$

$$= (\text{ilr}(x) - \text{ilr}(x_i))/|\mathbf{H}| (\text{ilr}(x) - \text{ilr}(x_i)).$$

In conclusion, for any parameterisation of the bandwidth matrix $\mathbf{H}$, the iln kernel density estimation is invariant under changes of bases.

References


