Assessing the effect of season, *montanera* length and sampling location on Iberian pig fat by compositional data analysis and standard multivariate statistics

Antonio Garrido-Fernández¹ and Manuel León-Camacho^{*2}

¹ Food Biotechnology Department; ² Lipid Characterization and Quality Department

Instituto de la Grasa, Campus Universitario Pablo de Olavide, Building, 46. Ctra. Sevilla-Utrera, km 1. 41013, Sevilla

Running title: Effect of season, *montanera* length, and sampling location on Iberian pig fat.

Corresponding author: Manuel León Camacho, Ph.D. e-mail: melon@ig.csic.es

22 **Abstract**

23 The work studies the effect of season, *montanera* length, and sampling location on the
24 Iberian pig fat, using compositional data (CoDa) analysis and standard statistics. CoDa
25 variation array and logratios involving C18:3: (C18:3/C17:1), (C18:3/C20:0), and
26 (C18:3/18:0) as well as the *ilr* balances (coordinates), based on C18:3, C17:1, and
27 C20:0, showed the highest variances. Discriminant Analysis (DA) led to similar
28 (conventional/coordinates) correct assignments regarding seasons (69/70%),
29 *montanera* length (71/70%) and sampling location (68/67%). Re-analyzing the
30 subcomposition of only the major fatty acids (FAs), led to slightly poorer results;
31 therefore, the removed FAs might play a role in segregation. Results are in line with
32 those from other authors and could indicate a partial capacity of pigs to control their
33 FA profiles. Overall, CoDa analysis provided useful information on data variability, the
34 effects of factors and, after conversion into coordinates, allowed applying standard
35 statistics while being respectful with the compositional sampling space.

36 **Key words:** Fatty acid (FA), Pig fat, Ham, Rump, Season, Compositional data (CoDa)
37 analysis, Hierarchical clustering analysis (HCA), Discriminant analysis (DA).

1. Introduction

The Iberian pig derived from *Sus mediterraneus* constitutes a racial group of native pigs that have inhabited in the Mediterranean silvopastoral system of the southwestern Iberian Peninsula (called the “Dehesa”) since ancient times (Rueda Sabater & Diéguez Carballo, 2007). The Iberian dry-cured ham is the most important meat product obtained from Iberian pig, elaborated following a traditional process, typical for the Mediterranean area. The extraordinary quality obtained is the result of an accurate combination of factors linked to breeding, fattening diet, fattening system of pigs in the “Dehesa”, (Dirinck, Van Opstaele, & Vandendriessche, 1997; Jurado, García, Timón, & Carrapiso, 2007; León-Camacho, Viera-Alcaide, & Vicario, 2007; Flores, Grimm, Toldrá, & Spanier, 1997), and the elaboration process of ham with long periods of drying and storage at cellars (Ruiz, Ventanas, Cava, Andrés, & García, 1999; Andrés, Cava, Ventanas, Muriel, & Ruiz, 2007).

The main contributors to the characteristic and intense flavour of this product are: i) meat quality, ii) rearing system, mainly during the final fattening period, called montanera (Mediterranean silvopastoral system), along which the pigs are fed with pasture and acorns from *Quercus subers*, *Quercus ilex*, and *Quercus faginea*, the predominant species in the “Dehesa”, iii) age of animals, iv) pig genotype, and v) conditions of the ripening process (Dirinck et al., 1997; Jurao et al., 2007; León-Camacho et al., 2007; Flores et al., 1997).

The subcutaneous lipidic fractions of the dry-cured hams have been considered as descriptors of the final fattening system (Gallardo, Narváez-Rivas, Pablos, Jurado, & León-Camacho, 2012; Narváez-Rivas, Ríos, Arteaga, Quilez, Barrero, & León-Camacho, 2008) and of the genotype (Viera-Alcaide, Vicario, Escudero-Gilete, Graciani-Constante, & León Camacho, 2008). Also, several authors have reported variations in fatty acid (FA) composition among the different anatomical regions, muscles or meat layers (Sharma, Gandemer, & Goutefongea, 1987; Franco, Escamilla, García, García Fontán, & Carballo, 2006; Monziols, Borneau, & Kouba, 2007; Narváez-Rivas, León-Camacho, & Vicario, 2009; Ros-Freixedes and Estany, 2014).

The composition in FA of the pig fat is usually determined by gas chromatography (GC) and expressed as weight percentages. Then, it represents a clear case of compositional data (constant sum of parts). Usually, its statistical analysis has been carried out by the standard multivariate methods developed for the Euclidean real space. Therefore, over the last decades, all these tools have been incorrectly applied to study this fat with possible inappropriate inferences (Aitchison, 1986). Among the numerous warnings published on the improper application of such techniques to Compositional Data (CoDa) are Aitchison (1986); Buccianti, Mateu Figueras, & Pawlowski-Glahn (2006); and Pawlowsky-Glahn & Buccianti (2011).

The first approach for the application of the exploratory CoDa analysis to the FA composition of the pig was reported by Ros, Reixach, Tor, & Estany (2011). Later, an overview of the basic techniques of the new statistic and its application to pig fat was presented by Ros-Freixedes & Stany (2014). They showed that the FA composition of muscles and subcutaneous fat did not display enough variability to become critical for standard statistics, but the CoDa analysis was useful to interpret the correlation structure among FAs. FA composition of *Calanus gracialis* was also used by Greenacre (2011) as a case study for the application of logratio analysis (LRA) and Correspondence Analysis to fish fat, which was included in a more exhaustive research on the lipid dynamics and life strategy of the Arctic Pelagic amphipods (Kraft, Graeve, Janssen, Greenacre, Falk-Petersen, 2015).

This work aims the investigation of the effect, during three consecutive seasons, of two fattening periods (*montanera* lengths) on the fatty acid composition of the Iberian pig fat from two anatomical locations (the subcutaneous fat from the rump and the adipose tissue covering the *B. femoris* muscle). The data were studied using CoDa methods and standard multivariate techniques for both the full fatty acid dataset, the subcomposition of the major FAs, and their respective CoDa derived coordinates.

2. MATERIALS AND METHODS

2.1 Animal and diets

Two hundred castrated male pure Iberian pigs, produced under the guidelines of the Designation of Origin “Jamón de Huelva” were used. The pigs were fed, according to the traditional *montanera* diet, with acorns from *Quercus subers*, *Quercus ilex*, and *Quercus faginea* and pasture, during two lengths of fattening periods: short *montanera* (M1< 90 days) and long *montanera* (M2, 91-120 days). The experiment included three consecutive seasons: 2003, 2004, and 2005.

2.2 Sampling

Two sampling locations of adipose tissue from each slaughtered animal were tested: i) One sample consisted of a chunk between 3 x 3 and 5x5 cm approximately and, at least, 6 mm of thickness (B.O.E, 2004), obtained by cutting the rump (R) at approximately 10 cm of the tail following the line of the back; in the laboratory, the skin and the loin were removed, and the adipose tissue of all chunks corresponding to each animal was punctured and homogenised, ii) the other sample was a mixture of two sub-samples from the subcutaneous adipose tissue covering the *B. femoris* muscle of both hams (H).

The above-mentioned raw samples were then melted in a microwave oven (De Pedro et al., 1997), heated for 3 min at 360W power, filtered and stored at -30°C until analysis.

2.3 Fatty acid analysis

FA methyl esters (FAME) were analysed by GC. FAMEs were extracted with n-hexane after cold methylation with 2N KOH in methanol, following the official method (B.O.E., 2004). Oleic acid was used as a reference to calculate the relative retention times. The quantification of individual FAs was carried out by evaluating the corresponding relative percentage according to the normalisation area procedure assuming an equal factor response for all FAs.

GC was performed with a Varian 3900 apparatus (Varian Co, Palo Alto, CA, USA) using a fused silica capillary HP 88 column (100m x 0,25 mm, 0,25 mm film thickness). The oven temperature was kept at 175 °C for 13 min and was then raised to 205 °C at a rate of 3.0 °C/min and held isothermally for 5.0 min. The injector

temperature was kept at 240 °C. A split ratio of 1:50 was used, and 1 µL were injected while the detector temperature was 250 °C. Hydrogen (131 kPa inlet pressure) was used as a carrier gas, while the make-up gas was nitrogen. Air and hydrogen with flow rates 450 and 40 mL/min, respectively, were used for the detector, which had a make-up flow of 40 mL/min of nitrogen.

2.4 Statistical analysis

2.4.1 Fatty acid dataset

The experimental design consisted of the combination of the levels of the three factors (season, *montanera* length, and sampling location), with the FA composition being the response and the pig distribution between the two *montanera* lengths made in agreement with the farmer. Overall, the number of samples gathered, analysed and, apparently, usable was 400.

After depuration of the results, the atypicality indices were estimated under the assumption of Additive Logistic Normal Distribution and a threshold of $p=0.95$, using CoDaPack (Department of Computer Science and Applied Mathematics, University of Girona, Spain, v2.01.14). As a result, 29 atypical observations were identified and removed. Therefore, the full dataset available for the analysis with both CoDa and the standard statistics consisted of 371 rows (samples) and 14 columns, which included three qualitative factors (season, *montanera* length, and sampling location, with 3, 2, and 2 levels, respectively) and the 11 FAs (FA matrix). In the FA matrix, each element of x_{ij} corresponded to the content of FA j for the observation i . The distribution of the samples among seasons and *montanera* lengths (in parenthesis) was 187 (M1=27, M2=160), 133 (M1=57, M2=76) and 51 (M1=0, M2=51) for 2003, 2004, and 2005, respectively, while there were 190 samples from rump and 181 from ham.

2.4.2 Overview of the CoDa statistic and its application to the Iberian pig data

2.4.2.1. Central tendency

The arithmetic mean and standard deviation do not fit with the Aitchison geometry as measures of central tendency and dispersion since they were defined for the Euclidean geometry in the real space (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018). Therefore, new alternatives were necessary. They were found in the concepts of centre, variation matrix and total variance (Aitchison 1986; Pawlowsky-Glahn et al., 2015).

For a dataset (X) of size n and D parts, the best representation of the centre is the closed geometric mean (barycenter), defined as:

$$g_m = C[g_1, g_2, \dots, g_D] \text{ where } g_j = \left(\prod_{i=1}^n x_{ij} \right)^{1/n} \quad (\text{Eq. 1})$$

where the C operator stands for closure; that is, for rescaling g_j to constant k (usually 1 or 100). The operation is performed part-wise. The range of values for each part is evaluated by estimating the corresponding 0, 25, 50 (median), 75 and 100 percentiles in the usual way. Centring moves the data to the barycenter. The operation is usually of interest when plotting since, after rescaling, the range of the observations is approximately the same, situation that improves the presentation. Besides, in the compositional biplot, centring moves the data to the origin, with the representation of parts being rays emerging from it.

2.4.2.2. Variation array

There is no similar counterpart in CoDa for measuring the variability. Usually, to account for the CoDa basic principle (data only carry relative information), compositions are transformed into logratios, with the log being used to facilitate mathematical handling (Aitchison, 1986). An overview of the logratio variability structure according to parts is the so-called variation array, which is formed by the variances of all pairwise logratios. Therefore, in CoDa, the relative variability of X is given by the matrix of logratio variances, originally defined by Aitchison (1986) as:

$$T = \begin{pmatrix} t_{11} & t_{12} & \dots & t_{1D} \\ t_{21} & t_{22} & \dots & t_{2D} \\ \dots & \dots & \dots & \dots \\ t_{D1} & t_{D2} & \dots & t_{DD} \end{pmatrix} \text{ where } t_{ij} = \text{var} \left(\ln \frac{x_i}{x_j} \right) \quad (\text{Eq. 2})$$

By definition, the diagonal terms of T are zeros, and the matrix is symmetric with respect to it. Therefore, only the upper half of the diagonal is habitually shown while the other lower half is used to include the average of the corresponding logratios $\left(\ln \frac{x_i}{x_j}\right)$, with a similar presentation structure. The cells of the variation array may be interpreted as the variability of the ratio between the two parts, with almost close to zero logratio variance approaching perfect proportionality. In this line, the variability of the pairwise logratios is usually considered as a measure of the association of the involved parts, although the term is still subjected to controversy (Filzmoser, Hron, & Templ, 2018).

The total variance, which summarizes the variability in a single value, is:

$$totvar(X) = \frac{1}{2D} \sum_{i=1}^D \sum_{j=1}^D var\left(\ln \frac{x_i}{x_j}\right) \quad (\text{Eq. 3})$$

which considers the redundant (double) information presented by the variance matrix. Then, the total variability of a compositional dataset does not depend on any particular coordinate representation.

2.4.2.2. Transformation of compositional data into coordinates

One of the most interesting options in CoDa analysis is the possibility of application of diverse logratio transformations (in agreement with the idea that compositions only carry relative information) which move the data into a new set of coordinates in the Euclidean real space, where they can be treated with standard multivariate tools. The additive logratio (alr) was the first to be introduced; but, it is rarely used currently since it depends on the part of reference. Nowadays, the two more familiar transformations are the centred and the isometric logratio transformations, with the last having the advantage of leading to orthonormal coordinates in the Euclidean real space.

a) Centred logratio (clr) transformation

This transformation has the property of preserving distances but leads to constrained data (sum of parts=0). It is defined as:

$$Z = clr(X) = \left[\ln \frac{x_1}{g_m(x)}, \dots, \ln \frac{x_D}{g_m(x)} \right] \quad (\text{Eq. 4})$$

where the denominator is the geometric mean of each observation. The variability of each *clr* component (*clr* variance) is then usually estimated part-wise from the FA matrix, although it is later transposed for vertical presentation at the end of the variation array matrix. The sum of the *clr* variances of all components also represents the total variance.

The centred *clr* coefficients have been used for the development of the exploratory CoDa biplot (Aithison and Greenacre, 2002). Since the denominator of the transformation is the geometric mean of each observation, the origin of the biplot also corresponds to the geometric mean of the data and the parts are represented by rays. As the transformation preserves distances, the observations can be interpreted in the usual way as when using data from the Euclidean real space. However, interpretation of the part loads should have in mind the special structure of CoDa, as could be observed later in the Results/Discussion section.

b) Isometric log ratio (*ilr*) transformation (coordinates)

The *ilr* transformed data are coordinates in an orthonormal basis. The first step is the selection of the basis. A particular case of selection is that linked to sequential binary partition (SBP) of the compositional vector (Egozcue, Pawlowsky-Glahn, Mateo-Figueras, & Barceló-Vidal, 2003; Pawlowsky-Glahn, Egozcue, & Tolosana-Delgado, 2015). In the first order partition, parts are split into two mutually exclusive groups. The process continues until the groups have a single part. For each order of partition, it is possible to define the balance between the two groups, coded as +1 and -1, for those parts included in the numerator and denominator, respectively, or 0 for the parts not included in any of these groups. The balance is then defined as the normalized logratio of the geometric mean of each group of parts, and it is estimated by the expression:

$$z_i = \sqrt{\frac{rs}{r+s}} \ln \frac{(x_{i1}x_{i2}\dots x_{ir})^{1/r}}{(x_{i1}x_{i2}\dots x_{is})^{1/s}} \quad (\text{Eq. 7})$$

where r and s are the number or parts marked with +1 and -1 in the respective SBP. The process for each observation ends when the last balance between the two final parts is estimated. Therefore, the number of *ilr* coordinates per observation is always D-1 (Pawlowsky-Glahn et al., 2015). There are many possibilities to define such an orthonormal basis system, even in case of SBP. For this reason, *ilr* is considered as a *class* of coordinates and is also common to refer to orthonormal (logratios) coordinates (Filzmoser et al., 2018). In this work, the SBP was built following the criterion of successively using as the numerator of Eq.7 the component with the highest logratio and *clr* variances, mainly for those of considerable values, and the rest in the denominator, condensing then most of the variance in the first balances. The calculations were made observation by observation for all the parts of the FA matrix. Therefore, assuming that C18:3 were the FA with the highest variance, the first balance (z_1) for the first observation would be:

$$z_1 \equiv \sqrt{\frac{1 \cdot 10}{1 + 10}} \ln \frac{(C18:3)^{1/10}}{(C14:0 \cdot C16:0 \cdot C16:1 \cdot C17:0 \cdot C17:1 \cdot C18:0 \cdot C18:1 \cdot C18:2 \cdot C20:0 \cdot C20:1)^{1/10}}$$

where the acronyms of FAs stand for the values of their parts.

Similarly, assuming C17:1 were the FA with the second highest variance, the second balance (z_2) for the first observation would then be:

$$z_2 \equiv \sqrt{\frac{1 \cdot 9}{1 + 9}} \ln \frac{(C17:1)^{1/1}}{(C14:0 \cdot C16:0 \cdot C16:1 \cdot C17:0 \cdot 1 \cdot C18:0 \cdot C18:1 \cdot C18:2 \cdot C20:0 \cdot C20:1)^{1/9}}$$

where the C18:3 FA (already used as the numerator in the first balance) has been removed from the calculations.

The process continued until the D-1 and D parts of the first observation are replaced with the last balance formed from the last two parts of the FA matrix, which is then estimated as:

$$z_{D-1} \equiv \sqrt{\frac{1 \cdot 1}{1 + 1}} \ln \frac{(C18:1)^{1/1}}{(C18:2)^{1/1}}$$

The process was repeated, observation by observation, until the 11 FAs (FA matrix) were substituted (observation-wise) with the 10 corresponding balances in the full dataset. These coordinates in the Euclidean real space were later subjected to

standard multivariate statistics. However, there is still some controversy on the convenience of this transformation since, except in the case of clear meaning of the balance logratio, interpretation of the results is frequently not straightforward, even after back-transformation (Greenacre, 2018).

Besides, the sequential binary partition and their corresponding balances may be represented in a system of coordinates in the form of a tree structure, usually known dendrogram, which allows an explicit interpretation of balances. In it, the vertical bars going up from each balance of this coordinate axis represent the variance of that specific coordinate, and the contact point is the coordinate mean (Pawlowsky-Glahn & Egozcue, 2006, 2011).

Summary compositional statistics, variation array, compositional biplot, compositional dendrogram as well as *clr* and *ilr* transformations were obtained using the CoDaPack software (Department of Computer Science and Applied Mathematics, University of Girona, Spain, v2.01.14).

Readers interested in a detailed explanation of the theory and application of the CoDa statistics should consult specialised texts (Pawlowsky-Glahn et al., 2011; Pawlowsky-Glahn et al. 2015; Van den Boogart & Tolosana Delgado, 2013; Filzmoser, Hron, & Templ 2018). Ros-Freixedes & Estany (2014) also includes an interesting overview of the basic CoDa tools.

2.4.3 Standard statistical analysis

The standard multivariate techniques (hierarchical cluster (HCA) and discriminant (DA) analyses) were applied to the full conventional FA dataset, the subcomposition of the major FAs (subcomposition.AFA), and their corresponding *ilr* transformed data (coordinates). The analyses were conducted, using centred and standardised data, under the following conditions:

-HCA: grouping of observations according to seasons, *montanera* length, and sampling location; Euclidean distance (dissimilarity); Wards' grouping methods; and automatic truncation (Entropy). The comparison between the results from the CoDa

analysis and standard statistics were based on the number of groups formed and the observations assigned to them.

DA: explanatory variables, FA matrix; dependent variables, season, *montanera* length, and sampling location; equal prior probabilities, significance level, $p=0.05$; tolerance, 0.0001; stepwise selection (p to enter and remove 0.05 and 0.10, respectively), except when applying bootstrapping; comparison among the within-class covariance matrices, Box test; comparison of the mean vectors of classes, Wilks' Lambda test; results included ,confusion matrix for the estimation sample and cross-validation; and, if applicable, display of centroids and confidence ellipses. When applying bootstrapping (which included all parts in the full dataset, subcomposition.AF and their *ilr* coordinates), the following specifications were used: simple sampling method, confidence interval level=0.95%, number of samples=1000. The results from using CoDa analysis and standard statistics were compared in this case by studying the proportion of success in the confusion matrix for the estimation sample and cross-validation, according to factor levels.

The software used was XLSTAT v.2016 (Addinsoft, Paris, France, v. 2017), for the standard multivariate analysis, and IBM SPSS Statistics for Windows, v.22 (IBM Corp., Armonk, N.Y., U.S.A), for the bootstrapping DA. The R package robCompositions was also applied for DA (Templ et al., 2011).

3. RESULTS AND DISCUSSION

3.1 Summary statistics

A first difference between the application of the standard and CoDa statistics to a dataset refers to the summary statistics since the arithmetic mean and standard deviation fit with the Euclidean real space. Instead, in CoDa, the central value and dispersion of the components are evaluated by their geometric means and the appropriate percentiles (Pawlowsky-Glahn et al., 2015; Filzmoser et al., 2018). Differences in the values found for the means (standard statistics) and central values (CoDa) were scant and affected mainly to the second decimal place while the same value in both types of data was observed in only C14:0, C17:0, and C20:0 (Table 1, for

season 2003). The ranges in the standard and CoDa (between the 25 and 75 percentiles) statistics show a limited variability within season 2003. Similar trends were observed for 2004 and 2005 well as for *montanera* length and sample locations and their respective levels, indicating a limited effect of factors on the fat compositions.

3.2 Variation array

The logratio variability according to components is the so-called variation array (Aitchison, 1986). As this matrix is symmetric with respect to the diagonal, it includes not only the logratio variances (upper diagonal) but the logratio means (lower diagonal). Besides, the *clr* and total variances are usually added as a last column on the right.

Therefore, the matrix includes three interesting pieces of information (Table 2):

i) The variance of the logratios of each component with respect to the other ones (Table 2). Those with the highest variability are expected to be the most appropriate for segregating the Iberian pig fat compositions according to season, *montanera* length, and sampling location since, apart from analytical errors, they may also include the variability due to the hypothetical differences between the levels of the factors (Egozcue & Pawlosky-Glahn, 2005; Pawlowsky-Glahn & Egozcue, 2011). In this case, the greatest potential segregation power could be expected from C18:3, C17:1, and C20:0.

ii) the *clr* variance for each FA when added as the last column on the right. It may be obtained directly or through the sum of the variances of all logratios divided by two times the number of parts in the composition (Thió-Henestrosa & Daunis-i-Estadella, 2011). Then, the parts with the largest logratio variances are also associated with those having the greatest *clr* variances (C18:3, C17:1, C20:0, C16:1, and C20:1) (Table 2). However, their interpretations should be cautious since they may include, as above-commented, not only the between but also the within level variability. The sum of *clr* variances also yields the total variance.

and iii) the means of the logratios of each component with respect to each other, which reflect the relative proportion between FAs in the compositions.

The variability observed for the unsaturated FAs agrees with results reported by Ros et al. (2011) and Ros-Freixedes & Estany (2014), who regardless of the origin of the sample, found high variabilities for logratios involving C20:4. Also, Narváez-Rivas et al. (2009) observed the largest differences among samples for linolenic acid, an essential FA for pigs, supplied exclusively by the feed. The supplementation of the pig diet with linseed (Bee, Jacot, Guex, & Biolley, 2008) modified the unsaturated FA profiles of the muscles and subcutaneous fat, increasing the levels of linolenic and eicosatrienoic acids.

Regarding logratios, a first observation shows the presence of both positive and negative signs which reminds the Euclidean real space. Furthermore, their direct analysis is claimed to lead to somewhat accurate results (Greenacre, 2018). As expected, the low mean (-5.6670) for $\ln(C20:0/18:1)$ or the high mean (5.3534) for $\ln(C18:1/C17:1)$, indicate that, as usual, C18:1 is in a sensibly higher concentration in pigs than C20:0 or C17:1, which are barely at detection limits. Therefore, the variation array provides also valuable information on the relative abundance of each FA which may be of interest for specific studies. For example, the selection of diets for the production of healthy pigs (low in saturated FAs) would be based on the logratios involving C18:1, C18:2, and C18:3 over the rest of saturated FAs, or on the logratio (saturated/unsaturated fat).

Since most of the logratio variances were related to unsaturated FAs in low concentrations and potential high variability, it was considered of interest to re-analyse the data in their absence and study the effect on the results. Therefore, this dataset (subcomposition.AFA) included C14:0, C16:0, C16:1, C18:0, C18:1, C18:2, and C20:1, the most Abundant FA). In its variation array (Table S1), the total variance decreased, but the greatest logratio and clr variances were for the monounsaturated C20:1 and C16:1 as well as C18:0; that is, most of the possible effects were still associated with unsaturated FAs. The largest logratio mean was for $\ln(C18:1/C14:1)$, also between monounsaturated FAs.

3.3 Graphical analyses

Ternary plot

This plot may show overall trends. Theoretically, the parts (a maximum of four) with the largest variance in the variation array or *clr* variance would provide the greatest segregation power. However, neither the full conventional centred dataset nor subcomposition.AFA showed any trend (Fig. S1A-B); thus, regardless of the way of expressing the composition, the graphs suggest quite a similarity among samples.

Compositional biplot

The biplot, adapted to CoDa by Aitchison & Greenacre (2002), is a convenient approach to present the compositional matrix, using a rank-2 approximation. The distribution of the components in the covariance biplot based on the full dataset was the same regardless of the grouping factor; consequently, only that regarding seasons, which explained 67.58% total variance, is displayed (Fig. 1). The centre of the diagram is the centroid of the compositions, the 11 FAs are represented by rays, and the cases by different colour circles associated with seasons. The largest links ($\approx$ the log ratio variance) were observed for $\ln(C18:3/C17:1)$, $\ln(C18:3/C20:0)$, and $\ln(C18:3/C18:0)$, which might include the variability due to possibly analytical errors. When two vertices coincide, or nearly so, this means that the variance of their logratios is close to zero, their ratio constant, and can be assumed that they carry redundant information (e.g. $\ln(16:1/17:0)$ and $\ln(C:18:0/C20:0)$). Short arrows (C14:0 or C18:1) may indicate a poor representation on this plane, a situation that is probable in this study since including the third PC, the explained variance increased up to $\approx 82\%$. The presence of collinear vertices may be related to subcompositions with one-dimensional variability biplot. In this way, C17:1, C17:0, C16:1, and C18:1 follows a linear trend as well as C18:3, C18:1, and C20:1. Besides, angles provide information on the correlation of subcompositions; the links between the $\ln(C18:3/C18:0)$ and $\ln(C17:1/C20:0)$ form an angle close to 90° , indicating the absence of correlation between these two logratios. Similarly, subcompositions including parts C18:3, C18:0 and C20:0 vs C17:1, C17:0, and C16:1. Thus, the information on variability provided by the biplot is in agreement with that of the variation array (Table 2). Most of the samples from season 2005 were

placed on the third and fourth quadrant of the covariance and form biplots, meaning that, possibly, the composition of season 2005 (all of them from M2) could be a bit peculiar and related to high levels of C18:3 (mainly) or C17:1 since no segregation in the biplot according to *montanera* lengths was observed (data not shown). The biplot was also very informative for exploring the relationships among the compositions from different muscles and the subcutaneous backfat of some of them (Ros-Freixedes & Estany, 2014), with the most outstanding logratios being between C20:4 vs the other polyunsaturated FAs (C18:3, C18:2, and C20:2) and C20:0. Interestingly, the biplot segregated intramuscular from subcutaneous fat. Regarding muscles, saturated and monounsaturated fat was always on the opposite side than polyunsaturated fat, as also observed in this work. Furthermore, Ros-Freixedes & Estany (2014) related the changes in logratio C18:0/C16:0, (C16:1/C16:0) and C18:1/C18:0) with the elongase and Δ^9 desaturase activities, interpreting such transformations as biologically consistent and in line with the findings of other authors.

When the biplot was built using the subcomposition.AFA, the relationships among components were, as expected, different. The greatest variances were also related to unsaturated FAs ($\ln(C16:1/C20:1)$ and $\ln(C16:1/C18:0)$) while the rest of acids were associated with the third PC. In any case, the segregation of samples was not improved (Fig. S2), indicating that the removal of FAs present in low concentrations had a limited interfering effect, if any, on the biplot segregation capacity.

3.4 Working in coordinates

A characteristic of the compositional data is that they may be transformed into coordinates in the Euclidean space where they can be subjected to standard statistical analysis without restriction (Egozcue et al., 2003).

Estimation of the ilr coordinates

The criteria for the selection of balances should be based on previous information on the components, the objectives of the work, or on the variation array and the biplot variances (Pawlowsky-Glahn & Egozcue, 2011). Ros-Freixedes & Estany

(2014) initiated the SBP according to the FA saturation but selected for the following partitions those FAs which led to interpretable results. As in this work, the interest was particularly focused only on the influence of the factors (and their levels) on the overall FA profiles, the magnitude of logratios variances was considered for the SBP. Therefore, the matrix for the sequential binary partition (Table S2), and the balances were deduced, sample by sample, from the FA matrix, following this criterion and the steps described in the CoDa statistical overview.

The averages of these balances (Table 3) reflect the relative abundance of their respective FAs in the numerators over those in the denominators. Only balance (z_{10}) showed a clear abundance of C18:1, the most abundant FA in pigs, over C18:2. Other two were rather equilibrated (z_4 and z_6), although with a different sign, while in the rest of balances the concentration (or geometric means) in the denominator were more abundant, particularly those including C18:1, the predominant FA in pigs. Concerning variances, their values were roughly decreasing as the order increased, with the four first balances accounting for the highest values, indicating that, according to their selection criterion, they could be the most qualified for the segregation of samples, although any interpretation based exclusively on variances should be treated with caution.

In this work, the balances were also represented in a system of coordinates in the form of a tree structure (dendrogram) for a more detailed study than that deduced just from their means. The dendrogram can be built according to the overall balances (Table 3) or the groups of the specific factors, displaying a boxplot for each level (Thió-Henestrosa & Daunís-i-Estadella, 2011). The plots as a function of season, *montanera* length, or sampling location showed similar trends; so, only the dendrogram regarding the season and its levels, which had the highest variability, is discussed (Fig. 2). Each coordinate is represented in the horizontal axis, which limits are fixed arbitrarily by the user but the same for each coordinate (Pawlowsky-Glahn & Egozcue, 2011), indicating its situation the relative prevalence of the FAs in the numerator over those in the denominator. In this case, the three boxes with their 0, 25, 55 (median), 75 and 100% percentiles below the horizontal line represent the seasons. Significant effects between these would be reflected as displaced non-overlapping boxes. In general,

most of the balances are displaced from the centre of the horizontal line to the right (Fig. 2). These positions indicate that, regardless of the levels of the factors, the presence of the FA in the numerator (e.g. C18:3, C17:1, C20:0, or C21:1, on the right of the balance plot) were in lower proportion than the geometric means of those in the denominator (on the left). The only exceptions were z_{10} which was positive, and z_4 and z_6 which, apart from the sign (negative and positive respectively) were quite centred. The overlapping of the season boxes in most of the balances means a limited effect of this factor on the FA profile of pig fat, except for z_1 (associated to C18:3) and, possibly, z_3 which might then be the two main contributors to their differentiation. The dendrogram (Fig. 2) may also illustrate on the relative presence of specific FA like C18:1 which was always more abundant than C18:2 and their logratio was almost constant due to its low variance (short vertical lines). Regardless of the season, the logratio $\ln(C14:0/C16:0)$ is also almost constant (low variance), with the C14:0 more abundant regardless of the season. Besides, the geometric mean of C14:0, C16:0, C17:0, and C18:0, all them saturated, is always less abundant than that of C16:1, C18:1 and C18:2 (Fig. 2). Any other relationship could be deduced similarly. The vertical lines above the overall and the specific balances for seasons indicate their corresponding variances, which are linked to their respective levels by the same colour (Fig. 2). The outstanding overall high variance of z_1 (mainly) and z_3 , could be related to the changes of C18:3 and C20:1 associated with seasons. However, z_2 variability could be mainly associated with the within-group variance due to the practical overlapping of seasons. Hence, in general, the variance values (within and, possibly, between groups) of only a few first balances could be qualified for the possible segregation between seasons. The effects of *montanera* length and sampling location were even lower since their balances had inferior variances. The fat composition of these pigs may then be strongly associated with their usual diet and metabolism and only partially influenced by season. The effects of *montanera* length and sampling location would be even lower. In any case, a detailed study of the biotransformations of the Iberian pig FAs was not a priority objective of this study.

When the balances were estimated based on the subcomposition.AFA, the sequential binary partition (Table S3) and the balance estimations led to different

average values and lower variances (Table S4), with a distinct dendrogram (Fig. S3). However, the greatest variances were also related to C20:1 (z_1), C16:1 (z_2), and C18:0 (z_3), mostly unsaturated and in low concentration but only z_3 showed a certain segregation capacity between seasons (displaced boxplots). In any case, the balances for each sample are their *ilr* transformed values (coordinates) and, together, replaced the full conventional dataset in the further multivariate analysis of subcomposition.AFA.

3.5 Segregation of samples by application of standard multivariate techniques

3.5.1 Results from Hierarchical Clustering Analysis

When the HCA was performed using the full conventional dataset and its coordinates, the analysis led to the formation of 4 and 3 groups (automatically established), respectively (Table S5). However, the samples from the different levels of the season, *montanera* type, and sampling location were never well classified even in case of using the *ilr* coordinates. This was because regardless of the dataset used, the FA profiles of the groups were fairly similar (Fig. S4 A-B).

Repeating the study with the subcomposition.AFA, regardless of the type of dataset used, the samples were grouped into 3 clusters, but the assignation of samples to them also was erratic (Table S6). In this case, the FA profiles of the subcomposition.AFA dataset did not reveal any difference among groups (Fig. S4 C) while was found, although limited, between some of their balances (Fig. S4 D), mainly for those derived from C20:1, C18:1, and C18:0.

3.5.2 Discriminant analysis

The analysis was also performed using the full conventional dataset, the subcomposition AFA, and their coordinates. The results are presented graphically as well as confusion matrices for both the proportions of correctly classified samples (used for the development of the model) or after cross-validation. As an additional check, the Fisher discriminant analysis was also applied (Templ et al., 2011).

For seasons, DA led to a certain degree of segregation but also to an evident overlapping of cases (Fig. S5, for only season 2003 samples) and a relatively close

situation of their centroids, regardless of working with the full conventional dataset (Fig. S5 A) or their coordinates (Fig. S5 B), but with the samples and centroids in opposite positions. An apparent more reliable analysis using bootstrapping showed an overall 69/70% (conventional/coordinates) correct assignments and 68/70, 66/67, and 82/78% for seasons 2003, 2004, and 2005, respectively, with proportions in cross-validation, usually a couple of points below (Table 4). Applying the CoDa Fisher discrimination test with the option “robust” (Templ et al., 2011), a proper segregation of 2005 season was achieved and a fairly good discrimination between season 2003 and 2004, although with some overlapping and a few samples (mainly from 2004 season) far from the groups (Fig. 3). Therefore, the FA profiles for the different seasons were not so different as for providing complete segregation of samples by DA, although the efficiency improved slightly ($\approx 1\%$ overall) using *ilr* coordinates and, mainly, applying robust CoDa techniques.

In case of the application of the same procedures to the conventional subcomposition.AFA and its *ilr* coordinates, the proportion of overall correct assignments (61%) according to seasons markedly decreased (season 2003 and 2004) or remained similar (season 2005) (Table S7) while the application of CoDa Fisher’s procedure led to certain segregation of samples as a function of season but also to abundant overlapping in the centre of the plot (Fig. S6). Hence, the use of the subcomposition.AFA, with only the major FAs, never improved the segregation power.

Regarding *montanera* length and sampling location (two levels), there was only one Factor. Using the full conventional or *ilr* coordinate datasets, DA led to some separation between the centroids of the two levels but with a manifest overlapping of cases (Fig. S7 A-B, for *montanera* length, and S8 A-B, for sampling location). However, DA with bootstrapping and both data set resulted in confusion matrices with similar overall correct assignments (71/70%, conventional/coordinates) for *montanera* length, with an equilibrated assignment between levels (73/69% and 70/70%, for M1 and M2), with validation values always below these percentages (Table 4). The classification regarding sampling location showed similar correct overall assignments (68/67%) using conventional dataset and coordinates and 2-4 points less in cross-validation (Table 4). For these factors, the Fisher procedure segregated several long *montanera* samples

(M2) correctly but was less efficient for classifying according to sampling location (Fig.S9 B). Therefore, the expected differences in the pig fat compositions between the two *montanera* lengths and sampling locations were never enough to produce total segregation between their levels. However, Narváez-Rivas et al. (2009) found significant differences for the majority of FAs between the two anatomical locations, linking the changes to the key enzymes involved in the lipogenesis; in the study, the season was also influential, increasing the ratio oleic/stearic in case of low natural resources for feeding, associated with the increase of the stearyl-CoA desaturase.

DA using the subcomposition.AFA and their coordinates led to similar results than those obtained with the full dataset (Fig. S7 C-D vs A-B, Fig. S8 C-D vs A-B) but with bootstrapping the overall correct assignments to seasons were (68/66%, conventional/coordinates) with M1 slightly better assigned ($\approx 70\%$) than M2 ($\approx 66\%$) with also close values in cross-validation (Table S7). Approximately similar correct assignments were observed for sampling location (66/66%, conventional/coordinates, for overall), with a range of approximately $\pm 1\%$, and slightly lower percentages for cross-validation. The Fisher DA, robust option, resulted in 74% and 75% matched pairs for *montanera* length and sampling location, respectively but did not plot for segregation between only two levels. In general, subcomposition.AFA led then to slightly worse results than the full FA dataset particularly in overall, 2003, and 2004 seasons.

Ros et al. (2011) also applied DA for segregating between the fat from different muscles (IMF) and the subcutaneous tissue (SF), performing the analysis with the conventional percentages and clr-transformed data (coordinates). There was an almost successful assignment of samples to SF and IMF; however, the overall classification of the samples according to the different muscles of origin was less effective, 63/83% and 70/76% for the conventional/clr coordinate datasets, but in favour of the last one. The use of only the two most influential FAs, in general, reduced the efficiency for segregating SF and IMF as well as the correct assignments among the different muscles, although coordinates also improved the discrimination with respect to the percentages. Their results agree with those found in this work, where the best results were obtained with the full dataset.

To summarise, this work explored the effect of diverse factors (season, *montanera* length, and sampling locations) on the FA profiles of Iberian pigs using CoDa techniques in comparison to following the standard statistical procedures. For this purpose, a full dataset (in weight percentages), with all FAs estimated, and a subcomposition.AFA composed of only the major FA were subjected to exploratory CoDa analysis, and tested, together with their corresponding coordinates, for their segregating power. Variation array and biplot of the database (371 samples) showed that the greatest variances for the full conventional dataset were observed for $\ln(C18:3/C17:1)$, $\ln(C18:3/C20:0)$, $\ln(C18:3/C18:0)$, $\ln(C17:1/C20:0)$ and $\ln(C17:1/C18:0)$ involving mainly to C18:3 and C17:1, situation that was also confirmed in the dendrogram deduced using these acids for the first orders of SBP. These findings are in agreement with Ros-Fleixedes & Estany (2014) who also observed great variability for logratios derived from C20:4, situated in the biplot opposed to other polyunsaturated FAs. In other studies, a diet supplemented with linseed modified the FA composition of diverse muscles, and the subcutaneous fat increased the linolenic and eicosatrienoic acids concentrations in all tissues (Bee et al., 2008). The results found in this work are in line with the hypothesis that the polyunsaturated FAs are usually those who experiment the greatest relative modifications (logratio variances), even considering the possibility of errors in their analyses. This is also in agreement with the observation that the segregation power of the subcomposition.AFA, with the major saturated or monounsaturated FAs, led to somewhat poorer classification. Other factors such as pig age or fat content may also be influential; notably, the activities of elongase and Δ^9 desaturase enzymes can markedly modify the long chain FAs relationships (Ros-Fleixedes & Estany (2014).

In this work, the season and *montanera* showed a limited effect on the Iberian pig FA profiles although a report (Bee et al., 2008) has shown that the concentration of FAs in tissues is altered by dietary fats differing in the degree of unsaturation and the dietary energy level. Linolenic acid in swine adipose and lean tissues derives exclusively from dietary sources and the concentration in tissues depends on the intake and overall fat deposition. However, it was less efficiently incorporated into adipose tissues when the feed was low in energy, and the dietary fat was rich in saturated FAs.

Finally, not great differences between the FA profiles of the two sampling locations used in this study were found. However, several authors (Narváez-Rivas et al., 2009; Ros et al., 2011; Ros-Fleixedes & Estany, 2014) reported some differences in the FA profiles among muscles and tissues. Monziols et al. (2007) found a gradient of decreasing unsaturation from the outer layer of subcutaneous adipose tissue to the inner layer. In Celta pigs, the sampling location also showed significantly higher saturated fat in the intramuscular than in the ventral fat, with the highest variations found in the linoleic and eicosadienoic acids. Interestingly, discriminant analysis based on the content of the C18:0, C18:2, and C20:2 differentiated between the two locations correctly (Franco et al., 2006). Possibly, a study including a larger number of sampling points could help for a more detailed investigation of this aspect (Ros-Fleixedes & Estany, 2014).

In general, CoDa provided interesting information through the exploratory analysis, but its application only led to slightly better segregation. The use of merely the major FAs did not improve the results; therefore, the factors of the design could have affected all FAs similarly. Since confusion matrices obtained from conventional and coordinates agreed reasonably, but always led to a considerable proportion (~30%) of erroneous assignments and such values can also be found in the literature, this could mean that, maybe, some other influential factors should be investigated or that the animals have mechanisms, possibly related to genotype, for adaptation to the environmental conditions. Besides, the improved results obtained using robust CoDa methods suggest the convenience of their wider application in the pig fat studies.

Conflict of interest

The authors declare no conflict of interest

Acknowledgements

This work was supported by the Spanish Government (project AGL2010-15494/ALI, partially financed by European regional development funds, ERDF), and Junta de Andalucía (through financial support to group AGR-125 and AGR-283).

References

- 648 Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*, reprinted in 2003
649 with additional material by The Blackburn Press. New Jersey (USA): Chapman & Hall
650 Ltd.
- 651 Aitchison, J. & Greenacre, M. (2002). Biplots for compositional data. *Journal of the*
652 *Royal Statistical Society, Series C (Applied Statistics)*, 51, 375-392.
- 653 Andrés, A.I., Cava, R., Ventanas, S., Muriel, E., & Ruiz, J. (2007). Effect of salt content
654 and processing conditions on volatile compounds formation throughout the ripening of
655 Iberian ham. *European Food Research and Technology*, 225, 677-684.
- 656 Bee, G., Jacob, S., Guex, G., & Biolley, C. (2008). Effects of two supplementation levels
657 of linseed combined with CLA or tallow on meat quality traits and fatty acid profile of
658 adipose and different muscle tissues in slaughter pigs. *Animals*, 2-5, 800-811.
- 659 Boletín Oficial del Estado (B.O.E.) (2004). Orden PRE/3844/2004 por la que se
660 establecen los métodos oficiales de toma de muestras en canales de cerdos ibéricos y
661 el método de análisis para la determinación de la composición de ácidos grasos de los
662 lípidos totales del tejido adiposo subcutáneo de cerdos ibéricos (pp. 38770-38775).
663 Madrid (Spain):B.O.E.
- 664 Buccianti, A., Mateu-Figueras, G., & Pawlowsky-Glahn, V.E. (Eds) 2006. *Compositional*
665 *Data Analysis in the Geosciences: From Theory to Practice*. Special publication, vol. 264.
666 London (U.K.): Geological Society.
- 667 De Pedro, E., Casillas, M., & Miranda, C. M. (1997) Microwave oven
668 application in the extraction of fat from the subcutaneous tissue
669 of Iberian pig ham. *Meat Science*, 45, 45–51.
- 670 Dirinck, P., Van Opstaele, F., & Vandendriessche, F. (1997). Flavour differences
671 between northern and southern European cured hams. *Food Chemistry*, 59, 511-521.
- 672 Egozcue, J.J. & Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in
673 Compositional Data Analysis. *Mathematical Geology*, 37, 795-828.

- 674 Egozcue, J.J., Pawlowsky-Glahn, V., Mateo-Figueras, G., & Barceló-Vidal, C. (2003).
675 Isometric logratio transformations for compositional data analysis. *Mathematical*
676 *Geology*, 35, 279-300.
- 677 Filzmoser, P., Hron, K., & Templ, M. (2018). Applied Compositional Data Analysis, with
678 worked examples. Cham (Switzerland): Springer Nature Switzerland AG.
- 679 Flores, M., Grimm, C.C., Toldrá, F., & Spanier, A.M. (1997). Correlations of Sensory and
680 Volatile Compounds of Spanish "Serrano" Dry-Cured Ham as a Function of Two
681 Processing Times". *Journal of Agriculture and Food Chemistry*, 45, 2178-2186.
- 682 Franco, I., Escamilla, M.C., García, J., García Fontán, M.C., & Carballo, J. (2006). Fatty
683 acid profile of the fat from Celta pig breed fattened using a traditional feed: Effect of
684 the location in the carcass. *Journal of Food Composition and Analysis*, 19, 792-799.
- 685 Gallardo, E., Narváez-Rivas, M., Pablos, F., Jurado, J.M., & León-Camacho, M. (2012).
686 Subcutaneous fat triacylglycerols profile from Iberian pigs as a tool to differentiate
687 between intensive and extensive fattening systems. *Journal of Agriculture and Food*
688 *Chemistry*, 60, 1645-1651.
- 689 Greenacre, M. (2011) Compositional data and correspondence analysis. In V.
690 Pawlowsky-Glahn, & A. Buccianti (eds), *Compositional Data Analysis: Theory and*
691 *Application*. London (U.K): Wiley & Sons.
- 692 Jurado, A., García, C., Timón, M.L., & Carrapiso, A.I. (2007). Effect of ripening time and
693 rearing system on amino acid-related flavour compounds of Iberian ham. *Meat*
694 *Science*, 75, 585-594.
- 695 Kraft, A, Graeve, M., Janssen, D., Greenacre, M., & Falk-Petersen, S. (2015). Artic
696 pelagic amphipods: lipid dynamics and life strategy. *Journal of Plankton Research*,
697 37,790-807.
- 698 León-Camacho, M., Viera-Alcaide, I., & Vicario, I.M. (2007). Acorn (*Quercus* spp.) fruit
699 lipids: saponifiable and unsaponifiable Fraction: A detailed study. *Journal of the*
700 *American Oil Chemists' Society*, 81, (2004) 447-453.

- 701 Monziols, M., Borneau, M., & Kouba, M. (2007). Comparison of the lipid content and
702 fatty acid composition of intermuscular and subcutaneous adipose tissues in pig
703 carcasses. *Meat Science*, 76, 54-60.
- 704 Narváez-Rivas, M., León-Camacho, M., & Vicario, I. M. (2009). Fatty acid and
705 triacylglycerol composition of the subcutaneous fat from Iberian pigs fattened on the
706 traditional feed: "montanera". Effect of anatomical location and length of feeding.
707 *Grasas y Aceites*, 60, 238-247.
- 708 Narváez-Rivas, M., Ríos, J.J., Arteaga, J.F., Quilez, J.F., Barrero, A.F., & León-Camacho,
709 M. (2008) Determination of ent-kaurene in subcutaneous fat of Iberian pigs by gas
710 chromatography multi-stage mass spectrometry with the aim to differentiate between
711 intensive and extensive fattening system. *Analytica Chimica Acta*, 624, 107-112.
- 712 Pawlowsky-Glahn, V., & Egozcue, J.J. (2006). *Compositional data and their analysis: An*
713 *introduction*, Special publication, vol. 264. London (UK): Geographical Society.
- 714 Pawlowsky-Glahn, V., & Egozcue, J.J. (2011). Exploring compositional data with the
715 CoDa-dendrogram. *Austrian Journal of Statistics*, 40, 103-113.
- 716 Pawlowsky-Glahn, V., & Buccianti, A. (eds). 2011. *Compositional Data Analysis. Theory*
717 *and application*. Chichester (U.K.): John Wiley & Sons Ltd.
- 718 Pawlowsky-Glahn, V., Egozcue, J.J. Tolosana-Delgado, R. (2015). *Modeling and analysis*
719 *of compositional data*. Chichester (U.K): John Wiley & Sons Ltd.
- 720 Ros, R., Reixach, J., Tor, M., Estany, J. (2011). Exploratory data analysis for fatty acids
721 composition in pig meat. In J.J. Egozcue, R. Tolosana Delgado, & M.I. Ortego (eds),
722 *Proceedings of the 4th International Workshop on Compositional Data Analysis* (pp.1-5).
723 San Feliu de Guixols (Girona, Spain): CODAWORK'11.
- 724 Ros-Freixedes, R., & Estany, J. (2014). On the compositional analysis of fatty acids in
725 pork. *Journal of Agricultural, Biological, and Environmental Statistics*, 19, 136-155.
- 726 Rueda Sabater, L. & Diéguez Garbayo, E. (2007). Manual del Cerdo Ibérico. 2nd Ed.
727 Zafra, Badajoz (Spain): Asociación Española De Criadores Cerdo Ibérico (AECERIBER).

- 728 Ruiz, J., Ventanas, J., Cava, R., Andrés A., & García, C. (1999). Volatile compounds of
729 dry-cured Iberian ham as affected by the length of the curing process. *Meat Science*,
730 52, 19-27.
- 731 Sharma, N., Gandemer, G., & Goutefongea, R. (1987). Comparative lipid composition of
732 porcine muscles at different anatomical locations. *Meat Science*, 19, 121-128.
- 733 Templ, M., Hron, K., & Filzmoser, P. (2011). robCompositions: An R-package for Robust
734 Statistical Analysis of Compositional Data. In V. Pawlowsky-Glahn, & A. Buccianti (Eds),
735 *Compositional Data Analysis: Theory and Application* (pp 341-355). Chichester (U.K.):
736 John Wiley & Sons Ltd.
- 737 Thió-Henestrosa, S., Daunis-i-Estadella, J. (2011) Exploratory analysis using
738 CoDaPack3D. In V. Pawlowsky-Glahn, & A. Buccianti (eds), *Compositional Data*
739 *Analysis. Theory and application* (p. 330-340). Chichester (U.K.): John Wiley & Sons
740 Ltd.
- 741 Van den Boogart, K.G. & Tolosana Delgado, R. (2013). Analyzing compositional data
742 with R. Berlin Heidelberg (Germany): Springer-Verlag.
- 743 Viera-Alcaide, I., Vicario, I.M., Escudero-Gilete, M.L., Graciani Constante, E., & León
744 Camacho, M. (2008). A multivariate study of the triacylglycerols composition of the
745 subcutaneous adipose tissue of Iberian pig in relation to the fattening diet and
746 genotype. *Grasas y Aceites*, 59, 327-336.

Legend of figures

Figure 1. Covariance biplot of Iberian pig fat, based on the *clr* transformed full conventional dataset, according to seasons.

Figure 2. Balance dendrogram of the Iberian pig fat, based on the sequential binary data partition of the full conventional dataset, according to seasons.

Figure 3. Segregation of samples using the Fisher's DA procedure (option robust), based on the full fatty acid profile of the Iberian pig fat. Projection of the samples onto the first two Fisher's scores according to seasons.

Table 1. Summary standard and CoDa statistics. Central tendency and dispersion for season 2003 (189 samples). Full conventional dataset (fatty acid concentrations expressed as weight percentages).

Fatty acid		Summary statistics					
		Classical			Compositional		
		Mean	SD	Min.-Max.	Central value	25 percentil	75 perce
Myristic	C14:0	1.16	0.13	0.81-1.55	1.16	1.07	1.24
Palmitic	C16:0	20.39	1.11	17.82-23.56	20.40	19.61	20.91
Palmitoleic	C16:1	2.00	0.32	1.27-2.89	1.98	1.76	2.20
Margaric	C17:0	0.29	0.05	0.17-0.43	0.29	0.26	0.32
Heptadecenoic	C17:1	0.27	0.07	0.13-0.49	0.26	0.23	0.31
Stearic	C18:0	10.11	1.11	7.46-13.81	10.08	9.23	10.83
Oleic	C18:1	55.30	1.61	51.64-59.04	55.40	54.15	56.48
Linoleic	C18:2	8.14	0.72	6.54-10.14	8.13	7.70	8.65
Linolenic	C18:3	0.53	0.13	0.32-0.94	0.52	0.42	0.62
Behenic	C20:0	0.20	0.03	0.15-0.28	0.20	0.18	0.22
Gadoleic	C20:1	1.60	0.25	1.01-2.27	1.58	1.42	1.74

Notes: SD, standard deviation; Min., minimum; Max., maximum,

Table 2. CoDa variation array matrix, including logratio variances, *clr* variance, and mean of the fatty acid logratios for the full dataset. Sample size: 371

		Variance ln(Xi/Xj)											clr varian ce
Xi\Xj		C14: 0	C16: 0	C16: 1	C17: 0	C17: 1	C18: 0	C18: 1	C18: 2	C18: 3	C20: 0	C20: 1	
C14: 0			0.008 4	0.017 5	0.043 2	0.055 2	0.030 3	0.014 6	0.016 9	0.074 2	0.044 2	0.037 9	0.0102
C16: 0	2.842 3			0.022 3	0.037 2	0.058 7	0.010 3	0.005 9	0.014 1	0.078 0	0.023 5	0.028 5	0.0052
C16: 1	0.526 2	- 2.316 1			0.045 9	0.039 7	0.053 0	0.027 2	0.031 1	0.106 5	0.064 9	0.058 4	0.0215
C17: 0	-	-	-				0.018	0.048	0.030	0.030	0.098	0.055	0.0199

0	1.406	4.248	1.932		1	2	7	9	1	7	5	
	0	2	1									
C17:	-	-	-	-	0.086	0.051	0.049	0.124	0.093	0.065		
1	1.502	4.345	2.028	0.096	8	2	1	3	8	3	0.0375	
	8	0	9	8								
C18:	2.099	-	1.573	3.505	3.602		0.019	0.032	0.100	0.014	0.036	
0	3	0.742	2	3	1		0	6	3	7	9	0.0184
	9											
C18:	3.850	1.008	3.324	5.256	5.353	1.751		0.007	0.067	0.025	0.019	
1	6	4	5	6	4	3		0	4	7	6	0.0035
C18:	1.955	-	1.429	3.361	3.458	-	-		0.047	0.040	0.030	
2	3	0.887	0	1	3	1	0.144	1.895	9	2	9	0.0064
		0					0	3				
C18:	-	-	-	0.641	0.738	-	-	-		0.108	0.102	
3	0.764	3.606	1.290	4	7	5	2.863	4.614	2.719	6	5	0.0616
	3	5	4				6	9	6			
C20:	-	-	-	-	-	-	-	-	-	-	-	
0	1.816	4.658	2.342	0.410	0.313	3.915	5.667	3.771	1.052		0.034	0.0251
	4	6	5	4	6	7	0	7	1		9	
C20:	0.295	-	-	1.701	1.797	-	-	-	1.059	2.111		
1	1	2.547	0.231	0	9	1.804	3.555	1.660	2	4	5	0.0206
		2	1			3	5	2				
Mean ln(X _i /X _j)												Total 0.2300 variance

Note: Outstanding logratio variances, *clr* variances, and mean logratios in bold.

Table 3. Overall balances and their variances after *ilr* transformation of the full conventional dataset. Sample size, 371.

Parameter	Balan ce 1 (<u>z</u> ₁)	Balan ce 2 (<u>z</u> ₂)	Balan ce 3 (<u>z</u> ₃)	Balan ce 4 (<u>z</u> ₄)	Balan ce 5 (<u>z</u> ₅)	Balan ce 6 (<u>z</u> ₆)	Balan ce 7 (<u>z</u> ₇)	Balan ce 8 (<u>z</u> ₈)	Balan ce 9 (<u>z</u> ₉)	Balan ce 10 (<u>z</u> ₁₀)
Mean	-	-	-	-	-	-	-	-	-	-
	1.381	2.305	2.910	1.042	1.606	1.074	2.009	2.478	1.940	1.3402
	2	5	2	6	2	4	8	6	6	
Variance	0.067	0.039	0.022	0.022	0.011	0.016	0.004	0.024	0.018	0.0035
	8	5	6	0	9	1	2	1	2	

Note: In parenthesis, symbols for the corresponding balances in the dendrogram

Table 4. Confusion matrix obtained after the classification of the samples by DA according to factor, using bootstrapping. Full conventional dataset and its *ilr* transformed coordinates.

<u>Conventional data</u>						<u>Coordinates (ilr transform)</u>			
Season						Season			
From/to	2003	2004	2005	Total	Correct assign. (%)	From/to	2003	2004	2005
2003	128 (125)	45 (46)	16 (18)	189	67.7 (66.1)	2003	133 (129)	43 (46)	13 (14)
2004	33 (35)	88 (85)	12 (13)	133	66.2 (63.9)	2004	32 (34)	89 (86)	12 (13)
2005	2 (3)	7 (7)	40 (39)	49	81.6 (79.6)	2004	4 (4)	7 (7)	38 (38)
Total	163 (163)	140 (138)	68 (70)	371	69.0 (67.1)	Total	169 (167)	139 (139)	63 (65)
<i>Montanera</i> length						<i>Montanera</i> length			
From/to	M1	M2	Total	Correct assign. (%)		From/to	M1	M2	
M1	63 (58)	23 (28)	86	73.3 (67.4)		M1	59 (56)	27 (30)	
M2	86 (87)	199 (198)	285	69.8 (69.5)		M2	86 (89)	199 (196)	
Total	149 (145)	222 (226)	371	70.6 (69.0)		Total	145 (145)	226 (226)	
Sampling location						Sampling location			
From/to	Hum	Rump	Total	Correct assign. (%)		From/to	Ham	Rump	
Hum	131 (125)	60 (66)	191	68.9 (65.4)		Ham	131 (126)	60 (65)	
Rump	60 (65)	120 (115)	180	66.7 (63.9)	32	Rump	62 (68)	118 (112)	

Total	191 (190)	180 (181)	371	67.7 (64.7)	Total	193 (194)	178 (177)
-------	--------------	--------------	-----	----------------	-------	--------------	--------------

Note: Results obtained using bootstrapping (n=1000). In parenthesis, cross-validation data. M1 and M2 refer to the two *montanera* lengths. Assign., assignation.

Highlights

-Formally, Compositional Data Analysis (CoDa) Analysis is appropriate to study Iberian pig fat

- Variation array and biplot showed the greatest variances for C18:3, C17:1, and C20:0
- Environmental factors explained only partially the changes in pig fat composition
- Besides montanera feeding, phenotype may play sensible role in pig fat composition
- Full fatty acid (FA) profile led to better /similar results than with the major ones.