Statistics and Operations Research Transactions, vol. 37, n. 1 (2013)

SORT 37 (1) January-June 2013, 3-18

Improved entropy based test of uniformity using ranked set samples

M. Mahdizadeh and N. R. Arghami*

Abstract

Ranked set sampling (RSS) is known to be superior to the traditional simple random sampling (SRS) in the sense that it often leads to more efficient inference procedures. Basic version of RSS has been extensively modified to come up with schemes resulting in more accurate estimators of the population attributes. Multistage ranked set sampling (MSRSS) is such a variation surpassing RSS. Entropy has been instrumental in constructing criteria for fitting of parametric models to the data. The goal of this article is to develop tests of uniformity based on sample entropy under RSS and MSRSS designs. A Monte Carlo simulation study is carried out to compare the power of the proposed tests under several alternative distributions with the ordinary test based on SRS. The results report that the new entropy tests have higher power than the original one for nearly all sample sizes and under alternatives considered.

MSC: 62G30; 62F03

Keywords: Information theory, ranked set sampling, test of fit.

1. Introduction

When the sampling units are difficult to measure but are reasonably simple and cheap to order according to the variable of interest, ranked set sampling (RSS) serves as an appealing alternative to the usual simple random sampling (SRS). Examples of this setup can be found in areas such as agriculture, environment and ecology. The RSS design works by ranking randomly drawn sampling units and quantifying a selected subset of them. McIntyre (1952) introduced this sampling technique while studying the yield of pasture in Australia. He suggested that a fairly accurate ordering of a set of adjacent plots by yield can be made using visual perception, although measuring the

* Department of Statistics, School of Mathematical Sciences, Ferdowsi University of Mashhad, P.O. Box 91775-1159, Mashhad, Iran. Corresponding author E-mail address: mahdizadeh.m@live.com
Received: June 2010
Accepted: December 2010

yield of each plot is expensive. As a similar situation, consider the following example mentioned by Gulati (2004). Suppose it is of interest to count the number of specific bacterial cells per unit volume in a cell suspension. A set of test tubes, containing the cell suspension, can be ordered by concentration using an optical device without actual measurement on them.

The RSS method can be elucidated as follows.

1. Draw k random samples, each of size k, from the target population.
2. Apply judgement ordering, by any cheap method, on the elements of the ith $(i = 1,\ldots,k)$ sample and identify the ith smallest unit.
3. Actually measure the k identified units in step 2.
4. Repeat steps 1-3, h times (cycles), if necessary, to obtain a ranked set sample of size $n = hk$.

The set of measured observations makes up a ranked set sample of size n denoted by $\{X_{[i]j} : i = 1,\ldots,k; j = 1,\ldots,h\}$, where $X_{[i]j}$ is the ith judgement order statistic from the jth cycle. To have better understanding of difference between the ranked set sample and simple ranked set sample of the same size, we consider the case of single cycle ($h = 1$) and perfect judgement ranking. In this case, the ranked set sample observations are also the respective order statistics. Let $X_1,\ldots,X_k$ be a simple random sample of size k from a continuous population with probability density function (PDF) $f(x)$ and cumulative distribution function (CDF) $F(x)$, and let $X_{[1]},\ldots,X_{[k]}$ denote a ranked set sample of size k obtained as described above.

In the SRS case, the k observations are independent and each of them represents a typical value from the population. Letting $X_{(1)} \leq \ldots \leq X_{(k)}$ be the order statistics associated with these SRS observations, we note that they are dependent random variables with joint PDF given by

$$g_{\mathrm{SRS}}(x_{(1)},\ldots,x_{(k)}) = k!\prod_{i=1}^{k} f(x_{(i)}).$$

In the RSS settings, additional information and structure is provided by through the judgement ranking process. The k measurements $X_{[1]},\ldots,X_{[k]}$ are also order statistics but in this case they are independent observations and each of them provides information about a different aspect of the population. The joint PDF for $X_{[1]},\ldots,X_{[k]}$ is given by

$$g_{\mathrm{RSS}}(x_{[1]},\ldots,x_{[k]}) = \prod_{i=1}^{k} f_i(x_{[i]}),$$

where $f_i(.)$ is the PDF for the ith order statistic of a simple random sample of size k from the target population. It is this extra structure provided by judgement ranking and

the independence of the resulting order statistics that enables RSS-based procedures to be more efficient than their RSS competitors with the same number of quantified units. A detailed discussion on the theory and applications of RSS can be found in the recent book by Chen et al. (2004).

Consider estimating the population mean under the aforesaid designs. Let $\bar{X}_{\mathrm{SRS}} = \sum_{i=1}^{k} X_i/k$ and $\bar{X}_{\mathrm{RSS}} = \sum_{i=1}^{k} X_{[i]}/k$ be the SRS and RSS sample mean, respectively. Hence, we have

$$\begin{aligned}\mathrm{E}(\bar{X}_{\mathrm{RSS}}) &= \frac{1}{k}\sum_{i=1}^{k}\Big\{\int_{-\infty}^{\infty} kx\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}f(x)\,dx\Big\}\\ &= \int_{-\infty}^{\infty} xf(x)\Big\{\sum_{i=1}^{k}\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}\Big\}\,dx. \qquad (1)\end{aligned}$$

Since the summation in equation (1) is just the sum over entire sample space of the probabilities for a binomial random variable with parameters $k-1$ and $F(x)$, it follows that

$$\mathrm{E}(\bar{X}_{\mathrm{RSS}}) = \int_{-\infty}^{\infty} xf(x)\,dx = \mu.$$

Letting $\mu_{[i]} = \mathrm{E}(X_{[i]})$, for $i = 1,\ldots,k$, we note that

$$\mathrm{E}(X_{[i]}-\mu)^2 = \mathrm{E}(X_{[i]}-\mu_{[i]}+\mu_{[i]}-\mu)^2 = \mathrm{E}(X_{[i]}-\mu_{[i]})^2 + (\mu_{[i]}-\mu)^2,$$

since the cross-product terms are zero. So

$$\mathrm{Var}(\bar{X}_{\mathrm{RSS}}) = \frac{1}{k^2}\Big\{\sum_{i=1}^{k}\mathrm{E}(X_{[i]}-\mu)^2 - \sum_{i=1}^{k}(\mu_{[i]}-\mu)^2\Big\}. \qquad (2)$$

Now, proceeding as we did with $\mathrm{E}(\bar{X}_{\mathrm{RSS}})$, we see that

$$\begin{aligned}\sum_{i=1}^{k}\mathrm{E}(X_{[i]}-\mu)^2 &= \sum_{i=1}^{k}\int_{-\infty}^{\infty} k(x-\mu)^2\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}f(x)\,dx\\ &= k\int_{-\infty}^{\infty}(x-\mu)^2 f(x)\Big\{\sum_{i=1}^{k}\binom{k-1}{i-1}[F(x)]^{i-1}[1-F(x)]^{k-i}\Big\}\,dx.\end{aligned}$$

Once again, using the binomial expansion, the interior sum is equal to 1 and we obtain

$$\sum_{i=1}^{k}\mathrm{E}(X_{[i]}-\mu)^2 = k\int_{-\infty}^{\infty}(x-\mu)^2 f(x)dx = k\sigma^2. \qquad (3)$$

Combining equations (2) and (3) yields

$$\mathrm{Var}(\bar{X}_{\mathrm{RSS}}) = \frac{\sigma^2}{k} - \frac{1}{k^2}\sum_{i=1}^{k}(\mu_{[i]} - \mu)^2 \leq \mathrm{Var}(\bar{X}_{\mathrm{SRS}}).$$

Al-Saleh and Al-kadiri (2000) extended the usual concept of RSS to to double ranked set sampling (DRSS) with the aim of constructing improved estimators of the population as compared with those associated with RSS and SRS. Subsequently, Al-Saleh and Al-Omari (2002) introduced multistage ranked set sampling (MSRSS), as a generalization of DRSS, and showed that estimators based on MSRSS dominate those obtained by DRSS. The MSRSS scheme can be summarized as follows.

1. Randomly identify k^{r+1} units from the population of interest, where r is the number of stages.
2. Allot the k^{r+1} units randomly into k^{r-1} sets of k^2 units each.
3. For each set in step 2, apply 1-2 of RSS procedure explained above, to get a (judgement) ranked set of size k. This step gives k^{r-1} (judgement) ranked sets, each of size k.
4. Without actual measuring of the ranked sets, apply step 3 on the k^{r-1} ranked set to gain k^{r-2} second stage (judgement) ranked sets, of size k each.
5. Repeat step 3, without any actual measurement, until an rth stage (judgement) ranked set of size k is acquired.
6. Actually measure the k identified units in step 5.
7. Repeat steps 1-6, h times, if necessary, to obtain an rth stage ranked set sample of size $n = hk$.

In analogy with the previous notation, the rth stage ranked set sample will be denoted by $\{X^{(r)}_{[i]j} : i = 1,\ldots,k; j = 1,\ldots,h\}$. Two special cases of $r = 1$ and $r = 2$ in MSRSS coincide with RSS and DRSS, respectively.

Goodness-of-fit tests are used to decide whether an observed sample can be considered as a set of independent realization from a given CDF F_0. More precisely, they are used to test the hypothesis $H_0 : F = F_0$, with F being the true CDF of the observations. For a review of goodness-of-fit tests based on SRS refer to the book by D'Agostino and Stephens (1986). Testing hypotheses on the parameters of classical distributions using ranked set samples have been developed in a large number of papers. However, this is not true in the case of test of fit, and a limited number of works are available on this topic. Stokes and Sager (1988) exploited RSS in estimating CDF. They proposed RSS analogue of Kolmogorov-Smirnov (KS) test and derived the null distribution of the test statistic.

Some distributions like normal, exponential and uniform have received much attention in the literature because of their tractable mathematical form. This is true in the

case of RSS and its variations. For example, estimation of parameters and quantiles of uniform distribution using generalized ranked-set sampling have been investigated (e.g., Adatia, 2003; Adatia and Ehsanes Saleh, 2004). In practical situations, however, the distributional form of the population is rarely known. Thus, application of these customized inferential methods is dependent on the availability of appropriate testing procedures for the assumptions of uniformity. Given a sample size, relative precision (RP) of the RSS estimator of the population mean with respect to its SRS counterpart (defined as the variance of the SRS mean divided by the variance of the RSS mean) differs according to the underlying distribution of the data, and is bounded above by $(k+1)/2$ for continuous distributions ($1 < \mathrm{RP} < (k+1)/2$) (where k is the set size with which the ranked set sample is collected), with the upper bound achieved only for the uniform distribution. We may be interested to know whether the RSS has the highest efficiency over SRS in estimating the population mean in a specific situation. This could be another reason for developing uniformity test based on RSS.

As an information-theoretic measure of uncertainty, Shannon (1948) proposed entropy of a distribution, and proved that the entropy of normal distribution exceeds that of any other distribution with a density having the same variance. Vasicek (1976) used this property to introduce a test of the composite hypothesis of normality, and impressed development of tests of fit for other distributions. Such entropy-based tests of fit are available for some other distributions. See Dudewicz and van der Meulen (1981), Gokhale (1983), Grzegorzewski and Wieczorkowski (1999), and Mudholkar and Tian (2002). In this paper, we tackle the problem of testing uniformity, with an entropy-based approach, when the researcher obtains data using RSS and MSRSS. Similar procedures for the inverse Gaussian law was suggested by Mahdizadeh and Arghami (2010).

The paper proceeds as follows. In Section 2, some basic notions from information theory are reviewed, entropy based tests of uniformity based on RSS and MSRSS are suggested, and critical values of the respective test statistics are provided for some sample sizes. Power properties of the new tests are assessed by means of simulations whose results are reported in Section 3. A summary completes the paper in Section 4.

2. The tests

Entropy of a distribution $F(x)$ with density function $f(x)$ is defined as

$$H(f) = -\int_{-\infty}^{\infty} f(x)\log f(x)\,dx. \tag{4}$$

Vasicek (1976) presented a nonparametric entropy estimator for $H(f)$ based on spacings of sample order statistics. The estimator called sample entropy is given by

$$V_{m,n}(f_X) = \frac{1}{n}\sum_{i=1}^{n} \log\big(\frac{n}{2m}(X_{(i+m)} - X_{(i-m)})\big), \tag{5}$$

where $X_{(1)},\ldots,X_{(n)}$ are the ordered values of a random sample of size n from F, $X_{(j)} = X_{(1)}$, if $j < 1$, $X_{(j)} = X_{(n)}$, if $j > n$ and the window size m is a positive integer such that $m \leq n/2$. This estimator is derived by expressing (4) in the form

$$H(f) = \int_0^1 \log\big(\frac{d}{du}F^{-1}(u)\big)\,du,$$

replacing the distribution F by the empirical distribution function, and using a difference operator instead of the differential operator.

Since entropy estimator (5) is based on spacings, one would need ordered values of the ranked set sample to estimate entropy in RSS. Imitating the SRS case, we first pool the units in all cycles and then form the estimator based on the ordered pooled sample. The MSRSS analogue of $V_{m,n}(f_X)$ turns out to be

$$V^{(r)}_{m,n}(f_X) = \frac{1}{n}\sum_{i=1}^{n} \log\big(\frac{n}{2m}(X^{(r)}_{(i+m)} - X^{(r)}_{(i-m)})\big),$$

where $X^{(r)}_{(a)}$ is the ath $(a = 1,\ldots,n)$ order statistic of the rth stage ranked set sample. From now on, the estimator (5) will be denoted by $V^{(0)}_{m,n}(f_X)$.

A simulation study was undertaken to compare the proposed estimators of entropy when the uniform U(0,1) is the underlying distribution. Table 1 displays simulated biases and root mean square errors (RMSEs) of $V^{(r)}_{m,n}$ for $r = 0,1,2$ based on 10,000 samples with $n = 10,20,30$, and $k = 10$ in MSRSS design (this setup is retained throughout the paper). It is seen that MSRSS improves entropy estimation with respect to SRS for given m and n. Besides, as the stage number increases, the absolute bias, and RMSE of the corresponding estimator diminishes.

Consider a random sample $X_1,\ldots,X_n$ from a population having a density function f with the support (0,1) and suppose it is of interest to verify $H_0 : X \sim \mathrm{U}(0,1)$ versus $H_1 : \sim H_0$. It is well-known that for an f concentrated on (0,1) we have $H(f) \leq 0$, and the maximum value of $H(f)$ is uniquely attained by the U(0,1) density (see Ash, 1965). Based on this result, Dudewicz and van der Meulen (1981) developed a test of H_0. Their test procedure is alternatively defined by the critical region

$$T_{m,n}(f_X) = \exp\big(V_{m,n}(f_X)\big) \leq T^*_{m,n,\alpha}(f_X),$$

where $T^*_{m,n,\alpha}(f_X)$ is the 100α percentile of the null distribution of $T_{m,n}(f_X)$. It can be shown, using convexity and Jensen's inequality, that $V_{m,n}(f_X) \leq 0$ for all f on (0,1).

Table 1: *Simulated biases and RMSEs of $V_{m,n}^{(r)}(f)$ $(r=0,1,2)$ for the U(0,1) distribution with $H(f)=0$.*

		SRS		RSS		DRSS	
n	m	Bias	RMSE	Bias	RMSE	Bias	RMSE
10	1	−0.5192	0.5709	−0.4007	0.4469	−0.3262	0.3692
	2	−0.4112	0.4478	−0.3085	0.3348	−0.2598	0.2778
	3	−0.4223	0.4532	−0.3272	0.3430	−0.2968	0.3067
	4	−0.4580	0.4866	−0.3715	0.3831	−0.3477	0.3541
	5	−0.5026	0.5282	−0.4256	0.4360	−0.4043	0.4101
20	1	−0.3955	0.4193	−0.3420	0.3646	−0.3088	0.3299
	2	−0.2718	0.2903	−0.2194	0.2351	−0.1894	0.2027
	3	−0.2547	0.2712	−0.2048	0.2160	−0.1826	0.1919
	4	−0.2609	0.2751	−0.2153	0.2242	−0.1987	0.2054
	5	−0.2783	0.2908	−0.2349	0.2420	−0.2212	0.2262
	6	−0.2972	0.3080	−0.2592	0.2650	−0.2478	0.2518
	7	−0.3230	0.3336	−0.2859	0.2908	−0.2755	0.2787
	8	−0.3468	0.3567	−0.3141	0.3184	−0.3041	0.3068
	9	−0.3772	0.3871	−0.3425	0.3468	−0.3344	0.3370
	10	−0.4041	0.4133	−0.3708	0.3747	−0.3637	0.3661
30	1	−0.3539	0.3697	−0.3210	0.3360	−0.2978	0.3118
	2	−0.2247	0.2373	−0.1917	0.2024	−0.1698	0.1795
	3	−0.1980	0.2089	−0.1642	0.1725	−0.1464	0.1538
	4	−0.1954	0.2049	−0.1639	0.1708	−0.1484	0.1542
	5	−0.2016	0.2101	−0.1719	0.1776	−0.1605	0.1651
	6	−0.2136	0.2211	−0.1850	0.1899	−0.1749	0.1788
	7	−0.2273	0.2342	−0.2000	0.2041	−0.1922	0.1954
	8	−0.2441	0.2509	−0.2179	0.2214	−0.2104	0.2131
	9	−0.2596	0.2655	−0.2354	0.2385	−0.2286	0.2308
	10	−0.2769	0.2826	−0.2543	0.2572	−0.2482	0.2501
	11	−0.2948	0.3003	−0.2736	0.2762	−0.2681	0.2698
	12	−0.3138	0.3191	−0.2921	0.2946	−0.2880	0.2897
	13	−0.3329	0.3381	−0.3117	0.3142	−0.3070	0.3086
	14	−0.3508	0.3559	−0.3323	0.3347	−0.3272	0.3287
	15	−0.3702	0.3753	−0.3520	0.3544	−0.3473	0.3487

Thus, we used the exponential of the original test statistic in the above for mathematical nicety.

In order to obtain the percentiles of the null distribution, $T_{m,n}(f_X)$ was calculated using the estimators $V_{m,n}^{(r)}(f_X)$ for $r=0,1,2$ based on 10,000 samples of size n generated from the U(0,1) distribution. The values were then used to determine $T_{m,n,0.1}^{*}(f_X)$ in different designs and for different sample sizes. Table 2 displays 0.1 critical points for the test statistics.

Table 2: *0.1 critical points for the test statistics under SRS, RSS and DRSS designs.*

n	m	SRS	RSS	DRSS	n	m	SRS	RSS	DRSS
10	1	0.4329	0.5186	0.5730	30	1	0.6089	0.6374	0.6557
	2	0.5213	0.6197	0.6765		2	0.7215	0.7575	0.7801
	3	0.5267	0.6272	0.6725		3	0.7508	0.7894	0.8129
	4	0.5119	0.6084	0.6458		4	0.7569	0.7982	0.8143
	5	0.4881	0.5769	0.6091		5	0.7553	0.7940	0.8094
20	1	0.5576	0.6003	0.6325		6	0.7491	0.7852	0.7980
	2	0.6642	0.7185	0.7518		7	0.7387	0.7748	0.7892
	3	0.6871	0.7432	0.7706		8	0.7276	0.7631	0.7758
	4	0.6865	0.7425	0.7667		9	0.7153	0.7506	0.7624
	5	0.6783	0.7317	0.7532		10	0.7039	0.7380	0.7485
	6	0.6645	0.7178	0.7365		11	0.6914	0.7235	0.7346
	7	0.6490	0.7005	0.7173		12	0.6767	0.7098	0.7211
	8	0.6324	0.6811	0.6980		13	0.6640	0.6955	0.7071
	9	0.6141	0.6613	0.6768		14	0.6501	0.6816	0.6912
	10	0.5968	0.6416	0.6574		15	0.6379	0.6671	0.6758

The test statistics use the entropy estimators and there is no criteria to select the optimal window size associated with a given sample size in order to calculate these estimators. As a guide mentioned by some authors, the window size producing the largest critical value for a given n is apt to yield the highest power. In this sense, the optimal window size, denoted by m^*, at the significance level 0.1 for sample sizes 10, 20 and 30 are approximately 3, 3 and 4, respectively. Figure 1 shows a comparison of

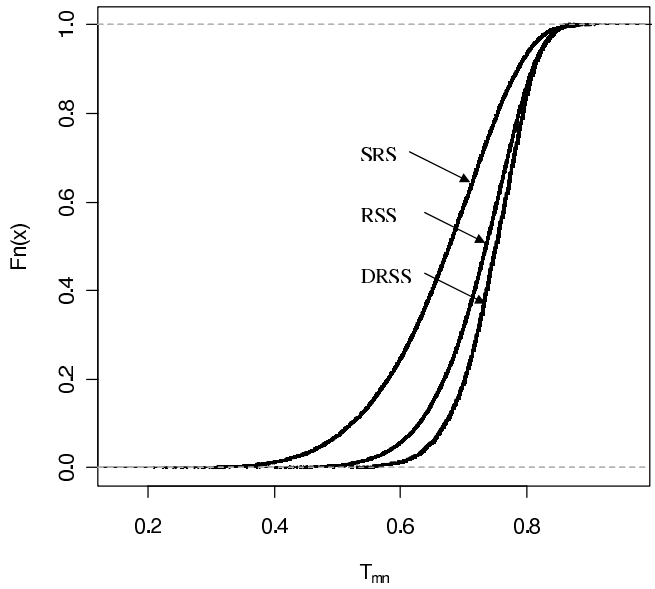


Figure 1: *This figures compares the CDF of null distribution of $T_{3,10}$ under SRS, RSS and DRSS designs.*

CDF of the test statistics in differen designs. It is observed that the null distribution of $T_{3,10}$ under SRS (RSS) is stochastically smaller than that under RSS (DRSS) (a similar trend is observed for sample sizes $n = 20, 30$). Thus, we expect the entropy test based on RSS (DRSS) to be more powerful than that based on SRS (RSS).

3. Simulation results

A Monte Carlo simulation experiment is carried out to compare power of the entropy tests. We considered three classes of alternatives presented by Stephens (1974) which have been used by many authors. These alternatives specified by their distribution functions are

$$A(k): F(z) = 1 - (1-z)^k \quad 0 \le z \le 1 \quad (k = 1.5, 1.75, 2),$$

$$B(k): F(z) = \begin{cases} 2^{k-1} z^k & 0 \le z \le 0.5 \\ 1 - 2^{k-1}(1-z)^k & 0.5 \le z \le 1 \end{cases} \quad (k = 1.5, 1.75),$$

and

$$C(k): F(z) = \begin{cases} 0.5 - 2^{k-1}(0.5-z)^k & 0 \le z \le 0.5 \\ 0.5 + 2^{k-1}(z-0.5)^k & 0.5 \le z \le 1 \end{cases} \quad (k = 2, 2.5).$$

As compared with uniform, the first and second family give points closer to 0 and 0.5, respectively. And the third family gives points clustered at 0 and 1. We also considered Beta(2,2) as a symmetric distribution.

Under each design, 10,000 samples of sizes $n = 10, 20, 30$ were generated from each alternative distribution and the power of the tests were estimated by proportion of the samples falling into the corresponding critical region. Tables 3–6 exhibit the estimated power of the tests.

The results manifest that given a sample size, the entropy tests based on RSS and DRSS are more powerful than that based on SRS irrespective of the alternative distribution. Moreover, improved tests are obtained by increasing the sampling effort. That is DRSS has the best performance among three considered designs as is the case of entropy estimation. This could be traced to the fact that the test statistic in each design is constructed based on the corresponding entropy estimator. It is notable that RSS and DRSS do not have much to offer when power of SRS design is less than 0.1. We observe that for $n = 10$, the value $m = 4$ is best (in the sense that it yields the highest power) for the tests under most alternatives except C (for which $m = 1$ is best). For $n = 20$, best m for alternatives A, B and C are respectively 7, 10 and 2, while for $n = 30$ these are 10, 15 and 3. Given a sample size, best m is different according to the alternative

Table 3: *Power comparison for the entropy tests of size 0.1 against alternatives A(1.5) and A(1.75).*

		A(1.5)			A(1.75)		
n	*m*	SRS	RSS	DRSS	SRS	RSS	DRSS
10	1	0.1745	0.1879	0.1924	0.2431	0.2716	0.2887
	2	0.2182	0.2668	0.3360	0.3147	0.4198	0.5609
	3	0.2325	0.3306	0.4635	0.3451	0.5276	0.7142
	4	0.2397	0.3814	0.5285	0.3570	0.5766	0.7628
	5	0.2436	0.3794	0.5017	0.3503	0.5728	0.7396
20	1	0.2298	0.2367	0.2620	0.3474	0.3843	0.4356
	2	0.3052	0.3530	0.4098	0.4786	0.6022	0.7178
	3	0.3292	0.4351	0.5174	0.5342	0.7264	0.8413
	4	0.3704	0.5064	0.6030	0.5760	0.7956	0.9032
	5	0.3728	0.5301	0.6386	0.5817	0.8207	0.9186
	6	0.3846	0.5693	0.6962	0.5932	0.8494	0.9451
	7	0.3817	0.5867	0.7092	0.5870	0.8575	0.9482
	8	0.3754	0.5801	0.7126	0.5821	0.8490	0.9436
	9	0.3720	0.5718	0.6996	0.5713	0.8358	0.9322
	10	0.3681	0.5536	0.6812	0.5608	0.8114	0.9156
30	1	0.2737	0.2871	0.2890	0.4554	0.4962	0.5216
	2	0.3795	0.4287	0.4882	0.6026	0.7538	0.8430
	3	0.4260	0.5468	0.6324	0.6748	0.8710	0.9406
	4	0.4556	0.6195	0.6958	0.7106	0.9175	0.9748
	5	0.4821	0.6536	0.7476	0.7382	0.9387	0.9824
	6	0.4926	0.6783	0.7662	0.7512	0.9435	0.9870
	7	0.5016	0.6985	0.8210	0.7533	0.9516	0.9936
	8	0.5137	0.7245	0.8344	0.7642	0.9578	0.9945
	9	0.5068	0.7352	0.8490	0.7618	0.9622	0.9934
	10	0.5184	0.7510	0.8538	0.7723	0.9651	0.9927
	11	0.5170	0.7486	0.8612	0.7674	0.9601	0.9954
	12	0.4996	0.7442	0.8569	0.7505	0.9570	0.9932
	13	0.4954	0.7355	0.8556	0.7410	0.9513	0.9942
	14	0.4825	0.7190	0.8230	0.7295	0.9372	0.9900
	15	0.4768	0.6925	0.7942	0.7153	0.9241	0.9786

Table 4: *Power comparison for the entropy tests of size 0.1 against alternatives A(2) and B(1.5).*

		A(2)			B(1.5)		
n	m	SRS	RSS	DRSS	SRS	RSS	DRSS
10	1	0.3181	0.3742	0.4107	0.1948	0.2245	0.2310
	2	0.4254	0.5969	0.7821	0.2716	0.3520	0.4633
	3	0.4635	0.7208	0.8913	0.3188	0.4795	0.6490
	4	0.4674	0.7648	0.9134	0.3425	0.5734	0.7572
	5	0.4612	0.7430	0.8882	0.3609	0.6026	0.7698
20	1	0.4983	0.5712	0.6344	0.2417	0.2672	0.2856
	2	0.6541	0.8348	0.9318	0.3458	0.4236	0.4978
	3	0.7103	0.9221	0.9830	0.3973	0.5324	0.6206
	4	0.7472	0.9532	0.9926	0.4564	0.6213	0.7230
	5	0.7624	0.9608	0.9952	0.4900	0.6792	0.7844
	6	0.7718	0.9680	0.9964	0.5140	0.7328	0.8465
	7	0.7697	0.9735	0.9972	0.5406	0.7760	0.8751
	8	0.7562	0.9658	0.9948	0.5510	0.8035	0.8924
	9	0.7445	0.9567	0.9931	0.5636	0.8142	0.9172
	10	0.7320	0.9453	0.9876	0.5727	0.8210	0.9204
30	1	0.6324	0.7255	0.7740	0.2911	0.3128	0.3397
	2	0.8009	0.9472	0.9872	0.4085	0.4939	0.5516
	3	0.8613	0.9861	0.9986	0.4821	0.6174	0.7050
	4	0.8870	0.9934	0.9998	0.5349	0.6960	0.7812
	5	0.9010	0.9963	1.0000	0.5719	0.7486	0.8305
	6	0.9084	0.9968	1.0000	0.6034	0.7764	0.8570
	7	0.9142	0.9980	0.9998	0.6170	0.8123	0.8996
	8	0.9175	0.9985	1.0000	0.6452	0.8375	0.9230
	9	0.9151	0.9984	0.9998	0.6636	0.8681	0.9408
	10	0.9182	0.9992	1.0000	0.6901	0.8894	0.9562
	11	0.9135	0.9981	1.0000	0.7004	0.9045	0.9636
	12	0.9064	0.9977	1.0000	0.7088	0.9173	0.9748
	13	0.8998	0.9964	1.0000	0.7190	0.9257	0.9782
	14	0.8890	0.9932	1.0000	0.7201	0.9212	0.9718
	15	0.8756	0.9925	0.9996	0.7236	0.9220	0.9706

Table 5: *Power comparison for the entropy tests of size 0.1 against alternatives B(1.75) and C(2).*

		B(1.75)			C(2)		
n	*m*	SRS	RSS	DRSS	SRS	RSS	DRSS
10	1	0.2696	0.3250	0.3514	0.2082	0.2431	0.2487
	2	0.3875	0.5339	0.6872	0.1430	0.1829	0.2151
	3	0.4612	0.6872	0.8495	0.0647	0.0483	0.0465
	4	0.5026	0.7831	0.9155	0.0475	0.0069	0.0009
	5	0.5219	0.8106	0.9287	0.0296	0.0025	0.0004
20	1	0.3815	0.4294	0.4830	0.4006	0.4483	0.5054
	2	0.5372	0.6652	0.7646	0.4324	0.5364	0.6298
	3	0.6054	0.7884	0.8820	0.3662	0.4892	0.5904
	4	0.6687	0.8607	0.9381	0.2751	0.3866	0.4613
	5	0.7018	0.8939	0.9562	0.1416	0.1989	0.2258
	6	0.7351	0.9210	0.9742	0.0626	0.0548	0.0476
	7	0.7600	0.9469	0.9834	0.0372	0.0062	0.0030
	8	0.7684	0.9573	0.9858	0.0261	0.0014	0.0002
	9	0.7842	0.9618	0.9902	0.0208	0.0007	0.0000
	10	0.7890	0.9624	0.9944	0.0149	0.0004	0.0000
30	1	0.4742	0.5268	0.5600	0.5574	0.6159	0.6627
	2	0.6537	0.7695	0.8597	0.6590	0.7601	0.8653
	3	0.7351	0.8835	0.9443	0.6512	0.7963	0.8977
	4	0.7867	0.9287	0.9725	0.6032	0.7654	0.8693
	5	0.8136	0.9508	0.9873	0.5244	0.6888	0.7830
	6	0.8362	0.9621	0.9897	0.4213	0.5626	0.6690
	7	0.8514	0.9723	0.9906	0.2789	0.4130	0.5117
	8	0.8725	0.9782	0.9950	0.1507	0.2194	0.2453
	9	0.8799	0.9847	0.9962	0.0652	0.0672	0.0705
	10	0.8980	0.9906	0.9967	0.0338	0.0097	0.0067
	11	0.9061	0.9912	0.9990	0.0241	0.0008	0.0000
	12	0.9078	0.9926	0.9993	0.0176	0.0003	0.0003
	13	0.9134	0.9953	1.0000	0.0142	0.0004	0.0000
	14	0.9142	0.9938	0.9993	0.0097	0.0001	0.0000
	15	0.9187	0.9947	0.9997	0.0075	0.0000	0.0000

Table 6: *Power comparison for the entropy tests of size 0.1 against alternatives C(2.5) and B(2,2).*

		C(2.5)			B(2,2)		
n	m	SRS	RSS	DRSS	SRS	RSS	DRSS
10	1	0.3168	0.3782	0.4105	0.2630	0.3174	0.3496
	2	0.2124	0.2951	0.3884	0.3767	0.5336	0.6941
	3	0.0820	0.0672	0.0713	0.4382	0.6843	0.8560
	4	0.0506	0.0066	0.0005	0.4821	0.7805	0.9158
	5	0.0292	0.0018	0.0001	0.5064	0.8028	0.9275
20	1	0.6375	0.7308	0.8150	0.3713	0.4160	0.4682
	2	0.6932	0.8286	0.9361	0.5364	0.6647	0.7724
	3	0.6184	0.7948	0.9178	0.6171	0.7995	0.8843
	4	0.4951	0.6870	0.8216	0.6796	0.8713	0.9420
	5	0.2755	0.4264	0.5384	0.7204	0.9035	0.9600
	6	0.0978	0.1192	0.1374	0.7446	0.9351	0.9784
	7	0.0480	0.0071	0.0030	0.7657	0.9483	0.9860
	8	0.0326	0.0010	0.0000	0.7780	0.9609	0.9876
	9	0.0254	0.0004	0.0000	0.7832	0.9626	0.9915
	10	0.0169	0.0001	0.0000	0.7894	0.9610	0.9928
30	1	0.8298	0.8974	0.9457	0.4624	0.5109	0.5486
	2	0.9129	0.9740	0.9963	0.6508	0.7693	0.8460
	3	0.9064	0.9781	0.9990	0.7440	0.8832	0.9476
	4	0.8806	0.9738	0.9987	0.7974	0.9346	0.9704
	5	0.8246	0.9464	0.9943	0.8294	0.9540	0.9835
	6	0.7350	0.8897	0.9718	0.8545	0.9654	0.9884
	7	0.5681	0.7825	0.9196	0.8679	0.9766	0.9942
	8	0.3442	0.5314	0.6851	0.8834	0.9825	0.9960
	9	0.1338	0.1970	0.2430	0.8920	0.9879	0.9982
	10	0.0523	0.0218	0.0210	0.9056	0.9915	0.9994
	11	0.0327	0.0012	0.0000	0.9142	0.9923	0.9986
	12	0.0242	0.0003	0.0000	0.9165	0.9942	0.9991
	13	0.0170	0.0001	0.0000	0.9178	0.9948	0.9994
	14	0.0122	0.0001	0.0000	0.9160	0.9936	0.9992
	15	0.0097	0.0000	0.0000	0.9181	0.9943	0.9994

distribution. As a remedy, we may use data histogram to determine best window size for implementing the tests. Table 7 compares the power of RSS entropy based test for uniformity, when m is best, with that of the KS test whose results are given in italic. It is seen that entropy test shows remarkable dominance over the KS test against alternatives B and B(2,2), whereas the KS test is better for alternatives A and C.

Table 7: *Power comparison for the entropy test and KS test of size 0.1 against several alternative distributions under RSS.*

	Distribution							
n	A(1.5)	A(1.75)	A(2)	B(1.5)	B(1.75)	C(2)	C(2.5)	B(2,2)
10	0.381	0.577	0.765	0.603	0.811	0.243	0.378	0.803
	0.629	*0.875*	*0.971*	*0.176*	*0.290*	*0.583*	*0.798*	*0.235*
20	0.587	0.858	0.974	0.821	0.962	0.536	0.829	0.961
	0.884	*0.993*	*1.000*	*0.327*	*0.566*	*0.845*	*0.975*	*0.482*
30	0.751	0.965	0.999	0.922	0.995	0.796	0.978	0.994
	0.970	*1.000*	*1.000*	*0.463*	*0.768*	*0.950*	*0.997*	*0.691*

Table 8: *0.1 critical points of the test statistics under MSRSS designs.*

	Stage Number		
$n(m^*)$	$r=2$	$r=3$	$r=4$
10(3)	0.6725	0.6910	0.7048
20(3)	0.7706	0.7892	0.7956
30(4)	0.8143	0.8236	0.8281

Table 9: *Power comparison for the entropy tests of size 0.1 against several alternative distributions under MSRSS designs.*

		Distribution							
$n(m^*)$	r	A(1.5)	A(1.75)	A(2)	B(1.5)	B(1.75)	C(2)	C(2.5)	B(2,2)
10(3)	2	0.4635	0.7142	0.8913	0.6490	0.8495	0.2487	0.4105	0.8560
	3	0.5371	0.7925	0.9467	0.7459	0.9011	0.2660	0.4419	0.9078
	4	0.5940	0.8593	0.9702	0.7762	0.9304	0.3171	0.5295	0.9517
20(3)	2	0.5174	0.8413	0.9830	0.6206	0.8820	0.5904	0.9178	0.8843
	3	0.5866	0.8945	0.9956	0.6874	0.9268	0.6780	0.9732	0.9282
	4	0.6218	0.9387	1.0000	0.7033	0.9409	0.7161	0.9846	0.9613
30(4)	2	0.6958	0.9748	0.9998	0.7812	0.9725	0.8693	0.9987	0.9704
	3	0.7340	0.9896	1.0000	0.8126	0.9893	0.9221	1.0000	0.9855
	4	0.7535	1.0000	1.0000	0.8290	0.9984	0.9407	1.0000	1.0000

Tables 2 and 3–6 were formed under MSRSS with $r = 3, 4$ to see whether further increase in power is achieved by increasing the stage number. Tables 8 and 9 contain 0.1 critical points and power of the tests, respectively. For a given n, the results are provided only for the optimal m, except for C family and $n = 10$ where $m = 1$ is applied. Also, results of DRSS design were included to ease comparison. From Table 9, we can see that as r increases, some improvement in power happens. The differences in results for $r = 2$ and $r = 3, 4$ are less pronounced in large sample size, and thus we may restrict ourselves to DRSS in practice.

4. Conclusion

This article was directed at the problem of developing tests of uniformity under RSS and MSRSS designs. In line with the available entropy based test of fit in SRS, our tests use sample entropy based on the pre-mentioned designs. Simulation studies accompany the presentation to explore power behaviour of the proposed tests in finite sample sizes. The results disclose that RSS and its variations outperform SRS in constructing powerful entropy based test of uniformity. The authors have developed similar tests for other distributions (e.g. uniform, beta, exponential, gamma, log-normal, Pareto, Rayleigh, Weibull, normal, Laplace, etc.) using improved entropy estimators (e.g., see Ebrahimi et al. (1994) and Novi Inverardi (2003)). The results will be reported in separate works.

Acknowledgements

The authors are grateful to the referees for their helpful comments that clearly improved this article. Partial support from "Ordered and Spatial Data Center of Excellence of Ferdowsi University of Mashhad" is acknowledged.

References

Adatia, A. (2003). Estimation of parameters of uniform distribution using generalized ranked set sampling. *Journal of Statistical Research*, 37, 193–202.

Adatia, A. and Ehsanes Saleh, A. K. Md. (2004). Estimation of quantiles of uniform distribution using generalized ranked-set sampling. *Pakistan Journal of Statistics*, 20, 355–368.

Al-Saleh, M. F. and Al-Kadiri, M. (2000). Double ranked set sampling. *Statistics & Probability Letters*, 48, 205–212.

Al-Saleh, M. F. and Al-Omari, A. I. (2002). Multistage ranked set sampling. *Journal of Statistical Planning and Inference*, 102, 273–286.

Ash, R. B. (1965). *Information Theory*. John Wiley & Sons, New York.

Chen, Z., Bai, Z. and Sinha, B. K. (2004). *Ranked set sampling: Theory and Applications*. Springer, New York.

D'Agostino, R. B. and Stephens, M. A. (1986). *Goodness-of-Fit Techniques*. Marcel Dekker, New York.

Dudewicz, E. J. and van der Meulen, E. C. (1981). Entropy-based tests of uniformity. *Journal of the American Statistical Association*, 76, 967–974.

Ebrahimi, N., Pflughoeft, K. and Soofi, E. S. (1994). Two measures of sample entropy. *Statistics & Probability Letters*, 20, 225–234.

Gokhale, D. V. (1983). On the entropy-based goodness-of-fit tests. *Computational Statistics and Data Analysis*, 1, 157–165.

Grzegorzewski, P. and Wieczorkowski, R. (1999). Entropy based goodness-of-fit test for exponentiality. *Communications in Statistics–Theory and Methods*, 28, 1183–1202.

Gulati, S. (2004). Smooth non-parametric estimation of the distribution function from balanced ranked set samples. *Environmetrics*, 15, 529–539.

Mahdizadeh, M. and Arghami, N. R. (2010). Efficiency of ranked set sampling in entropy estimation and goodness-of-fit testing for the inverse Gaussian law. *Journal of Statistical Computation and Simulation*, 80, 761–774.

McIntyre, G. A. (1952). A method of unbiased selective sampling using ranked sets. *Australian Journal of Agricultural Research*, 3, 385–390.

Mudholkar, G. S. and Tian, L. (2002). An entropy characterization of the inverse Gaussian distribution and related goodness-of-fit test. *Journal of Statistical Planning and Inference*, 102, 211–221.

Novi Inverardi, P. L. (2003). MSE comparison of some different estimators of entropy. *Communications in Statistics–Simulation and Computation*, 32, 17–30.

Stokes, S. L. and Sager, T.W. (1988). Characterization of a ranked-set sample with application to estimating distribution function. *Journal of the American Statistical Association*, 83, 374–381.

Shannon, C. E. (1948). A mathematical theory of communications. *Bell System Technical Journal*, 27, 379–423, 623–656.

Stephens, M. A. (1974). EDF statistics for goodness of fit and some comparisons. *Journal of the American Statistical Association*, 69, 730–737.

Vasicek, O. (1976). A test of normality based on sample entropy. *Journal of the Royal Statistical Society: Series B*, 38, 54–59.

SORT 37 (1) January-June 2013, 19-28

A note on the Fisher information matrix for the skew-generalized-normal model

Reinaldo B. Arellano-Valle[1], Héctor W. Gómez[2] and Hugo S. Salinas[3]

Abstract

In this paper, the exact form of the Fisher information matrix for the skew-generalized normal (SGN) distribution is determined. The existence of singularity problems of this matrix for the skew-normal and normal particular cases is investigated. Special attention is given to the asymptotic properties of the MLEs under the skew-normality hypothesis.

MSC: 62E20

Keywords: Asymptotic distribution, Kurtosis, maximum likelihood estimation, singular information matrix, Skewness.

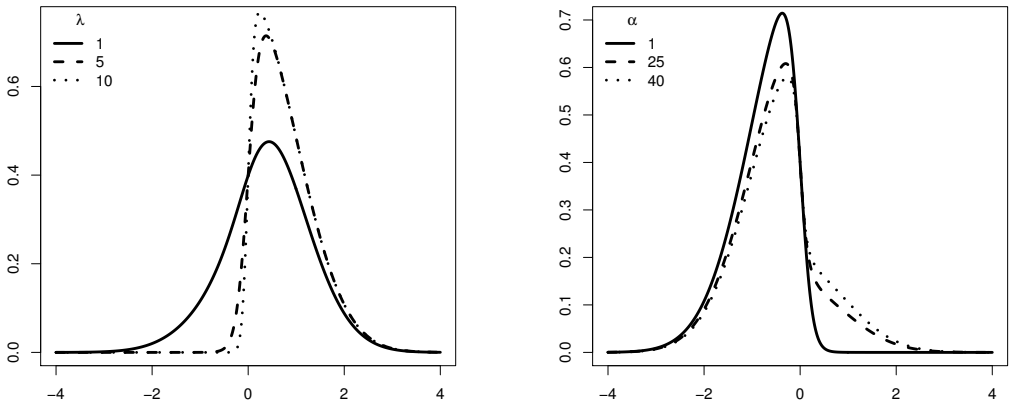
1. Introduction

Arellano-Valle, Gómez and Quintana (2004) introduced the skew-generalized-normal (SGN) distribution with density

$$f(z;\lambda,\alpha) = 2\phi(z)\Phi\left(\frac{\lambda z}{\sqrt{1+\alpha z^2}}\right),\ z\in\mathbb{R},\ \lambda\in\mathbb{R},\ \alpha\geq 0, \tag{1}$$

and denoted by $\mathrm{SGN}(\lambda,\alpha)$, where $\phi(\cdot)$ and $\Phi(\cdot)$ are the density function and cumulative distribution function of the standardized normal distribution, respectively. The skewness of the SGN distribution (1) is regulated by the parameters λ and α, so that it reduces

[1] Departamento de Estadística, Facultad de Matemáticas, Pontificia Universidad Católica de Chile, Chile.
E-mail: reivalle@mat.puc.cl
[2] Departamento de Matemáticas, Facultad de Ciencias Básicas, Universidad de Antofagasta, Antofagasta, Chile.
E-mail: hgomez@uantof.cl.
[3] Departamento de Matemática, Facultad de Ingeniería, Universidad de Atacama, Chile.
E-mail: hsalinas@mat.uda.cl.
Received: March 2011
Accepted: November 2011



(a) SGN$(\lambda, \alpha = 1)$ (b) SGN$(\lambda = -5, \alpha)$

Figure 1: *Examples of the skew-generalized normal density.*

to the skew-normal (SN) distribution when $\alpha = 0$ and to the normal (N) one when $\lambda = 0$. Note, however, that the value of α is irrelevant when $\lambda = 0$. The same occurs with λ for the limiting case when $\alpha \to \infty$. In both of these situations, the normality is attained from the SGN model, producing there a local identifiability problem. Further special models can also be obtained by reparametrizing α in terms of λ or viceversa. For example, by making $\alpha = \lambda^2$ we obtain in (1) the so-called skew-curved normal (SCN) distribution in Arellano-Valle *et al.* (2004). This flexibility of the SGN distribution allows to incorporate a wide range of models in a neighbourhood of the normal distribution. Figure 1 shows the behaviour of the SGN density for different values of the parameters λ and α. Only positive values of λ are considered in the plots of Figure 1(a); when the sign of λ is reversed, the density is reflected about the origin, as in Figure 1(b).

Further properties of the SGN model are investigated by Arellano-Valle *et al.* (2004). In particular, they gave formulas for the moments of a SGN random variable, $Z \sim$ SGN(λ, α). They showed that the even moments of Z are equal to the corresponding even moments of a standardized normal random variable. For the odd moments of Z, they obtained expressions involving an implicit formula,

$$E(Z^{2k+1}) = 2c_k - 2^k \Gamma(1+k)(2/\pi)^{1/2}, \quad k = 0, 1, 2, \ldots,$$

where $c_k := c_k(\lambda, \alpha) = \int_0^\infty u^k \, \phi(\sqrt{u}) \, \Phi\left(\frac{\lambda\sqrt{u}}{\sqrt{1+\alpha u}}\right) du$. The location-scale extension of the SGN distribution (1) was also obtained by Arellano-Valle *et al.* (2004) by letting $X = \mu + \sigma Z$, where $Z \sim$ SGN(λ, α), and where $\mu \in \mathbb{R}$ and $\sigma > 0$ are the location and scale parameters, respectively. In this case, the notation $X \sim$ SGN$(\mu, \sigma, \lambda, \alpha)$ will be used. Finally, for the mean and variance of $X \sim$ SGN$(\mu, \sigma, \lambda, \alpha)$ we can note that $E(X) = \mu + \sigma\mu_z$ and $Var(X) = (1 - \mu_z^2)\sigma^2$, where $\mu_z = 2c_0 - (2/\pi)^{1/2}$ is the mean of Z.

The Fisher information matrix has an important role in statistical analysis (classical and Bayesian) as well as in information theory. In the location-scale skew-normal distribution, however, the Fisher information matrix is singular (Azzalini, 1985) when the skewness/shape parameter is zero, i.e., under the normality hypothesis. This fact violates the standard regularity conditions leading to the asymptotic normal distribution of the MLEs. A situation of this type falls under nonstandard asymptotic theory studied by Rotnitzky et al. (2000), who showed that in these circumstances the rate of convergence estimate is slower than the usual one. Motivated by this fact, we consider it important to obtain and analyse the behaviour of the Fisher information matrix in an generalization of the skew-normal distribution.

In this note, we determine the exact form of the Fisher information matrix for the skew-generalized-normal (SGN) distribution. Next, we examine the existence of singularity problems of this matrix for the skew-normal and normal special cases, giving a special attention to the asymptotic properties of the MLEs under the skew-normality hypothesis ($\lambda = 0$).

This paper is organized as follows. The elements of the expected information matrix for the full location-scale SGN model are derived in Section 2. Solutions for the singularity problems in the full information matrix for the normal particular cases are also discussed there. The technical details are given in an Appendix.

2. Maximum likelihood estimation

This section is related to the asymptotic properties of the MLEs of the location-scale SGN model. Specifically, the ingredients to compute the expected information matrix for the full location-scale SGN model are given. Hence, the study is focused on the asymptotic behaviour of the MLEs for the particular skew-normal and normal models.

2.1. Likelihood score functions

Let $X_1,\ldots,X_n$ be a random sample drawn from the SGN$(\mu,\sigma,\lambda,\alpha)$ distribution. The log-likelihood function for $\theta = (\mu,\sigma,\lambda,\alpha)^\top$ is $\sum_{i=1}^n l(\theta,X_i)$, where $l(\theta,X)$ is the log-likelihood for θ based on a single observation X, that is,

$$l(\theta,X) := \log f(X;\theta) = \frac{1}{2}\log\left(\frac{2}{\pi}\right) - \log(\sigma) - \frac{Z^2}{2} + \log\Phi(W), \tag{2}$$

where $Z = (X-\mu)/\sigma$ and $W = W(Z) = \lambda Z/(1+\alpha Z^2)^{1/2}$. The score function is $\sum_{i=1}^n S_\theta(\theta,X_i)$, where $S_\theta(\theta,X) = \partial l(\theta,X)/\partial\theta$ is the vector $(S_\mu,S_\sigma,S_\lambda,S_\alpha)^\top$ with elements

$$S_\mu = \frac{Z}{\sigma} - \frac{1}{\sigma}\frac{\phi(W)}{\Phi(W)}\frac{\partial W}{\partial Z}, \quad S_\sigma = -\frac{1}{\sigma} + \frac{Z^2}{\sigma} - \frac{1}{\sigma}\frac{\phi(W)}{\Phi(W)}\frac{\partial W}{\partial Z}Z,$$

$$S_\lambda = \frac{\phi(W)}{\Phi(W)}\frac{\partial W}{\partial \lambda} \quad \text{and} \quad S_\alpha = \frac{\phi(W)}{\Phi(W)}\frac{\partial W}{\partial \alpha},$$

where $\partial W/\partial Z = \lambda/(1+\alpha Z^2)^{3/2}$, $\partial W/\partial \lambda = Z/(1+\alpha Z^2)^{1/2}$ and $\partial W/\partial \alpha = -\lambda Z^3/2(1+\alpha Z^2)^{3/2}$.

2.2. Fisher information matrix

By definition, the SGN-expected information matrix for θ can be computed as $\mathbf{I}_\theta = E[\mathbf{S}_\theta \mathbf{S}_\theta^\top]$, where $\mathbf{S}_\theta$ is the SGN-score vector above. Thus, the elements $I_{\theta_i\theta_j} = E[\mathbf{S}_{\theta_i}\mathbf{S}_{\theta_j}^\top]$ of this matrix are shown in the Appendix to be

$$I_{\mu\mu} = \frac{1}{\sigma^2} + \frac{\lambda^2}{\sigma^2}\eta_{03}, \quad I_{\mu\sigma} = \frac{2}{\sigma^2}(c_1 - c_0) - \frac{(2/\pi)^{1/2}}{\sigma^2} - \frac{2\lambda}{\sigma^2}\rho_{23} + \frac{\lambda}{\sigma^2}\rho_{03} + \frac{\lambda^2}{\sigma^2}\eta_{13},$$

$$I_{\mu\lambda} = \frac{1}{\sigma}\rho_{21} - \frac{\lambda}{\sigma}\eta_{12}, \quad I_{\mu\alpha} = -\frac{\lambda}{2\sigma}\rho_{43} + \frac{\lambda^2}{2\sigma}\eta_{33}, \quad I_{\sigma\sigma} = \frac{2}{\sigma^2} + \frac{\lambda^2}{\sigma^2}\eta_{23},$$

$$I_{\sigma\lambda} = -\frac{\lambda}{\sigma}\eta_{22}, \quad I_{\sigma\alpha} = \frac{\lambda^2}{2\sigma}\eta_{43}, \quad I_{\lambda\lambda} = \eta_{21}, \quad I_{\lambda\alpha} = -\frac{\lambda}{2}\eta_{42} \text{ and } I_{\alpha\alpha} = \frac{\lambda^2}{4}\eta_{63},$$

where the coefficients ρ_{nm} and η_{nm} are defined in Proposition 1 given in the Appendix. These coefficients must be computed numerically.

For the nonnormal cases with $\lambda \neq 0$ and $0 \leq \alpha < \infty$, the above information matrix is always nonsingular, so that the usual $\sqrt{n}$-asymptotic behaviour holds for the MLEs. In particular, estimation of the standard errors of the parameter estimates can be taken from the diagonal elements of the inverse Fisher information matrix. Moreover, the submatrix of the full information matrix corresponding to the vector of parameters $(\mu,\sigma,\lambda)^\top$ coincides with the SN-information matrix obtained by Azzalini (1985). In addition, for the skew-normal special case with $\alpha = 0$, the full associated information matrix is also nonsingular. See Section 2.3 below.

For the normal case that follows when $\lambda = 0$, the information matrix of $\theta = (\mu,\sigma,\lambda,\alpha)^\top$ is

$$\begin{pmatrix} \frac{1}{\sigma^2} & 0 & \frac{2(2/\pi)^{1/2}}{\sigma}d_1(\alpha) & 0 \\ & \frac{2}{\sigma^2} & 0 & 0 \\ & & \frac{2}{\pi}d_2(\alpha) & 0 \\ & & & 0 \end{pmatrix},$$

where $d_1(\alpha) = \int_0^\infty \frac{z^2\phi(z)}{(1+\alpha z^2)^{1/2}}dz$ and $d_2(\alpha) = \frac{1}{\alpha}\left(1 - (2\pi/\alpha)^{1/2}e^{\frac{1}{2\alpha}}\Phi(-\alpha^{-1/2})\right)$ (see Corollary 1 in the Appendix). Although the first three columns of this matrix are linearly

independent, it leads to a singular information matrix because of a final column (corresponding to the parameter α) of 0s. This fact is obvious from (1), since α is non-identifiable when $\lambda = 0$. Properties of the MLEs when the SGN model reduces to the normal case are considered in Section 2.4.

2.3. Properties of the MLEs in the skew-normal case

Suppose that the parameter vector is $\theta^* = (\mu^*, \sigma^*, \lambda^*, 0)^\top$, that is, the data are drawn from the $\mathrm{SN}(\mu^*, \sigma^*, \lambda^*)$ distribution. At $\theta = \theta^*$, the components of the score vector S_θ are

$$S^*_\mu = \frac{1}{\sigma^*}\left[Z^* - \lambda^* \frac{\phi(\lambda^* Z^*)}{\Phi(\lambda^* Z^*)}\right], S^*_\sigma = \frac{1}{\sigma^*}\left[Z^{*2} - 1 - \lambda^* \frac{\phi(\lambda^* Z^*)}{\Phi(\lambda^* Z^*)} Z^*\right],$$
$$S^*_\lambda = \frac{\phi(\lambda^* Z^*)}{\Phi(\lambda^* Z^*)} Z^* \quad \text{and} \quad S^*_\alpha = -\frac{\lambda^*}{2} \frac{\phi(\lambda^* Z^*)}{\Phi(\lambda^* Z^*)} Z^{*3},$$

where $Z^* = (X - \mu^*)/\sigma^*$. Linear dependence does not exist between the elements of the score function when $\lambda^* \neq 0$. Consequently, the information matrix is not singular in this case. In the full parameter case, the vector $n^{1/2}(\hat\mu - \mu^*, \hat\sigma - \sigma^*, \hat\lambda - \lambda^*, \hat\alpha)$, where $(\hat\mu, \hat\sigma, \hat\lambda, \hat\alpha)$ is the MLE of $(\mu, \sigma, \lambda, \alpha)$, converges in distribution to (Y_1, Y_2, Y_3, Y_4), where $(Y_1, Y_2, Y_3, Y_4)^\top$ is a multivariate normal random vector with mean vector $(0,0,0,0)^\top$ and covariance matrix

$$\begin{pmatrix} \frac{1}{\sigma^2}(1+\lambda^2 a_0) & \frac{1}{\sigma^2}\left(\frac{\lambda(2/\pi)^{1/2}(1+2\lambda^2)}{(1+\lambda^2)^{3/2}} + \lambda^2 a_1\right) & \frac{1}{\sigma}\left(\frac{(2/\pi)^{1/2}}{(1+\lambda^2)^{3/2}} - \lambda a_1\right) & \frac{1}{2\sigma}\left(-\frac{3\lambda(2/\pi)^{1/2}}{(1+\lambda^2)^{5/2}} + \lambda^2 a_3\right) \\ & \frac{1}{\sigma^2}(2+\lambda^2 a_2) & -\frac{\lambda}{\sigma} a_2 & \frac{\lambda^2}{2\sigma} a_4 \\ & & a_2 & -\frac{\lambda}{2} a_4 \\ & & & \frac{\lambda^2}{4} a_6 \end{pmatrix}^{-1}$$

where $a_k := a_k(\lambda) = \frac{1}{\pi}\int_0^\infty z^k \phi(\sqrt{1+2\lambda^2}\, z)\left[\frac{(-1)^k}{\Phi(-\lambda z)} + \frac{1}{\Phi(\lambda z)}\right] dz$ for $k = 0, 1, 3, 4, 6$, which have to be evaluated numerically (see Proposition 2 in the Appendix).

2.4. Properties of the MLEs in the normal case

Suppose now that the parameter vector is $\theta^* = (\mu^*, \sigma^*, 0, \alpha^*)^\top$, that is, the data are obtained from a $\mathrm{N}(\mu^*, \sigma^{*2})$ distribution. At $\theta = \theta^*$, the components of S_θ are

$$S^*_\mu = \frac{Z^*}{\sigma^*}, \quad S^*_\sigma = \frac{Z^{*2} - 1}{\sigma^*}, \quad S^*_\lambda = \frac{(2/\pi)^{1/2} Z^*}{(1 + \alpha^* Z^{*2})^{1/2}} \quad \text{and} \quad S^*_\alpha = 0,$$

where $Z^* = (X - \mu^*)/\sigma^*$.

In this case, the components of $(S^*_\mu, S^*_\sigma, S^*_\lambda)$ are linearly independent at least that $\alpha^* = 0$, and so the singularity of the information matrix of θ^* is due to the fact that $S^*_\alpha = 0$. Moreover, the score component of interest S^*_λ cannot be expressed as a linear combination of the components of $(S^*_\mu, S^*_\sigma, S^*_\alpha)$, that is, there is not a vector $\mathbf{c} \neq 0$ of constants such that $S^*_\lambda = \mathbf{c}^\top (S^*_\mu, S^*_\sigma, S^*_\alpha)^\top$, and so the condition (28) considered by Rotnitzky *et al.* (2000) is not satisfied. Consequently, the methodology proposed by these authors cannot be applied to study the asymptotic properties of the MLEs in the normal case ($\lambda = 0$), since there is no vector $\mathbf{c} \neq (0,0,0,0)^\top$ to initialize the iterative process in order to obtain an appropriate reparametrization for which the information matrix is of full rank. As was mentioned above, this conclusion derives from the fact that α is non-identifiable when $\lambda = 0$.

If, in addition, $\alpha^* = 0$, i.e., $\theta^* = (\mu^*, \sigma^*, 0, 0)^\top$, we then find in the above score functions the relation $S^*_\lambda = (2/\pi)^{1/2} \sigma^* S^*_\mu$. Hence, at θ^* the full information matrix has rank 2, which violates the condition (27) of Rotnitzky *et al.* (2000).

A similar fact occurs when $\alpha \to \infty$, which is another form to obtain the normal model. That is, for $\theta^* = (\mu^*, \sigma^*, \lambda^*, \infty)^\top$, we have whatever the value of λ^* that S^*_μ and S^*_σ are as before, but $S^*_\lambda = S^*_\alpha = 0$. Therefore, again the the methodology proposed by Rotnitzky *et al.* (2000) is not appropriated to study the asymptotic properties of the MLEs in the normal case.

However, if the objective is to study the normality hypothesis only, then a natural and convenient strategy is the following:

a) Use the SGN model to test the skew-normality hypothesis $\alpha = 0$ (see Section 2.3).
b) If the skew-normal model is not rejected, then use this model to test the normality hypothesis $\lambda = 0$. In this case, the Rotnitzky *et al.* (2000) methodology (see Chiogna, 2005) as well as the centred parametrization (see Azzalini, 1985) can be used.

Acknowledgments

The research of R. B. Arellano-Valle was supported by Grant FONDECYT (Chile) 1120121|1090411. The work of H. W. Gómez was supported by Grant FONDECYT (Chile) 1090411 and the work of H. S. Salinas was supported by Grant DIUDA (Chile) 221229. The authors thank the editor and two referees whose constructive comments led to a far improved presentation.

Appendix

This appendix provides preliminary calculations needed to derive the elements of the SGN expected information matrix. To simplify the notation, let $W := W(Z) = \frac{\lambda Z}{\sqrt{1+\alpha Z^2}}$ and $R = R(W) = \frac{\phi(W)}{\Phi(W)}$, where $Z \sim \mathrm{SGN}(\lambda, \alpha)$.

Proposition 1 *Let* $\rho_{nm} = E_Z\left(\frac{Z^n R}{(1+\alpha Z^2)^{m/2}}\right)$ *and* $\eta_{nm} = E_Z\left(\frac{Z^n R^2}{(1+\alpha Z^2)^m}\right)$, $n,m = 0,1,\ldots$, *where* $Z \sim SGN(\lambda,\alpha)$. *Then,*

$$\rho_{nm} = \begin{cases} 0, & for\ n = 2k+1\ (odd), \\ E_Y\left(\frac{2Y^{2k}\phi(W(Y))}{(1+\alpha Y^2)^{m/2}}\right), & for\ n = 2k\ (even), \end{cases}$$

and

$$\eta_{nm} = E_Y\left(\left[\frac{(-1)^n}{\Phi(-W(Y))} + \frac{1}{\Phi(W(Y))}\right]\frac{Y^n\phi^2(W(Y))}{(1+\alpha Y^2)^m}\right),$$

where $Y \sim 2\phi(y)I(y \geq 0)$.

Proof: For $n = 2k+1$, we have after a simple algebra that

$$\rho_{nm} = 2\int_{-\infty}^{\infty} \frac{z^{2k+1}\phi(z)}{(1+\alpha z^2)^{m/2}}\phi\left(\frac{\lambda z}{\sqrt{1+\alpha z^2}}\right)dz = 2\int_{-\infty}^{\infty} z h_0(z)dz = 0,$$

since for all $k,m = 0,1,\ldots$, the function $h_0(z) = \frac{z^{2k}\phi(z)}{(1+\alpha z^2)^{m/2}}\phi\left(\frac{\lambda z}{\sqrt{1+\alpha z^2}}\right)$ is even. Similarly, for $n = 2k$, we have

$$\begin{aligned}\rho_{nm} &= 2\int_{-\infty}^{\infty} \frac{z^{2k}\phi(z)}{(1+\alpha z^2)^{m/2}}\phi\left(\frac{\lambda z}{\sqrt{1+\alpha z^2}}\right)dz \\ &= 2\int_0^{\infty} \frac{2y^{2k}\phi(y)}{(1+\alpha y^2)^{m/2}}\phi\left(\frac{\lambda y}{\sqrt{1+\alpha y^2}}\right)dy = 2E_Y\left(\frac{Y^{2k}\phi(W(Y))}{(1+\alpha Y^2)^{m/2}}\right).\end{aligned}$$

Finally, for η_{nm} we have

$$\begin{aligned}\eta_{nm} &= 2\int_{-\infty}^{\infty} \frac{z^n\phi(z)}{(1+\alpha z^2)^m}\frac{\phi^2\left(\frac{\lambda z}{\sqrt{1+\alpha z^2}}\right)}{\Phi\left(\frac{\lambda z}{\sqrt{1+\alpha z^2}}\right)}dz \\ &= 2\int_0^{\infty} \frac{(-y)^n}{\Phi\left(\frac{-\lambda y}{\sqrt{1+\alpha y^2}}\right)}h_1(y)dy + 2\int_0^{\infty} \frac{y^n}{\Phi\left(\frac{\lambda y}{\sqrt{1+\alpha y^2}}\right)}h_1(y)dy \\ &= 2\int_0^{\infty} \left[\frac{(-1)^n}{\Phi(-W(y))} + \frac{1}{\Phi(W(y))}\right]\frac{y^n\phi^2(W(y))}{(1+\alpha y^2)^m}\phi(y)dy \\ &= E_Y\left(\left[\frac{(-1)^n}{\Phi(-W(Y))} + \frac{1}{\Phi(W(Y))}\right]\frac{Y^n\phi^2(W(Y))}{(1+\alpha Y^2)^m}\right),\end{aligned}$$

where it is used that the function $h_1(t) = \frac{\phi^2(W(t))\phi(t)}{(1+\alpha t^2)^m}$ is even for all $m = 0, 1, \ldots$; concluding thus the proof. ■

From Proposition 1 we have, after some straightforward algebra, that the entries $I_{\theta_i\theta_j} = E(S_{\theta_i}S_{\theta_j})$ of the information matrix $\mathbf{I}_\theta$ are as follows:

$$I_{\mu\mu} = E\left(\frac{Z^2}{\sigma^2} - \frac{2\lambda ZR}{\sigma^2(1+\alpha Z^2)^{3/2}} + \frac{\lambda^2R^2}{\sigma^2(1+\alpha Z^2)^3}\right) = \frac{1}{\sigma^2} + \frac{\lambda^2}{\sigma^2}\,\eta_{03},$$

$$\begin{aligned} I_{\mu\sigma} &= E\left(-\frac{Z}{\sigma^2} + \frac{Z^3}{\sigma^2} - \frac{2\lambda Z^2R}{\sigma^2(1+\alpha Z^2)^{3/2}} + \frac{\lambda R}{\sigma^2(1+\alpha Z^2)^{3/2}} + \frac{\lambda^2ZR^2}{\sigma^2(1+\alpha Z^2)^3}\right) \\ &= \frac{2}{\sigma^2}(c_1 - c_0) - \frac{(2/\pi)^{1/2}}{\sigma^2} - \frac{2\lambda}{\sigma^2}\,\rho_{23} + \frac{\lambda}{\sigma^2}\,\rho_{03} + \frac{\lambda^2}{\sigma^2}\,\eta_{13}, \end{aligned}$$

$$I_{\mu\lambda} = E\left(\frac{Z^2R}{\sigma(1+\alpha Z^2)^{1/2}} - \frac{\lambda ZR^2}{\sigma(1+\alpha Z^2)^2}\right) = \frac{1}{\sigma}\,\rho_{21} - \frac{\lambda}{\sigma}\,\eta_{12},$$

$$I_{\mu\alpha} = E\left(-\frac{\lambda Z^4R}{2\sigma(1+\alpha Z^2)^{3/2}} + \frac{\lambda^2Z^3R^2}{2\sigma(1+\alpha Z^2)^3}\right) = -\frac{\lambda}{2\sigma}\,\rho_{43} + \frac{\lambda^2}{2\sigma}\,\eta_{33},$$

$$\begin{aligned} I_{\sigma\sigma} &= E\left(\frac{1}{\sigma^2} - \frac{2Z^2}{\sigma^2} + \frac{2\lambda ZR}{\sigma^2(1+\alpha Z^2)^{3/2}} + \frac{Z^4}{\sigma^2} - \frac{2\lambda Z^3R}{\sigma^2(1+\alpha Z^2)^{3/2}} + \frac{\lambda^2Z^2R^2}{\sigma^2(1+\alpha Z^2)^3}\right) \\ &= \frac{1}{\sigma^2} - \frac{2}{\sigma^2} + \frac{3}{\sigma^2} + \frac{\lambda^2}{\sigma^2}\,\eta_{23} = \frac{2}{\sigma^2} + \frac{\lambda^2}{\sigma^2}\,\eta_{23}, \end{aligned}$$

$$I_{\sigma\lambda} = E\left(-\frac{ZR}{\sigma(1+\alpha Z^2)^{1/2}} + \frac{Z^3R}{\sigma(1+\alpha Z^2)^{1/2}} - \frac{\lambda Z^2R^2}{\sigma(1+\alpha Z^2)^2}\right) = -\frac{\lambda}{\sigma}\,\eta_{22},$$

$$I_{\sigma\alpha} = E\left(\frac{\lambda Z^3R}{2\sigma(1+\alpha Z^2)^{3/2}} - \frac{\lambda Z^5R}{2\sigma(1+\alpha Z^2)^{3/2}} + \frac{\lambda^2Z^4R^2}{2\sigma(1+\alpha Z^2)^3}\right) = \frac{\lambda^2}{2\sigma}\,\eta_{43},$$

$$I_{\lambda\lambda} = E\left(\frac{Z^2R^2}{(1+\alpha Z^2)}\right) = \eta_{21},$$

$$I_{\lambda\alpha} = E\left(-\frac{\lambda Z^4R^2}{2(1+\alpha Z^2)^2}\right) = -\frac{\lambda}{2}\,\eta_{42},$$

$$I_{\alpha\alpha} = E\left(\frac{\lambda^2Z^6R^2}{4(1+\alpha Z^2)^3}\right) = \frac{\lambda^2}{4}\,\eta_{63}.$$

Corollary 1 *If $\lambda = 0$, then the entries $I_{\mu\lambda}$ and $I_{\lambda\lambda}$ of the information matrix $\mathbf{I}_\theta$ reduce to*

$$I_{\mu\lambda} = \frac{2(2/\pi)^{1/2}}{\sigma} d_1(\alpha) \quad \textit{and} \quad I_{\lambda\lambda} = \frac{2}{\pi} d_2(\alpha), \tag{3}$$

where $d_1(\alpha) = \int_0^\infty \frac{z^2\phi(z)}{(1+\alpha z^2)^{1/2}} dz$ *and* $d_2(\alpha) = \frac{1}{\alpha}\left(1 - (2\pi/\alpha)^{1/2} e^{\frac{1}{2\alpha}} \Phi(-\alpha^{-1/2})\right)$.

Proof: In fact, if $\lambda = 0$, then $W \equiv 0$, and so $R \equiv (2/\pi)^{1/2}$. Thus, $I_{\mu\lambda} = \frac{1}{\sigma} E\left(\frac{Z^2}{(1+\alpha Z^2)^{1/2}} R\right)$ $= \frac{(2/\pi)^{1/2}}{\sigma} \int_{-\infty}^{\infty} \frac{z^2\phi(z)}{(1+\alpha z^2)^{1/2}} dz = \frac{2(2/\pi)^{1/2}}{\sigma} \int_0^\infty \frac{z^2\phi(z)}{(1+\alpha z^2)^{1/2}} dz$ since the function $\frac{z^2\phi(z)}{(1+\alpha z^2)^{1/2}}$ is even. Note that this integral has be computed numerically when $\alpha > 0$. For $I_{\lambda\lambda}$ we have $I_{\lambda\lambda} = E\left(\frac{Z^2}{1+\alpha Z^2} R^2\right) = \frac{2}{\pi} \int_{-\infty}^{\infty} \frac{z^2\phi(z)}{1+\alpha z^2} dz = \frac{4}{\pi} \int_0^\infty \frac{z^2\phi(z)}{1+\alpha z^2} dz$ since that function $\frac{z^2\phi(z)}{1+\alpha z^2}$ is even. Hence, the result follows by noting from Mathematica (Wolfram Research, 2008) that

$$d_2(\alpha) := 2\int_0^\infty \frac{z^2\phi(z)}{1+\alpha z^2} dz = \frac{1}{\alpha} - \frac{(\pi/2)^{1/2} e^{\frac{1}{2\alpha}} (1 - \mathrm{erf}(\frac{\sqrt{2}}{2\sqrt{\alpha}}))}{\alpha^{3/2}},$$

for $\alpha > 0$, where $\mathrm{erf}\left(\frac{\sqrt{2}}{2} t\right) = 2\Phi(t) - 1$. ∎

Proposition 2 *Let $Z \sim SGN(\lambda, 0)$. Then*

$$a_k(\lambda) := E_Z\left(Z^k \left\{\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\right\}^2\right) = \frac{1}{\pi} \int_0^\infty z^k \phi(\sqrt{1+2\lambda^2} z) \left[\frac{(-1)^k}{\Phi(-\lambda z)} + \frac{1}{\Phi(\lambda z)}\right] dz.$$

Proof: Since $\phi^2(\lambda z)\phi(z) = \frac{1}{2\pi}\phi(\sqrt{1+2\lambda^2} z)$ we have after a simple algebra that

$$\begin{aligned} E_Z\left(Z^k \left\{\frac{\phi(\lambda Z)}{\Phi(\lambda Z)}\right\}^2\right) &= 2\int_{-\infty}^{\infty} \frac{z^k \phi^2(\lambda z)\phi(z)}{\Phi(\lambda z)} dz \\ &= \frac{1}{\pi} \int_{-\infty}^{\infty} \frac{z^k \phi(\sqrt{1+2\lambda^2} z)}{\Phi(\lambda z)} dz \\ &= \frac{1}{\pi} \int_0^\infty z^k \phi(\sqrt{1+2\lambda^2} z) \left[\frac{(-1)^k}{\Phi(-\lambda z)} + \frac{1}{\Phi(\lambda z)}\right] dz. \end{aligned}$$

∎

References

Arellano-Valle, R. B., Gómez, H. W. and Quintana, F. A. (2004). A new class of skew-normal distributions. *Communications in Statistics: Theory and Methods*, 33(7), 1465–1480.

Azzalini, A. (1985). A class of distributions which includes the normal ones. *Scandinavian Journal of Statistics*, 12, 171–178.

Chiogna, M. (2005). A note on the asymptotic distribution of the maximum likelihood estimator for the scalar skew-normal distribution. *Statistical Methods & applications*, 14, 331–341.

Rotnitzky, A., Cox, D. R., Bottai, M. and Robins, J. (2000). Likelihood-based inference with singular information matrix. *Bernoulli*, 6, 243–284.

Wolfram Research, Inc., Mathematica, Version 7.0, Champaign, IL (2008).

SORT 37 (1) January-June 2013, 29-56

The normal distribution in some constrained sample spaces

Glòria Mateu-Figueras[1], Vera Pawlowsky-Glahn[1] and Juan José Egozcue[2]

Abstract

Phenomena with a constrained sample space appear frequently in practice. This is the case, for example, with strictly positive data, or with compositional data, such as percentages or proportions. If the natural measure of difference is not the absolute one, simple algebraic properties show that it is more convenient to work with a geometry different from the usual Euclidean geometry in real space, and with a measure different from the usual Lebesgue measure, leading to alternative models that better fit the phenomenon under study. The general approach is presented and illustrated using the normal distribution, both on the positive real line and on the *D*-part simplex. The original ideas of McAlister in his introduction to the lognormal distribution in 1879, are recovered and updated.

MSC: 60A10, 60E10, 62E10.

Keywords: Additive logistic normal distribution, Aitchison measure, Lebesgue measure, lognormal distribution, orthonormal basis, simplex.

1. Introduction

In general, continuous multivariate observations are assumed to be real random vectors which density functions are defined with respect to the Lebesgue measure. The Lebesgue measure is compatible with the inner vector space structure of real space and thus natural in $\mathbb{R}$. When random vectors are defined on a constrained sample space, $E \subset \mathbb{R}^D$, methods and concepts used in real space can lead to nonsensical results. For example, for positive random variables, the usual confidence interval $\bar{x} \pm kS$, where S is the standard deviation, can include negative values. In the case of random compositions, i.e., of random vectors

[1] Dept. d'Informàtica i Matemàtica Aplicada, Campus Montilivi, Universitat de Girona, Spain. Email: gloria.mateu@udg.edu
[2] Dept. de Matemàtica Aplicada III, Universitat Politècnica de Catalunya, Barcelona, Spain.
Received: December 2011
Accepted: September 2012

defined on the simplex that represent proportions of some whole, problems appear if correlations between components are used. This is a well-known problem stated by Pearson (1897) and called *spurious correlation*. A way to avoid these problems is to use transformations, such as the logarithm on the positive real line or logratio transformations on the simplex (Aitchison, 1986). There is a long history behind the logarithmic and the logratio transformations. The well-known lognormal and the logistic normal distributions were introduced in $\mathbb{R}_+$ and in the simplex, respectively, through these transformations. In this contribution, we revise those definitions and propose a common and new theory to introduce a normal distribution in constrained sample spaces. In particular, we focus on a general constrained sample space, $E \subset \mathbb{R}^D$, which admits a meaningful Euclidean vector space structure, possibly different from the usual structure of real space. The idea, previously used in Eaton (1983), is that, for any Euclidean vector space E with a one-to-one transformation to $\mathbb{R}^D$, a measure λ_E, compatible with its structure, is obtained from the respective structure of $\mathbb{R}^D$ and its Lebesgue measure. This allows us to define density functions on E considering the measure λ_E or, equivalently, the corresponding density functions of the coordinates.

Every one-to-one transformation between a set E and real space induces a real Euclidean vector space structure in E, with associated measure λ_E. Particularly interesting are those transformations that are meaningful and related to the measure of difference between observations. This idea can be found in Galton (1879), as an introduction to the logarithmic transformation as a means to acknowledge Fechner's law, according to which *perception equals log(stimulus)*. The idea was then formalised by McAlister (1879). This approach has acquired a growing importance in applications, due to the fact that some *constrained sample spaces*, which are subsets of some real space — like $\mathbb{R}_+$ or the simplex — can be structured as Euclidean vector spaces (Pawlowsly-Glahn and Egozcue, 2001). It is important to emphasize that this approach implies using a measure which is different from the usual Lebesgue measure.

The advantage of this approach is that it opens the door to study statistical models using a measure which is considered to be appropriate or natural for the studied phenomenon, instead of the ordinary Lebesgue measure. Here we apply this idea to the normal distribution on the two mentioned constrained sample spaces, the positive real line, $\mathbb{R}_+$, and the simplex, $\mathcal{S}^D$. They are well known as the lognormal distribution and the additive logistic normal distribution when expressed with respect to the Lebesgue measure. We focus on their representation when the reference measure is the measure associated to the Euclidean vector space structure of the sample space. While the probability law is the same, the change of representation produces a change in some characteristic values of the distribution. Also, some invariance properties of normal distributions appear as natural within the structure of the sample space. These properties usually get lost when representing these distributions with respect to the Lebesgue measure. The idea of using not only an interpretable space structure, but also to change the measure, is a powerful tool because it leads to results coherent with the interpretation of the measure of difference, and because they are mathematically more straightforward.

Section 2 describes some technical details in an abstract setting concerning Euclidean vector spaces, their reference measure, and the definition of the normal probability density functions on them. Sections 3 and 4 present the application of these concepts to the positive real line and the simplex, respectively, as well as some examples on normal modelling in these constrained spaces.

2. Probability densities in Euclidean vector spaces

Let $E \subseteq \mathbb{R}^D$ be the sample space for a random vector $\mathbf{X}$, i.e. each realization of $\mathbf{X}$ is in E. Assume there exists a one-to-one, differentiable, mapping $h : E \to \mathbb{R}^d$ with $d \leq D$. In the case of the positive real line, $E = \mathbb{R}_+$ and $d = D = 1$, i.e. $\mathbb{R}$ and $\mathbb{R}_+$ have the same dimension. This is not the case of the simplex $\mathcal{S}^D$, which consists of vectors of D positive components adding up to a fixed constant. Only $d = D-1$ components are required to specify a point in it, i.e. the dimension of $\mathcal{S}^D$ is $d = D-1 < D$. The mapping h allows to define a Euclidean vector structure on E just translating the standard properties of $\mathbb{R}^d$ into E. The existence of the mapping h implies some characteristics of E. An important one is that E must have some border set so that h transforms neighbourhoods of this border into neighbourhoods of infinity in $\mathbb{R}^d$. For instance, a sphere in $\mathbb{R}^3$ with a defined pole can be transformed into $\mathbb{R}^2$, but, if no pole is defined, this is no longer possible.

The vector addition or internal operation $\oplus$ and the scalar multiplication or external operation $\odot$ in E are defined as

$$\mathbf{x} \oplus \mathbf{y} = h^{-1}(h(\mathbf{x}) + h(\mathbf{y})) \,, \quad \alpha \odot \mathbf{x} = h^{-1}(\alpha \cdot h(\mathbf{x})) \,,$$

for $\mathbf{x}, \mathbf{y} \in E$ and $\alpha \in \mathbb{R}$. With these definitions, E is a d-dimensional vector space. The metric structure is induced by the inner product $\langle \mathbf{x}, \mathbf{y} \rangle_E = \langle h(\mathbf{x}), h(\mathbf{y}) \rangle$. It implies the norm, $\|\mathbf{x}\|_E = \|h(\mathbf{x})\|$, and the distance, $\mathrm{d}_E(\mathbf{x}, \mathbf{y}) = \mathrm{d}(h(\mathbf{x}), h(\mathbf{y}))$, thus completing the Euclidean vector space structure of E. This structure is derived from the inner product, norm and distance in $\mathbb{R}^d$, denoted as $\langle \cdot, \cdot \rangle$, $\| \cdot \|$ and $\mathrm{d}(\cdot, \cdot)$, respectively. By construction, $h(\mathbf{x})$ is the vector of coordinates of $\mathbf{x} \in E$. The coordinates correspond to the orthonormal basis in E given by the images of the canonical basis in $\mathbb{R}^d$ by h^{-1}. The origin of the space E is then $h^{-1}(\vec{\mathbf{0}})$ where $\vec{\mathbf{0}}$ is the neutral element of $\mathbb{R}^d$ with respect to the ordinary sum. The Lebesgue measure in $\mathbb{R}^d$, λ_d, induces a measure in E, denoted λ_E, using the fact that h is one-to-one and setting $\lambda_E(h^{-1}(B)) = \lambda_d(B)$, for any Borelian B in $\mathbb{R}^d$ (Eaton, 1983). This idea was used in Pawlowsky-Glahn (2003) to define the Aitchison measure on the simplex.

In order to define probability density functions (pdf's) in E, a reference measure is needed. A pdf is the Radon–Nikodym derivative of a probability measure P with respect to a measure on E. When the reference measure is λ_E, we denote the pdf as $f^E = dP/d\lambda_E$. When E is viewed as a subset of $\mathbb{R}^D$, the pdf with respect to the Lebesgue measure λ_D could be eventually considered. However, if $d < D$, the random vector $\mathbf{X}$

cannot be absolutely continuous with respect to λ_D and the pdf does not exist. Our approach, and a more natural way to define a pdf for $\mathbf{X}$, is to start with a pdf for the (random) coordinates $\mathbf{Y} = h(\mathbf{X})$ in $\mathbb{R}^d$. Assume that $f_{\mathbf{Y}}$ is the pdf of $\mathbf{Y}$ with respect to the Lebesgue measure, λ_d, in $\mathbb{R}^d$, i.e. $\mathbf{Y}$ is absolutely continuous with respect to λ_d and the pdf is the Radon–Nikodym derivative $f_{\mathbf{Y}} = dP/d\lambda_d$. The random vector $\mathbf{X}$ is recovered from $\mathbf{Y}$ as $\mathbf{X} = h^{-1}(\mathbf{Y})$. When $D > d$, $h^{-1}(\mathbf{Y})$ can be expressed using only d of its components. Let h_d^{-1} be such a restriction and $\mathbf{X}_d = h_d^{-1}(\mathbf{Y})$. The inverse mapping is denoted by $h_d(\mathbf{X}_d)$ and it holds that $h_d(\mathbf{X}_d) = h(\mathbf{X})$. This means that more than d components of $\mathbf{X}$ are redundant. When $D = d$, the restriction of h^{-1} reduces to $h_d^{-1} = h^{-1}$. For instance, to recover a vector in the simplex of D components from its representation using $d = D-1$ coordinates, one can recover $d = D-1$ components. The remaining one is obtained from the constant sum of all components. The pdf of $\mathbf{X}_d$ with respect to the Lebesgue measure in $\mathbb{R}^d$ is computed using the Jacobian rule

$$f_{\mathbf{X}_d}(\mathbf{x}_d) = \frac{dP}{d\lambda_d}(\mathbf{x}_d) = f_{\mathbf{Y}}(h_d(\mathbf{x}_d)) \cdot \left| \frac{\partial h_d(\mathbf{x}_d)}{\partial \mathbf{x}_d} \right| , \tag{1}$$

where the last term is the d-dimensional Jacobian of h_d. The next step is to express the pdf with respect to λ_E, the compatible measure in E. The chain rule for Radon–Nikodym derivatives implies

$$f^E_{\mathbf{X}_d}(\mathbf{x}_d) = \frac{dP}{d\lambda_E}(\mathbf{x}_d) = \frac{dP}{d\lambda_d}(\mathbf{x}_d) \cdot \frac{d\lambda_d}{d\lambda_E}(\mathbf{x}_d) , \tag{2}$$

and, due to the inverse function theorem, the last derivative is

$$\frac{d\lambda_d}{d\lambda_E}(\mathbf{x}_d) = \left| \frac{\partial h_d^{-1}(h_d(\mathbf{x}_d))}{\partial \mathbf{y}} \right| = \left| \frac{\partial h_d(\mathbf{x}_d)}{\partial \mathbf{x}_d} \right|^{-1} , \tag{3}$$

Substituting (2) and (3) into (1),

$$f^E_{\mathbf{X}}(\mathbf{x}) = \frac{dP}{d\lambda_E}(\mathbf{x}) = f_{\mathbf{Y}}(h(\mathbf{x})) , \tag{4}$$

where the subscripts d have been suppressed, as they only play a role when computing the Jacobian. Difficulties using $f^E_{\mathbf{X}}$, arising from the fact that the integral $\mathrm{P}(A) = \int_A f^E_{\mathbf{X}}(\mathbf{x}) d\lambda_E(\mathbf{x})$ is not an integral with respect to the Lebesgue measure in $\mathbb{R}^d$ but with respect to the Lebesgue type measure in E, are solved working with coordinates. Particularly, they are solved working with coordinates with respect to an orthonormal basis in E. Using (4) the probability of an event $A \subseteq E$ can be computed as $\mathrm{P}(A) = \int_{h(A)} f_{\mathbf{Y}}(h(\mathbf{x}))\, d\lambda_d(h(\mathbf{x}))$ or, in simpler notation, $\mathrm{P}(A) = \int_{h(A)} f_{\mathbf{Y}}(\mathbf{y})\, d\mathbf{y}$.

The representation of the distribution of random variables by pdf's defined with respect to the measure λ_E requires a review of the moments and other characteristics of the pdf's. Following Eaton (1983), the expectation and variance of $\mathbf{X}$ are defined as follows:

Definition 2.1 *Let* $\mathbf{X}$ *be a random variable supported on* E *and* $h : E \to \mathbb{R}^d$ *the coordinate function defined on* E*. The expectation in* E *is*

$$\mathrm{E}^E[\mathbf{X}] = h^{-1}\left(\int_{\mathbb{R}^d} \mathbf{y}\, f_{h(\mathbf{X})}(\mathbf{y})\, d\mathbf{y}\right) \tag{5}$$

$$= h^{-1}\left(\mathrm{E}[h(\mathbf{X})]\right), \tag{6}$$

provided the integral in (5) exists in the Lebesgue sense.

Intuitively, the expectation (5) in E consists of representing the elements of E using coordinates and to integrate using the pdf of the coordinates; the result is transformed back into E. Equation (6) summarizes this result using the standard definition of expectation of the coordinates in $\mathbb{R}^d$.

The variance involves only real expectations and can be identified with the variance of coordinates. Special attention deserves the metric or total variance (Aitchison, 1986; Pawlowsly-Glahn and Egozcue, 2001). Assuming the existence of the integrals, the metric variability of $\mathbf{X}$ with respect to a point $\mathbf{z} \in E$ is defined as $\mathrm{Var}^E[\mathbf{X}, \mathbf{z}] = \mathrm{E}[\mathrm{d}_E^2(\mathbf{X}, \mathbf{z})]$. The minimum metric variability is attained for $\mathbf{z} = \mathrm{E}^E[\mathbf{X}]$, thus supporting the definition in (5)–(6). The metric variance is then

$$\mathrm{Var}^E[\mathbf{X}] = \mathrm{E}[\mathrm{d}_E^2(\mathbf{X}, \mathrm{E}^E[\mathbf{X}])]\ . \tag{7}$$

The mode of a pdf is normally defined as its maximum value, although local maxima are also frequently called modes. However, the shape and, particularly, the maximum values depend on the reference measure taken in the Radon-Nikodym derivatives of the density. Since the Lebesgue measure in the coordinate space, $\mathbb{R}^d$, corresponds to the measure λ_E in E, the mode can be defined as

$$\mathrm{Mode}^E[\mathbf{X}] = \underset{\mathbf{x}\in E}{\mathrm{argmax}}\{f_{\mathbf{X}}^E(\mathbf{x})\} = h^{-1}\left(\underset{\mathbf{y}\in\mathbb{R}^d}{\mathrm{argmax}}\{f_{h(\mathbf{X})}(\mathbf{y})\}\right)\ .$$

3. The positive real line

The real line, with the ordinary sum and product by scalars, has a vector space structure. The ordinary inner product and the Euclidean distance are compatible with these

operations, i.e. they satisfy the translation invariance and the homogeneity properties. But this geometry is not suitable for the positive real line. Confront, for example, some meteorologists with two pairs of samples taken at two rain gauges, $\{5;10\}$ and $\{100;105\}$ in mm, and ask for the difference; quite probably, in the first case they will say there was double the total rain in the second gauge compared to the first, while in the second case they will say it rained a lot but approximately the same. They are assuming a relative measure of difference. Consequently, for them the natural measure of difference is not the usual Euclidean one, and the ordinary vector space structure of $\mathbb{R}$ does not behave suitably for the problem. In fact, problems might appear shifting a positive number (vector) by a negative real number (vector); or multiplying a positive number (vector) by an arbitrary real number (positive or negative scalar), because results can be outside $\mathbb{R}_+$.

There are two operations, $\oplus$, $\odot$, which induce a vector space structure in $\mathbb{R}_+$ (Pawlowsly-Glahn and Egozcue, 2001). In fact, given $x, x^* \in \mathbb{R}_+$, the internal operation, which plays an analogous role to addition in $\mathbb{R}$, is the usual product $x \oplus x^* = x \cdot x^*$ and, for $\alpha \in \mathbb{R}$, the external operation, which plays an analogous role to the product by scalars in $\mathbb{R}$, is $\alpha \odot x = x^\alpha$. An inner product, compatible with $\oplus$ and $\odot$ is $\langle x, x^* \rangle_+ = \ln x \cdot \ln x^*$, which induces a norm, $\|x\|_+ = |\ln x|$, and a distance, $\mathrm{d}_+(x, x^*) = |\ln x^* - \ln x|$, and thus the complete Euclidean vector space structure in $\mathbb{R}_+$. Since $\mathbb{R}_+$ is a one-dimensional vector space, there are only two orthonormal bases: the unit-vector (e) and its inverse element with respect to the internal operation (e^{-1}). From now on the first option is considered and it will be denoted by e. Any $x \in \mathbb{R}_+$ can be expressed as $x = \ln x \odot e = e^{\ln x}$ which reveals that $h(x) = \ln x$ is the coordinate of x with respect to the basis e. The measure λ_+ in $\mathbb{R}_+$ can be defined so that, for λ_1 the Lebesgue measure in $\mathbb{R}^1$, and an interval $(a,b) \subset \mathbb{R}_+$, $\lambda_+(a,b) = \lambda_1(\ln a, \ln b) = |\ln b - \ln a|$ and $d\lambda_+ / d\lambda_1 = 1/x$ (Mateu-Figueras, 2003). Following the notation in Section 2, all these definitions can be obtained by setting $E = \mathbb{R}_+$, $D = d = 1$ and $h(x) = \ln x$. The generalization to $E = \mathbb{R}_+^D$ is straightforward: for $\mathbf{x} \in \mathbb{R}_+^D$, the coordinate function can be defined as $h(\mathbf{x}) = \ln(\mathbf{x})$, where the logarithm applies component-wise.

3.1. The normal distribution on $\mathbb{R}_+$

Using the algebraic-geometric structure in $\mathbb{R}_+$ and the measure λ_+, the normal distribution on $\mathbb{R}_+$ was defined in Mateu-Figueras et al. (2002) through the density function of orthonormal coordinates.

Definition 3.1 *Let $(\Omega, \mathcal{F}, P)$ be a probability space. A random variable $X : \Omega \longrightarrow \mathbb{R}_+$ is said to have a normal on $\mathbb{R}_+$ distribution with two parameters μ and σ^2, written $\mathcal{N}_+(\mu, \sigma^2)$, if its density function with respect to λ_+ is*

$$f_X^+(x) = \frac{dP}{d\lambda_+}(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2} \frac{(\ln x - \mu)^2}{\sigma^2} \right), \qquad x \in \mathbb{R}_+. \tag{8}$$

The density (8) is the usual normal density applied to coordinates $\ln x$ as implied by (4) and it is a density in $\mathbb{R}_+$ with respect to the λ_+ measure. This density function is completely restricted to $\mathbb{R}_+$ and its expression corresponds to the law of frequency introduced by McAlister (1879). The probability law corresponding to the density (8) is that of the lognormal distribution, denoted Λ, where μ and σ^2 are the logarithmic mean and variance. The continuous line in Figure 1 represents the density function (8) for $\mu = 0$ and $\sigma^2 = 1$. Note that the areas under the log-normal density f_X are proportional to probabilities, whereas areas under f_X^+, as shown in the figure, are not. In the case of f_X^+ a probability is proportional to the ordinate of the curve times the length of dx, i.e. times $\lambda_+(x, x+dx) = |\ln(x+dx) - \ln(x)|$.

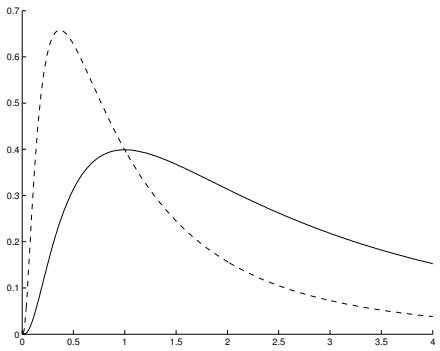


Figure 1: *Density function f_X^+ (——) and f_X (- - - -) with $\mu = 0$ and $\sigma = 1$.*

According to this approach, the normal distribution in $\mathbb{R}_+$ exhibits properties analogous to the normal distribution in $\mathbb{R}$, the most relevant of which are summarized in the following properties. The corresponding proofs are presented in the appendix.

Property 3.1 *Let $X \sim \mathcal{N}_+(\mu, \sigma^2)$, $a \in \mathbb{R}_+$ and $b \in \mathbb{R}$. Then, the random variable $X^* = a \oplus (b \odot X) = a \cdot X^b$ is distributed as $\mathcal{N}_+(\ln a + b\mu, b^2\sigma^2)$.*

Property 3.2 *Let $X \sim \mathcal{N}_+(\mu, \sigma^2)$ and $a \in \mathbb{R}_+$. Then, $f^+_{a \oplus X}(a \oplus x) = f_X^+(x)$, where f_X^+ and $f^+_{a \oplus X}$ represent the probability density functions of the random variables X and $a \oplus X = a \cdot X$, respectively.*

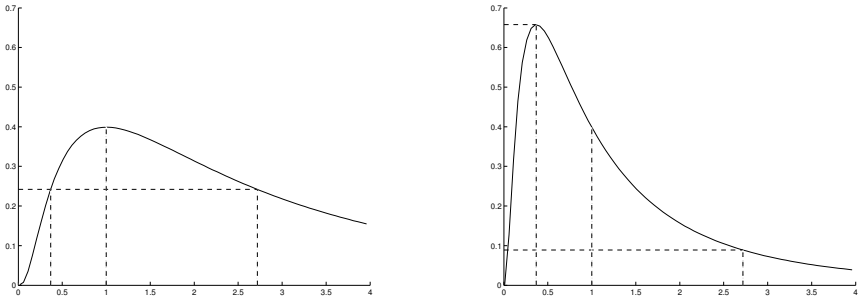
Property 3.3 *If $X \sim \mathcal{N}_+(\mu, \sigma^2)$, then $\mathrm{E}^+[X] = \mathrm{Med}^+[X] = \mathrm{Mode}^+[X] = e^{\mu}$.*

Property 3.4 *If $X \sim \mathcal{N}_+(\mu, \sigma^2)$, then $\mathrm{Var}^+[X] = \sigma^2$.*

Note that property 3.1 implies that the family $\mathcal{N}_+(\mu, \sigma^2)$ is closed under the operations in $\mathbb{R}_+$ and property 3.2 asserts the equivariance under translations in $\mathbb{R}_+$.

The expected value, the median and the mode are elements of the support space $\mathbb{R}_+$, but the variance is only a numerical value which describes the dispersion of X. We are

used to taking the square root of σ^2 as a means to represent intervals centred at the mean and with radius equal to some standard deviations. Such an interval, centred at $\mathrm{E}^+[X] = e^{\mu}$ and with length $2k\sigma$, is $(e^{\mu-k\sigma}, e^{\mu+k\sigma})$, as $\mathrm{d}_+(e^{\mu-k\sigma}, e^{\mu+k\sigma}) = 2k\sigma$. This kind of interval is used in practice (Ahrens, 1954); for instance, under log-normality assumption, predictive intervals in $\mathbb{R}_+$ are computed on log-transformed data, and then back-transformed using exponentiation. In figure 2(a) we represent the interval $(e^{\mu-\sigma}, e^{\mu+\sigma})$ for a $\mathcal{N}_+(\mu, \sigma^2)$ density function with $\mu = 0$ and $\sigma^2 = 1$. It can be shown that it is of minimum length in $\mathbb{R}_+$, and also an isodensity interval, as the distribution is symmetric around e^{μ} in $\mathbb{R}_+$. This symmetry might seem paradoxical, in view of the shape of the density function. But still, it is symmetric within the Euclidean vector space structure of $\mathbb{R}_+$, although certainly not within the space structure of $\mathbb{R}$.



(a) *(b)*

Figure 2: *Dashed line: interval* $(e^{\mu-\sigma}, e^{\mu+\sigma})$*; (a)* f_X^+*, corresponding to* $\mathcal{N}_+(\mu = 0, \sigma^2 = 1)$*, (b)* f_X *corresponding to* $\Lambda(\mu = 0, \sigma^2 = 1)$*.*

An important aspect of this approach is that consistent estimators and exact confidence intervals for the expected value are easy to obtain. It is enough to take exponentials of those obtained from normal theory using log-transformed data, i.e. using the coordinates with respect to the orthonormal basis. Thus, let $x_1, x_2, \ldots, x_n$ be a random sample and $y_i = \ln x_i$ for $i = 1, 2, \ldots, n$. Then, the optimal estimator for the mean of a normal in $\mathbb{R}_+$ population is the geometric mean $(x_1 x_2 \cdots x_n)^{1/n}$, that equals to $e^{\bar{y}}$. An exact $(1-\alpha)100\%$ confidence interval for the mean is $(e^{\bar{y}-t_{\alpha/2}S/\sqrt{n}}, e^{\bar{y}+t_{\alpha/2}S/\sqrt{n}})$, where S denotes the standard deviation of the log transformed sample and $t_{\alpha/2}$ the $(1-\alpha/2)$ t-student ($n-1$ d.f.) quantile.

The normal distribution plays a relevant role in statistics mainly due to its relationship with the central limit theorem. The central limit theorem for the log-normal model is well-known (Aitchison and Brown, 1957). The sums of random variables are replaced by multiplications of positive variables and the limiting distribution is the log-normal. As the central limit theorem concerns the limiting probability law and the multiplication of random variables, it can be translated into terms of the normal in $\mathbb{R}_+$. Let X_1, X_2, ..., X_n be a sequence of random variables in $\mathbb{R}_+$. Define the coordinates of X_i as $Y_i = \ln X_i$ for

$i = 1,2,\ldots$, and assume they are mutually independent and identically distributed with mean μ and variance σ^2. The standardized variable $(1/n)\sum_{i=1}^{n}(Y_i-\mu)/\sigma$ converges in law, for $n \to \infty$, to a random variable with standard normal distribution due to the central limit theorem. Transforming back using $h^{-1} = \exp$, the central limit theorem for the normal in $\mathbb{R}_+$ yields: the random standardized geometric mean

$$Z_n = \frac{1}{n\sigma} \odot \bigoplus_{i=1}^{n} [X_i \ominus \exp(\mu)] = \left[\prod_{i=1}^{n} \left(\frac{X_i}{\exp(\mu)} \right)^{1/\sigma} \right]^{1/n},$$

converges in law, as $n \to \infty$, to a random variable distributed $\mathcal{N}_+(1,0)$.

The role of the operations $\oplus$ and $\odot$ (in $\mathbb{R}_+$ they are the multiplication and powering) in the central limit theorem is remarkable. Its relevance relies on the fact that the operations on random variables involved are interpretable and of frequent use.

3.2. Normal on $\mathbb{R}_+$ versus lognormal

The lognormal distribution has long been recognized as a useful model in the evaluation of random phenomena whose distribution is positive and skew, and specially when dealing with measurements in which the random errors are multiplicative rather than additive. The history of this distribution dates back to 1879, when Galton (1879) observed that the *law of frequency of errors* was incorrect in many groups of phenomena. This observation was based on Fechner's law which, in its approximate and simplest form, is *sensation=log(stimulus)*. According to this law, an error of the same magnitude in excess or in deficiency (in the absolute sense) is not equally probable; therefore, he proposed the geometric mean as a measure of the most probable value instead of the arithmetic mean. This remark was followed by the memoir of McAlister (1879), where a mathematical development concluding with the lognormal distribution was performed. He proposed a practical and easy method for the treatment of a data set grouped around its geometric mean: *convert the observations into logarithms and treat the transformed data set as a series round its arithmetic mean*, and introduced a density function called the *law of frequency* which is the normal density function applied to the log-transformed variable, i.e. the density (8). In order to compute probabilities in given intervals, he introduced also the *law of facility*, nowadays known as the lognormal density function (9).

A unified treatment of lognormal theory is presented in Aitchison and Brown (1957); recent developments are compiled in Crow and Shimizu (1988). A great number of authors use the lognormal model from an applied point of view. Their approach assumes $\mathbb{R}_+$ to be a subset of the real line with the usual Euclidean geometry restricted to it. This is how everybody understands the sentence *an error of the same magnitude in excess or in deficiency* in the same way. One might ask oneself why there is much to say about the lognormal distribution if the data analysis can be referred to the intensively studied normal distribution by taking logarithms. One of the generally accepted reasons is that

parameter estimates are biased if obtained from the inverse transformation. As noted above, the normal on $\mathbb{R}_+$ distribution is well known as the lognormal distribution. But the proposed change of representation produces differences in some properties which are studied below.

Recall that a positive random variable X is said to be lognormally distributed with two parameters μ and σ^2 if $Y = \ln X$ is normally distributed with mean μ and variance σ^2. We write $X \sim \Lambda(\mu, \sigma^2)$. Its probability density function is

$$f_X(x) = \begin{cases} \frac{1}{\sqrt{2\pi}\sigma x} \exp\left(-\frac{1}{2}\left(\frac{\ln x - \mu}{\sigma}\right)^2\right) & x > 0, \\ 0 & x \leq 0. \end{cases} \tag{9}$$

Comparing (9) with (8), subtle differences can be observed. One is that (9) includes a case for the zero and for the negative values of the random variable. This fact is paradoxical, because the lognormal model is completely restricted to $\mathbb{R}_+$. It is forced by the fact that $\mathbb{R}_+$ is considered as a subset of $\mathbb{R}$ with the restricted structure and, consequently, the variable is assumed to be a real random variable, hence the name *lognormal distribution in* $\mathbb{R}$. Another difference lies in the coefficient $1/x$, the Jacobian, which is necessary to work with real analysis in $\mathbb{R}$. In the lognormal case the Jacobian is necessary because the density is written with respect to the Lebesgue measure, but in the normal in $\mathbb{R}_+$ case the Jacobian is not necessary as the density is expressed with respect to λ_+. More obvious differences are that (9) is not equivariant under translations and is not symmetric around the mean. Note that for the lognormal case, $\mathrm{E}[X] = e^{\mu + \frac{1}{2}\sigma^2}$, the $\mathrm{Med}[X] = e^{\mu}$ and $\mathrm{Mod}[X] = e^{\mu - \sigma^2}$. Using our approach a different expected value and a different mode are obtained, while the value for the median is the same. The dashed line in Figure 1 illustrates the probability density function (9) for $\mu = 0$ and $\sigma^2 = 1$. It clearly differs from the density function (8) plotted in continuous line.

As for the normal in $\mathbb{R}_+$ case, an interval centered at the mean and with radius equal to some standard deviations can be represented for the lognormal in $\mathbb{R}$. Considering $\mathbb{R}_+$ as a subset of $\mathbb{R}$ with an Euclidean structure, such an interval is: $(\mathrm{E}[X] - k\mathrm{Stdev}[X], \mathrm{E}[X] + k\mathrm{Stdev}[X])$. But it has no sense, because the lower bound might take a negative value. For example, for $\mu = 0$ and $\sigma^2 = 1$, the above interval with $k = 1$ is $(-0.512, 3.810)$. This is the reason why sometimes intervals $(e^{\mu - k\sigma}, e^{\mu + k\sigma})$ are used, which are considered to be *non-optimal*, because they are neither isodensity intervals, nor do they have minimum length. In Figure 2(b) we represent the interval $(e^{\mu-\sigma}, e^{\mu+\sigma})$ for $\Lambda(\mu, \sigma^2)$ with $\mu = 0$ and $\sigma^2 = 1$. It is clear that in the bounds of the interval the density function takes different values.

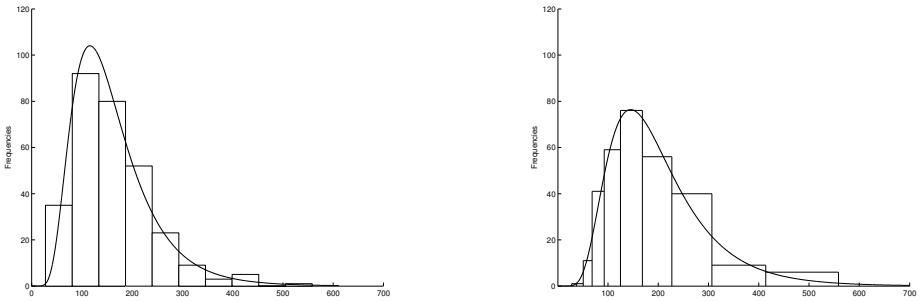
Consistent estimators and exact confidence intervals for the mean and the variance of a lognormal variable are difficult to compute. Early methods are summarised by Aitchison and Brown (1957) and Crow and Shimizu (1988). In the literature an extensive number of procedures and discussions can be found. It is not the objective of this paper

to summarise them all and to provide a complete set of formulas. For the mean, the term $e^{\bar{y}}$ multiplied by a term depending on σ, expressed as an infinite series or tabulated in a set of tables, is obtained in most cases (Aitchison and Brown, 1957; Krige, 1981; Clark and Harper, 2000). For example, Sichel (1996) optimal estimator for the mean of a lognormal population is used by Clark and Harper (2000). This estimator is obtained as $e^{\bar{y}}\gamma$, where γ is a bias correction factor depending on the variance and the size of the data set. It is tabulated in a set of tables. A similar bias correction factor is used to obtain confidence intervals on the population mean (Clark and Harper, 2000). Nevertheless, in practical situations, sometimes the sample geometric mean, $e^{\bar{y}}$, is used to represent the mean and in some cases also to represent the mode of a lognormally distributed variable (Herdan, 1960). But, as adverted by Crow and Shimizu (1988), those affirmations cannot be justified using the lognormal theory. On the contrary, using the normal in $\mathbb{R}_+$ approach, those affirmations are completely justified.

3.3. Example

The differences between using a density with respect to the Lebesgue measure λ or a density with respect to the measure λ_+ can be best appreciated in practice. In order to compare the classical lognormal estimators with those obtained by the normal in $\mathbb{R}_+$ approach, we have simulated 300 samples representing sizes of oil fields in thousands of barrels, a geological variable often lognormally modelled (Davis, 1986). The objective with this simple example is to estimate a suitable location parameter and a corresponding confidence interval and to compare the results obtained using the lognormal approach with the results obtained using the proposed approach. Using the classical lognormal procedures and Table A2 provided by Aitchison and Brown (1957) we obtain 161.96 as an estimate for the mean. Afterwards, using Tables 1,2 and 3 given by Krige (1981), we obtain 162.00 and $(150.31, 176.78)$ as an estimate and approximate 90% confidence interval for the mean. Also, using Tables 7, 8(b) and 8(e) provided by Clark and Harper (2000), we could apply Sichel's bias correction to obtain 161.86 and $(144.07, 188.39)$ as the optimal estimator and confidence interval for the mean in the context of the lognormal approach.

Using the normal in $\mathbb{R}_+$ approach we obtain 145.04 as the estimate for the mean and $(138.70, 151.68)$ as the exact 90% confidence interval for the mean. Logically, different values are obtained, as different methodologies are used. The mean is obtained as $e^{\bar{y}}$ and it is not necessary to apply any bias correction, as unbiasedness is in this case equivalent to unbiasedness in coordinates. The confidence interval is obtained as $(e^{\bar{y}-t_{0.05}S/\sqrt{n}}, e^{\bar{y}+t_{0.05}S/\sqrt{n}})$ where S denotes the sample standard deviation of the log transformed sample. Note that only exponentials of the mean and the 90% confidence interval obtained from normal theory using log-transformed data are taken. As can be observed, the differences to those obtained using the lognormal approach are important. With the normal in $\mathbb{R}_+$ a much more conservative result is obtained, although it is consistent with the assumed geometry of $\mathbb{R}_+$.



(a) *(b)*

Figure 3: *Simulated sample, $n = 300$. (a) Histogram and fitted lognormal density; the Lebesgue-lengths of the bins are equal. (b) Display analogous to a histogram and fitted normal in $\mathbb{R}_+$ density; the λ_+-lengths of the bins are equal.*

In order to compare graphically the normal in $\mathbb{R}_+$ and the lognormal approaches we can represent the histogram with the corresponding fitted densities. In Figures 3(a) and 3(b) the histogram with the fitted lognormal and normal in $\mathbb{R}_+$ densities are shown. Note that the intervals of the histogram are of equal length in both cases, as the absolute Euclidean distance is used in (a) and the relative distance in $\mathbb{R}_+$ is used in (b) to compute them. Thus, (b) is a display analogous to a histogram, based on the structure defined in Section 3. Finally, in Figure 4 the histogram of the log-transformed data or, equivalently, of the coordinates with respect to the orthonormal basis, with the fitted normal density, is provided. This last figure is adequate using both methodologies, but in this case we have chosen exactly the same intervals as in Figure 3(b). This is only possible using the normal on $\mathbb{R}_+$ approach, as the intervals on the positive real line have corresponding intervals in the space of coordinates.

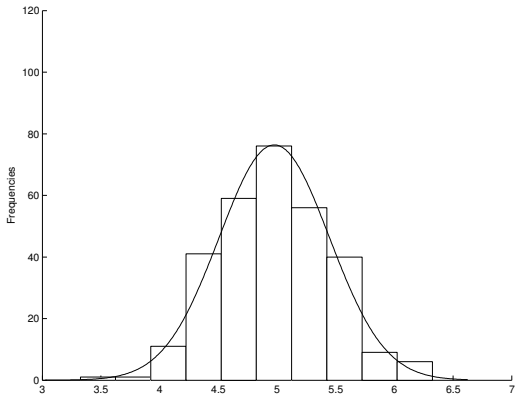


Figure 4: *Simulated sample, $n = 300$. Histogram of the log-transformed sample with the fitted normal density. The bins of the histogram are the log-images of the bins shown in Figure 3(b).*

The normal on $\mathbb{R}_+$ density model and its properties have been applied in a spatial context, and the results have been compared with those obtained with the classical lognormal kriging approach (Tolosana-Delgado and Pawlowsky-Glahn, 2007). Using this approach, problems of non-optimality, robustness and preservation of distribution disappear.

4. The simplex

Compositional data are parts of some whole which carry only relative information. Typical examples are parts per unit, percentages, ppm, or moles per liter. When constrained to sum to a constant, their sample space is the D-part simplex, $\mathcal{S}^D = \{\mathbf{x} = (x_1, x_2, \ldots, x_D)^\mathsf{T} : x_1 > 0, x_2 > 0, \ldots, x_D > 0; \sum_{i=1}^D x_i = \kappa\}$, where $^\mathsf{T}$ stands for transpose and κ is a constant, set to 1 for simplicity in (Aitchison, 1982).

The simplex $\mathcal{S}^D$ has a $(D-1)$-dimensional Euclidean vector space structure (Billheimer et al., 2001; Pawlowsly-Glahn and Egozcue, 2001) with the following operations. Let $\mathcal{C}(\cdot)$ denote the closure operation which normalises any vector $\mathbf{x}$ to a constant sum (Aitchison, 1982), $\mathbf{x}, \mathbf{x}^* \in \mathcal{S}^D$, and $\alpha \in \mathbb{R}$. The internal operation, called *perturbation*, is defined as $\mathbf{x} \oplus \mathbf{x}^* = \mathcal{C}(x_1 x_1^*, x_2 x_2^*, \ldots, x_D x_D^*)^\mathsf{T}$, with inverse $\mathbf{x} \ominus \mathbf{x}^* = \mathcal{C}(x_1/x_1^*, x_2/x_2^*, \ldots, x_D/x_D^*)^\mathsf{T}$. The external operation, called *powering*, is defined as $\alpha \odot \mathbf{x} = \mathcal{C}(x_1^\alpha, x_2^\alpha, \ldots, x_D^\alpha)^\mathsf{T}$, and the inner product as

$$\langle \mathbf{x}, \mathbf{x}^* \rangle_a = \frac{1}{D} \sum_{i<j} \ln \frac{x_i}{x_j} \ln \frac{x_i^*}{x_j^*} \,. \tag{10}$$

The associated squared distance

$$\mathrm{d}_a^2(\mathbf{x}, \mathbf{x}^*) = \frac{1}{D} \sum_{i<j} \left(\ln \frac{x_i}{x_j} - \ln \frac{x_i^*}{x_j^*} \right)^2 ,$$

is relative and satisfies standard properties of a distance (Martín-Fernández et al., 1998), i.e. $\mathrm{d}_a(\mathbf{x}, \mathbf{x}^*) = \mathrm{d}_a(\mathbf{a} \oplus \mathbf{x}, \mathbf{a} \oplus \mathbf{x}^*)$ and $\mathrm{d}_a(\alpha \odot \mathbf{x}, \alpha \odot \mathbf{x}^*) = \mid \alpha \mid \mathrm{d}_a(\mathbf{x}, \mathbf{x}^*)$. The corresponding geometry is known as *Aitchison geometry*, and therefore the subindex a is used.

The inner product (10) and its associated norm, $\|\mathbf{x}\|_a = \sqrt{\langle \mathbf{x}, \mathbf{x} \rangle_a}$, ensure the existence of orthonormal basis $\{\mathbf{e}_1, \mathbf{e}_2, \ldots, \mathbf{e}_{D-1}\}$ which, for a given basis, lead to a unique expression of a composition $\mathbf{x}$ as a linear combination,

$$\mathbf{x} = (\langle \mathbf{x}, \mathbf{e}_1 \rangle_a \odot \mathbf{e}_1) \oplus (\langle \mathbf{x}, \mathbf{e}_2 \rangle_a \odot \mathbf{e}_2) \oplus \ldots \oplus (\langle \mathbf{x}, \mathbf{e}_{D-1} \rangle_a \odot \mathbf{e}_{D-1}).$$

In inner product spaces, an orthonormal basis is not uniquely determined. It is not straightforward to determine which one is the most appropriate to solve a specific problem, but a promising strategy, based on binary partitions, has been developed in

(Egozcue and Pawlowsky-Glahn, 2005). Here, whenever a specific basis is needed, the basis given in (Egozcue et al., 2003) is used. In this basis, the coordinates of $\mathbf{x} \in \mathcal{S}^D$ are

$$y_i = \frac{1}{\sqrt{i(i+1)}} \ln\left(\frac{x_1 x_2 \cdots x_i}{x_{i+1}^i}\right), \qquad i = 1,2,\ldots,D-1\,. \tag{11}$$

The coordinates in this particular basis are denoted $\mathrm{ilr}(\mathbf{x})$ to emphasise the fact that this coordinate transformation is an isometric mapping from $\mathcal{S}^D$ to $\mathbb{R}^{D-1}$ and that the coordinates are log-ratios (Egozcue et al., 2003). The important point is that, once an orthonormal basis has been chosen, all standard statistical methods can be applied to the coordinates and transferred to the simplex preserving their properties. This is what we call the *principle of working on coordinates* (Mateu-Figueras et al., 2011). As stated in Section 2., the Lebesgue measure in the space of coordinates induces a measure in $\mathcal{S}^D$, denoted here as λ_a and called Aitchison measure on $\mathcal{S}^D$. This measure is absolutely continuous with respect to the Lebesgue measure on $\mathbb{R}_{D-1}$, and the relationship between them is $|d\lambda_a/d\lambda_{D-1}| = (\sqrt{D}\; x_1 x_2 \cdots x_D)^{-1}$ (Pawlowsky-Glahn, 2003). Following the notation in Section 2., all these definitions can be obtained by setting $E = \mathcal{S}^D$ and $d = D-1$.

For later use, the concept of subcomposition is required. For $1 < C < D$, a C-part subcomposition, $\mathbf{x}_S$, from a D-part composition, $\mathbf{x}$, can be obtained as $\mathbf{x}_S = \mathcal{C}(\mathbf{S}\mathbf{x})$, where $\mathbf{S}$ is a $C \times D$ selection matrix with C elements equal to 1 (one in each row and at most one in each column) and the remaining elements equal to 0 (Aitchison, 1986). A subcomposition can be regarded as a composition in a simplex with fewer parts, and thus as an element of a space of lower dimension.

4.1. Some basic statistical concepts in the simplex

A random composition $\mathbf{X}$ is a random vector with $\mathcal{S}^D$ as sample space. In the literature laws of probability over $\mathcal{S}^D$ using the Lebesgue measure can be found. Consequently, the probabilities or any moment are computed using the classical definition. But some usual elements appear to be of little use when working with real situations. One typical example is the expected value, in the sense that frequently it does not lie within the bulk of the sample. As an alternative, the geometric interpretation of the expected value has been used to define the centre, $\mathrm{cen}[\mathbf{X}]$, of a random composition. It is defined as the composition that minimises the expression $\mathrm{E}[\mathrm{d}_a^2(\mathbf{X}, \mathrm{cen}[\mathbf{X}])]$ (Aitchison, 1997; Pawlowsly-Glahn and Egozcue, 2001). The result is $\mathrm{cen}[\mathbf{X}] = \mathcal{C}(\exp(\mathrm{E}[\ln \mathbf{X}]))$, which can be rewritten as (Egozcue et al., 2003) $\mathrm{cen}[\mathbf{X}] = \mathrm{ilr}^{-1}(\mathrm{E}[\mathrm{ilr}(\mathbf{X})])$, or, in general terms, as

$$\mathrm{cen}[\mathbf{X}] = h^{-1}\left(\mathrm{E}\left[h(\mathbf{X})\right]\right).$$

The centre of a random composition is equal to the expectation in $\mathcal{S}^D$ defined in Equation (5) in Section 2. This is an important result because if a law of probability on $\mathcal{S}^D$ is defined using the classical approach, this equality does not hold.

As already mentioned, traditionally the simplex has been considered as a subset of real space and, consequently, the laws of probability have been defined using the standard approach. This is the case for families of distributions like the Dirichlet (Monti et al., 2011), the additive logistic normal (Aitchison, 1982), the additive logistic skew-normal (Mateu-Figueras et al., 2005), or those defined using the Box-Cox family of transformations (Barceló-Vidal, 1996). Except for the Dirichlet, these laws of probability are defined using transformations from the simplex to real space. Two of these transformations will appear later herein, the additive log-ratio (alr) and the centred log-ratio (clr),

$$\operatorname{alr}(\mathbf{x}) = \left(\ln\left(\frac{x_1}{x_D}\right), \ldots, \ln\left(\frac{x_{D-1}}{x_D}\right)\right)^{\mathsf{T}}, \tag{12}$$

$$\operatorname{clr}(\mathbf{x}) = \left(\ln\left(\frac{x_1}{g(\mathbf{x})}\right), \ldots, \ln\left(\frac{x_D}{g(\mathbf{x})}\right)\right)^{\mathsf{T}}, \tag{13}$$

where $g(\mathbf{x})$ is the geometric mean of the components of $\mathbf{x}$. The relationship between the alr and the clr transformations is provided by Aitchison (1986, p.92). The relationships between the alr, clr and ilr transformations are provided by Egozcue et al. (2003).

4.2. The normal distribution on $\mathcal{S}^D$

Aitchison (1986) introduced the additive logistic normal (aln) distribution. A random variable on the simplex is aln distributed if the alr transformed random composition (12) has a multivariate normal distribution. The alr transformation is a representation of a composition using coordinates with respect to an oblique basis of the simplex (Egozcue et al., 2011). An equivalent definition can be formulated using orthonormal coordinates. Additionally, the Aitchison measure on the simplex is used to obtain the corresponding density function. Although the following definition is formally different from that of the aln, it corresponds to the same probability law with a different parametrisation.

Consider an orthonormal basis on $\mathcal{S}^D$ and let $h(\cdot)$ be the corresponding orthonormal coordinates.

Definition 4.1 *Let $(\Omega, \mathcal{F}, p)$ be a probability space. A random composition $\mathbf{X} : \Omega \longrightarrow \mathcal{S}^D$ is said to have a normal on $\mathcal{S}^D$ distribution, with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, if its moment generating function is*

$$M(\mathbf{t}) = \mathrm{E}[\exp(\mathbf{t}^{\mathsf{T}} h(\mathbf{X}))] = \exp\left(\boldsymbol{\mu}^{\mathsf{T}}\mathbf{t} + \frac{1}{2}\mathbf{t}^{\mathsf{T}}\boldsymbol{\Sigma}\mathbf{t}\right),$$

where $\mathbf{t}$ *is a* $D-1$ *real vector. The* $D-1$ *vector* $\boldsymbol{\mu}$ *is the mean expressed in coordinates and the* $(D-1)\times(D-1)$ *matrix* $\boldsymbol{\Sigma}$ *is the covariance matrix of the coordinates.*

The parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ depend on the specific orthonormal basis selected. If the covariance matrix $\boldsymbol{\Sigma}$ is non singular, it can be inverted and the regular normal distribution can be defined as follows:

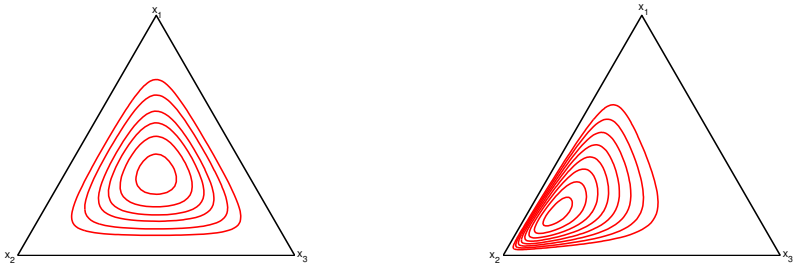
Definition 4.2 *Let* $(\Omega,\mathcal{F},p)$ *be a probability space. A random composition* $\mathbf{X}:\Omega\longrightarrow\mathcal{S}^D$ *is said to have a regular normal on* $\mathcal{S}^D$ *distribution, with parameters* $\boldsymbol{\mu}$ *and* $\boldsymbol{\Sigma}$*, if its density function is*

$$f_{\mathbf{X}}^{\mathcal{S}}(\mathbf{x})=(2\pi)^{-(D-1)/2}|\boldsymbol{\Sigma}|^{-1/2}\exp\left(-\frac{1}{2}\left(h(\mathbf{x})-\boldsymbol{\mu}\right)^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}\left(h(\mathbf{x})-\mu\right)\right). \qquad (14)$$

The notation $\mathbf{X}\sim\mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ is used. The subscript $\mathcal{S}$ indicates that it is a density on the simplex, i.e. a Radon-Nykodym derivative with respect to the Aitchison measure on $\mathcal{S}^D$; the superscript D indicates the number of parts of the composition. Figure 5 shows the isodensity curves of two normal densities on $\mathcal{S}^3$ taking the particular basis given by Egozcue et al. (2003) and using a ternary diagram as a convenient and simple way for representing 3-part compositions (see Aitchison, 1986, p.6). The isodensity curves are not equidistant, the levels are only chosen in order to clearly show the shape of the density function. To understand Figure 5, it should be remarked that the areas on the ternary diagram are computed according λ_a, which significantly differs from the usual Lebesgue area intuitively assigned to the triangle interior. The differences of assigned areas are specially dramatic near the edges of the triangle.

The density (14) is the usual normal density applied to coordinates $h(\mathbf{x})$ as implied by (4). It is a density in $\mathcal{S}^D$ with respect to the λ_a measure. The same strategy is used by Mateu-Figueras and Pawlowsky-Glahn (2007) to define the skew-normal in $\mathcal{S}^D$ law.

The main properties of this model follow. A complete proof of each property can be found in the appendix. The proofs are straightforward for a reader familiar with compositional data analysis.



(a) *(b)*

Figure 5: *Isodensity plots of two* $\mathcal{N}_{\mathcal{S}}^{3}(\mu,\boldsymbol{\Sigma})$ *with (a)* $\mu=(0,0)$*, (b)* $\boldsymbol{\mu}=(-1,1)$ *and* $\boldsymbol{\Sigma}=Id$*.*

Property 4.1 *Let* $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu},\boldsymbol{\Sigma})$, $\boldsymbol{a} \in \mathcal{S}^D$ *and* $b \in \mathbb{R}$. *Then, the D-part random composition* $\mathbf{X}^* = \mathbf{a} \oplus (b \odot \mathbf{x})$ *has a* $\mathcal{N}_{\mathcal{S}}^{D}(h(\mathbf{a}) + b\boldsymbol{\mu}, b^2\boldsymbol{\Sigma})$ *distribution.*

Property 4.2 *Let* $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *and* $\mathbf{a} \in \mathcal{S}^D$. *Then* $f^{\mathcal{S}}_{\mathbf{a}\oplus\mathbf{X}}(\mathbf{a} \oplus \mathbf{x}) = f^{\mathcal{S}}_{\mathbf{X}}(\mathbf{x})$, *where* $f^{\mathcal{S}}_{\mathbf{X}}$ *and* $f^{\mathcal{S}}_{\mathbf{a}\oplus\mathbf{X}}$ *represent the density functions of the random compositions* $\mathbf{X}$ *and* $\mathbf{a} \oplus \mathbf{X}$, *respectively.*

Property 4.3 *Let* $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *and* $\mathbf{X}_P = \mathbf{P}\,\mathbf{X}$, *the random composition* $\mathbf{X}$ *with the parts reordered by a permutation matrix* $\mathbf{P}$. *Then* $\mathbf{X}_P \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu}_P,\boldsymbol{\Sigma}_P)$ *with* $\boldsymbol{\mu}_P = \mathbf{U}^{\mathsf{T}}\mathbf{P}\mathbf{U}\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_P = (\mathbf{U}^{\mathsf{T}}\mathbf{P}\mathbf{U})\boldsymbol{\Sigma}(\mathbf{U}^{\mathsf{T}}\mathbf{P}^{\mathsf{T}}\mathbf{U})$, *where* $\mathbf{U}$ *is a* $D \times (D-1)$ *matrix with the* clr *transformation of an orthonormal basis of* $\mathcal{S}^D$ *as columns.*

Property 4.4 *Let* $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu},\boldsymbol{\Sigma})$ *and* $\mathbf{X}_S = \mathcal{C}(\mathbf{S}\mathbf{X})$, *the C-part random subcomposition obtained from the* $C \times D$ *selection matrix* $\mathbf{S}$. *Then* $\mathbf{X}_S \sim \mathcal{N}_{\mathcal{S}}^{C}(\boldsymbol{\mu}_S,\boldsymbol{\Sigma}_S)$, *with* $\mu_S = \mathbf{U}^{*\mathsf{T}}\mathbf{S}\mathbf{U}\boldsymbol{\mu}$, $\boldsymbol{\Sigma}_S = (\mathbf{U}^{*\mathsf{T}}\mathbf{S}\mathbf{U})\boldsymbol{\Sigma}(\mathbf{U}^{\mathsf{T}}\mathbf{S}^{\mathsf{T}}\mathbf{U}^{*})$, *where* $\mathbf{U}$ *is a* $D \times (D-1)$ *matrix with the* clr *transformation of an orthonormal basis of* $\mathcal{S}^D$ *as columns and* $\mathbf{U}^*$ *is a* $C \times (C-1)$ *matrix with the* clr *transformation of an orthonormal basis of* $\mathcal{S}^C$ *as columns.*

Property 4.5 *Let* $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. *Then, the expected value in* $\mathcal{S}^D$ *is*

$$\operatorname{cen}[\mathbf{X}] = \mathrm{E}^a[\mathbf{X}] = h^{-1}(\boldsymbol{\mu}) \,,$$

independently of the orthonormal basis of $\mathcal{S}^D$ *for which the coordinate mapping* h *is defined.*

Property 4.6 *Let* $\mathbf{X} \sim \mathcal{N}_{\mathcal{S}}^{D}(\boldsymbol{\mu},\boldsymbol{\Sigma})$. *The metric variance of* $\mathbf{X}$ *is* $\operatorname{Var}^a[\mathbf{X}] = \operatorname{trace}(\boldsymbol{\Sigma})$.

Property 4.1 states that the normal on $\mathcal{S}^D$ law is closed under perturbation and powering. Property 4.2 states that it is equivariant under perturbation. This is important, because when working with compositional data the centring operation (Martín-Fernández et al., 1999), a perturbation using the inverse of the centre of the data set, is often applied in practice to better visualise and interpret the pattern of variability (von Eynatten et al., 2002). Properties 4.3 and 4.4 show that the normal on $\mathcal{S}^D$ family is closed under permutation and subcompositions.

Given a compositional data set the estimates of parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ can be computed applying the maximum likelihood procedure to the coordinates. The estimated values $\hat{\boldsymbol{\mu}}$ and $\widehat{\boldsymbol{\Sigma}}$ allow to compute the estimates of the centre (expected value in $\mathcal{S}^D$) and metric variance of the random composition $\mathbf{X}$, as

$$\begin{aligned} \widehat{\mathrm{E}^a[\mathbf{X}]} &= (\hat{\mu}_1 \odot \mathbf{e}_1) \oplus \cdots \oplus (\hat{\mu}_{D-1} \odot \mathbf{e}_{D-1}) = h^{-1}(\hat{\boldsymbol{\mu}}) \,, \\ \widehat{\operatorname{Var}^a[\mathbf{X}]} &= \operatorname{trace}\left(\widehat{\boldsymbol{\Sigma}}\right) . \end{aligned}$$

To validate the distributional assumption of normality on $\mathcal{S}^D$, some goodness-of-fit tests of the multivariate normal distribution have to be applied to the coordinates of the sample data set. There is a large battery of possible tests but, as suggested by Aitchison (1986), it is reasonable to start testing the normality of each marginal using empirical distribution function tests. Unfortunately, the univariate normality of each component is a necessary but not sufficient condition for the normality of the whole vector. Also, these univariate tests depend on the orthonormal basis chosen. This difficulty does not depend on the proposed methodology, as the same problem appears when working with laws of probability defined using transformations and the Lebesgue measure in $\mathcal{S}^D$ (Aitchison et al., 2003). The multivariate normal model can also be validated considering the Mahalanobis distance $(h(\mathbf{X})-\hat{\boldsymbol{\mu}})^{\mathsf{T}}\widehat{\boldsymbol{\Sigma}}^{-1}(h(\mathbf{X})-\hat{\boldsymbol{\mu}})$, which is sampled from a χ^2_{D-1}-distribution if the fitted model is appropriate. In this case, the dependence on the chosen orthonormal basis disappears (Stevens, 1986). The use of empirical distribution function tests is also suggested in (Aitchison, 1986).

As mentioned at the beginning of this section, the parametrisation used depends on the selected orthonormal basis. In fact, the vector $\boldsymbol{\mu}$ and the matrix $\boldsymbol{\Sigma}$ are the mean and the covariance matrix of the coordinates random vector $h(\mathbf{X})$. Nevertheless, the distribution can also be defined using object parameters (Tolosana-Delgado, 2005; Eaton, 1983). The idea under the object parametrisation is to define the model independently of the coordinates used for representation. The mean vector of the coordinates $\boldsymbol{\mu}$ is the coordinate representation of a composition in $\mathcal{S}^D$, $\boldsymbol{\mu}_{\mathcal{S}} = h^{-1}(\boldsymbol{\mu})$, that does not depend on the selected basis. The covariance matrix $\boldsymbol{\Sigma}$ can be interpreted as the representation of a symmetric positive semidefinite endomorphism $\Sigma_{\mathcal{S}}$ on $\mathcal{S}^D$. For each choice of basis in $\mathcal{S}^D$ the endomorphism has a different matrix representation $\boldsymbol{\Sigma}$, but the endomorphism itself remains the same. Then, as proposed by Tolosana-Delgado (2005), the normal on $\mathcal{S}^D$ distribution can be defined as

Definition 4.3 *A random composition* $\mathbf{X}$ *is said to follow a normal on* $\mathcal{S}^D$ *distribution with a given mean vector* $\boldsymbol{\mu}_{\mathcal{S}}$ *and a positive semidefinite symmetric endomorphism* $\Sigma_{\mathcal{S}}$ *on* $\mathcal{S}^D$*, if for any testing vector* $\mathbf{x}$*, the projection* $\langle \mathbf{x}, \mathbf{X}\rangle_a$ *follows a univariate normal distribution on* $\mathbb{R}$ *with expectation* $\langle \mathbf{x}, \boldsymbol{\mu}_{\mathcal{S}}\rangle_a$ *and variance* $\langle \mathbf{x}, \Sigma_{\mathcal{S}}\mathbf{x}\rangle_a$*.*

If $\Sigma_{\mathcal{S}}$ is positive definite, the density with respect to λ_a is

$$f^{\mathcal{S}}_{\mathbf{X}}(\mathbf{x}) = (2\pi)^{-(D-1)/2}|\Sigma_{\mathcal{S}}|^{-1/2}\exp\left(-\frac{1}{2}\langle \mathbf{x}\ominus\boldsymbol{\mu}_{\mathcal{S}}, \Sigma_{\mathcal{S}}^{-1}(\mathbf{x}\ominus\boldsymbol{\mu}_{\mathcal{S}})\rangle_a\right),$$

where $|\Sigma_{\mathcal{S}}|$ is the determinant of the endomorphism $\Sigma_{\mathcal{S}}$.

As noted by Tolosana-Delgado (2005), given a basis, object definitions may be identified with coordinate ones, proving that the coordinate approach gives the same results whichever basis is used.

4.3. The central limit theorem in S^D

The relevant role of the normal distribution for real vectors is due to the central limit theorem and related properties. The normal distribution in the simplex satisfies a central limit theorem in S^D, as stated in (Aitchison, 1986) to characterize the logistic normal distribution. In the present context, the multivariate central limit theorem (Kocherlakota and Kocherlakota, 1982) holds for coordinates. Let $\mathbf{X}_1$, $\mathbf{X}_2$, ..., $\mathbf{X}_n$ be a sequence of random compositions in $\mathcal{S}^D$. Consider their coordinates with respect to an arbitrary orthonormal basis $\mathbf{Y}_i = h(\mathbf{X}_i) \in \mathbb{R}^{D-1}$, $i = 1,2,\ldots,n$. Assume that the coordinate vectors $\mathbf{Y}_i$ are mutually independent and identically distributed with mean $\boldsymbol{\mu}$ and covariance matrix $\boldsymbol{\Sigma}$; then, being $\overline{\mathbf{Y}}_n = n^{-1}\sum_{i=1}^n \mathbf{Y}_i$, the random vector $\sqrt{n}\left(\overline{\mathbf{Y}}_n - \boldsymbol{\mu}\right)$ converges in distribution to the multivariate normal $\mathcal{N}(0,\boldsymbol{\Sigma})$ as $n \to \infty$. These random vectors are coordinates of the random compositions

$$\sqrt{n} \odot \left(\overline{\mathbf{X}}_n \ominus h^{-1}(\mu)\right) \ , \ \overline{\mathbf{X}}_n = \frac{1}{n} \odot \bigoplus_{i=1}^{n} \mathbf{X}_i = \mathcal{C}\left(\exp\left(\frac{1}{n}\sum_{i=1}^{n} \ln \mathbf{X}_i\right)\right) \ . \tag{15}$$

The random compositions (15) converge in distribution to $\mathcal{N}_{\mathcal{S}}^D(0,\boldsymbol{\Sigma})$, and the multivariate central limit theorem holds in the simplex. The reference operation is the perturbation and the corresponding average equals to the closed geometric mean composition, that is, the geometric mean vector followed by the closure operation. This result justifies the name of normal in the simplex for the studied distribution. However, the relevance of a central limit theorem in this context relays on the interpretation of the average $\overline{\mathbf{X}}_n$ or just the perturbation of random variables in the simplex. Perturbation has many intuitive interpretations depending on the applied context. The following example of concentrations may be illustrative. Suppose that $\mathbf{Z}$ contains the concentrations of D pollutants in a sample of water. The sample is filtered using a permeable membrane with transfer function $\mathbf{X}$, i.e. the components are multiplicative factors applied to the concentrations in $\mathbf{Z}$. The perturbation $\mathbf{Z} \oplus \mathbf{X}$ expresses the output concentrations after applying the filter $\mathbf{X}$. As the filtering membrane is replaced by another similar one after each filtering process, we can assume that $\mathbf{X}$ is random. In order to express the random effect of a filtering membrane we perform a sequence of filtering experiments with similar but different transfer functions. The mean transfer function, say $\overline{\mathbf{X}}_n$, will be approximately distributed as a normal in the simplex as predicted by the central limit theorem.

4.4. The normal on S^D vs the additive logistic normal

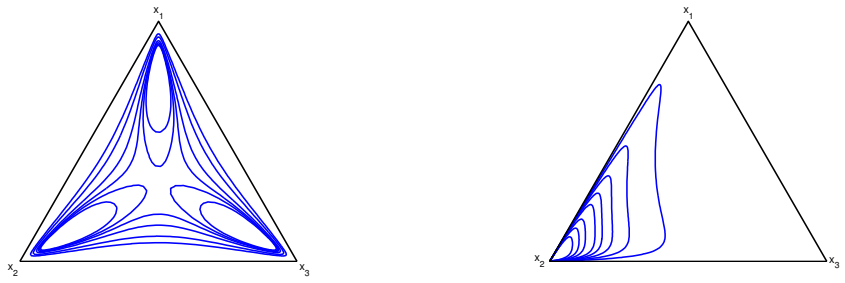
The normal on the simplex is well known as the logistic normal distribution. Nevertheless, the proposed change of representation produces differences in some properties. In this section we study these changes.

The approach used in (Aitchison, 1982) to define the additive logistic normal law on the simplex is standard: transform the random composition from the simplex to real

space, define the density function of the transformed vector, return to the simplex using the change of variable theorem. The result is a density function for the initial random composition with respect to the Lebesgue measure. Therefore, a random composition is said to have an additive logistic normal distribution (aln) when the additive log-ratio (alr) transformed vector – see Eq. (12) – has a normal distribution. Note that this definition does not explicitly state that the change of variable theorem is used. But this is the principal difference between the approach based on working with transformations, and the new approach, based on working with coordinates.

The aln model was initially defined using the additive log-ratio transformation. Using the matrix relationship between log-ratio transformations (Egozcue et al., 2003) the density function in terms of an isometric log-ratio transformation is obtained. Consequently, we can define the logistic normal distribution with parameters $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$, with density function:

$$f_{\mathbf{X}}(\mathbf{x}) = \frac{(2\pi)^{-(D-1)/2} \mid \boldsymbol{\Sigma} \mid^{-1/2}}{\sqrt{D} x_1 x_2 \cdots x_D} \exp\left(-\frac{1}{2} (\mathrm{ilr}(\mathbf{x}) - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathrm{ilr}(\mathbf{x}) - \mu) \right). \qquad (16)$$



(a) *(b)*

Figure 6: *Isodensity plots of two logistic normal densities with (a)* $\boldsymbol{\mu} = (0,0)$*, (b)* $\boldsymbol{\mu} = (-1,1)$ *and* $\boldsymbol{\Sigma} = Id$*.*

To easily compare both approaches we will use the normal density on the simplex taking the basis given by Egozcue et al. (2003) and consequently the ilr vector stated in (11). Nevertheless, any orthonormal basis could be considered, as the vector $\mathrm{ilr}(\mathbf{x})$ can be obtained from $h(\mathbf{x})$ and the corresponding change of basis matrix. The only difference between expressions (14) and (16) is the term $(\sqrt{D} x_1 x_2 \cdots x_D)^{-1}$, the Jacobian of the isometric log-ratio transformation that reflects the change of the measure on $\mathcal{S}^D$. The influence of this term can be observed in the isodensity curves in Figure 6, where areas on the triangle are computed using the ordinary Lebesgue measure. These curves can be compared with the curves in Figure 5, where areas where computed using λ_a. The differences between Figures 5 and 6 are obvious; in particular the tri-modality in Figure 6(a). This behaviour is not exclusive of the logistic normal density, bimodality is also present with Beta or Dirichlet densities when their parameters are close to 0 and the Lebesgue measure is considered. In Figure 6(b) a single mode can be observed;

nevertheless, the position and the shape of the curves are not the same as in Figure 5(b), the corresponding normal on $\mathcal{S}^3$.

Another difference is the moments of any order. The expression of the density function plays a fundamental role when any moment is computed. The density (16) is a classical density, consequently moments are computed using the standard definition. Obviously, the results are not the same as in the normal on $\mathcal{S}^D$ case. For example, the expected value of an aln density, denoted as $\mathrm{E}[\mathbf{X}]$, exists but numerical procedures have to be applied (see Aitchison, 1986, p.116) to find it and the result is not the same as in property 4.5. Using our approach, the centre of a random composition, denoted as $\mathrm{cen}[\mathbf{X}]$ (Aitchison, 1997), is obtained when the expected value $\mathrm{E}^a[\mathbf{X}]$ is computed. Consequently, it is not necessary to define new characteristic parameters. Using the classical definition, e.g. the expected value, a representative location parameter is obtained. Remember that the centre of a random composition was introduced by Aitchison (1997) because he perceived that the usual expected value $\mathrm{E}[\mathbf{X}]$ is not a representative location parameter. This discrepancy appears because Aitchison (1982) adopts perturbation and powering as operations in the sample space, but uses the density function with respect to the Lebesgue measure, thus assuming for the density a measure not compatible with the operations.

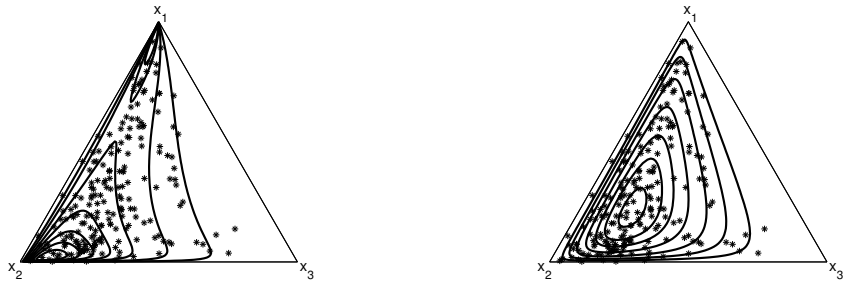
Some coincidences can be found as well. The closure under perturbation, powering, permutation and subcompositions of the logistic normal density model is proven by Aitchison (1986), and stated in Properties 4.1,4.3 and 4.4 for the normal on $\mathcal{S}^D$ density model. Nevertheless, the logistic normal class is not equivariant under perturbation, i.e. $f_{\mathrm{a}\oplus\mathbf{X}}(\mathrm{a}\oplus\mathbf{x}) \neq f_{\mathbf{X}}(\mathbf{x})$.

In summary, the essential differences between both approaches are the shape of the probability density function, in some cases leading to multi-modality for the standard approach; the moments which characterise the density, particularly important in practice for the expected value and the variance; and equivariance under perturbation.

4.5. Example

To illustrate the differences between a density with respect to the Lebesgue measure λ and a density with respect to the measure λ_a in $\mathcal{S}^D$, a GDP data set will be used. The data set used is taken from the National Accounts Statistics database and is available on the United Nations Statistic Division web page http://unstats.un.org/unsd/snaama/dnllist.asp. We use the information corresponding to the year 2009 for 208 countries. The GDP data set is based on the international standard industrial classification (ISIC) of all economic activities. The original data contains the percentages of each economic activity for all countries divided in six categories.

The goal is to compare some characteristics corresponding to the logistic normal and the normal densities on the simplex. In order to provide some useful comprehensive figures a three-part compositional data set is preferred. For this reason the three-part subcomposition (x_1, x_2, x_3) is used, where



(a) *(b)*

Figure 7: *GDP data with isodensity curves of the fitted (a) logistic normal and (b) normal on $\mathcal{S}^3$ densities.*

- $x_1 =$ agriculture, hunting, forestry, fishing (ISIC A-B),
- $x_2 =$ mining, manufacturing, utilities (ISIC C-E),
- $x_3 =$ construction (ISIC F).

Following the suggestions by Aitchison (1986), a battery of 12 tests of goodness-of-fit are used. They are based on the Anderson-Darling, Cramér-von Mises and Watson statistics, applied to the coordinates of the three-part sample data set. In particular, the tests are applied to the marginal distributions, to the bivariate angle distribution and to the radial distribution. Taking a 5% significance level, no significant departure from normality is obtained by any of these tests.

Parameters of the two density models, the normal on $\mathcal{S}^3$ and the logistic normal, are equal. This is a direct consequence of the definition of densities and hence likelihoods. In this case, after taking a suitable ilr transformation (eq. 11 was used), the maximum likelihood estimates for both density models are:

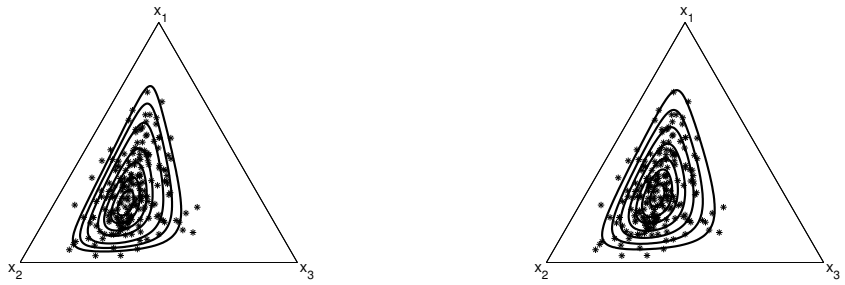
$$\hat{\boldsymbol{\mu}} = (-0.715, 0.521)^{\mathsf{T}}, \qquad \hat{\boldsymbol{\Sigma}} = \begin{pmatrix} 1.303 & 0.452 \\ 0.452 & 0.680 \end{pmatrix}.$$

Figures 7(a) and 7(b) show the sample in a ternary diagram and the isodensity curves of the fitted logistic normal density and the normal in $\mathcal{S}^3$ density. Different features are observed. The logistic normal density shows two modes whereas the normal in $\mathcal{S}^3$ exhibits a single mode. When contours and sample are plotted in the coordinate space (Figure 9(a)) differences disappear, as the probability density in ilr-coordinates, a bivariate normal in $\mathbb{R}^2$, is equal for the two density models.

After plotting the contours of the density with respect to the Lebesgue measure (Figure 7(a)) showing two modes, one might think about the existence of two subpopulations that could explain the bimodality of the logistic normal density. However, using the available information on the data set concerning geography or development of countries, no coherent reasons were found for the observed bimodality shown in Figure

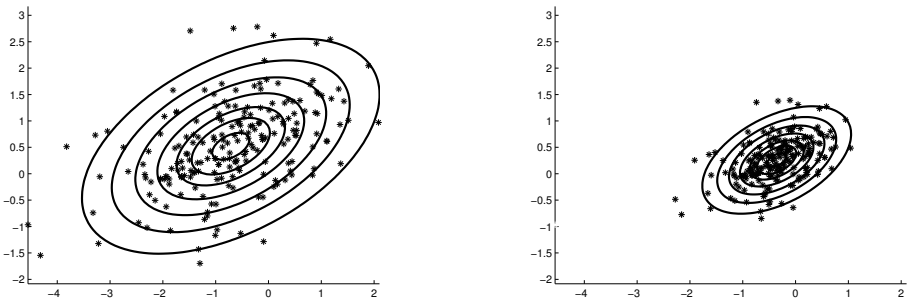
7(a). The bimodality in this case is only due to the measure of reference chosen in the simplex.

For illustration purposes, changes of density contours under powering are shown in Figure 8 for both density models. Let $\mathbf{X}$ denote the three-part random variable of the example with the estimated parameters $\hat{\boldsymbol{\mu}}$ and $\hat{\boldsymbol{\Sigma}}$ shown above, and consider $\mathbf{X}_\alpha = \alpha \odot \mathbf{X}$ for $\alpha = 1/2$. In Figures 8(a) and 8(b) the isodensity contours with respect to the Lebesgue measure and the Aitchison measure in the ternary diagram are represented. As can be observed, in the logistic normal case, the bimodality disappears. In other words, the power transformation, which should only move the centre of the density and modify the variability, can eliminate or in other cases generate arbitrary modes if the Lebesgue measure is considered (Mateu-Figueras and Pawlowsky-Glahn, 2008). This undesirable behaviour of modes and isodensity contours prevents the use of the logistic normal density on the simplex and all statistics depending on it, e.g. expectation and covariance with respect to the Lebesgue measure on the simplex, predictive regions, etc.



(a) *(b)*

Figure 8: *Power transformed GDP data with isodensity curves of the fitted (a) logistic normal and (b) normal on $\mathcal{S}^3$ densities.*



(a) *(b)*

Figure 9: ilr *coordinates of the (a) GDP data set and (b) the power transformed data set with the corresponding fitted normal densities.*

5. Conclusions

A particular Euclidean vector space structure of the positive real line and of the simplex, together with the associated measure, allow us to define parametric models with desirable properties. Normal density models on $\mathbb{R}_+$ and on $\mathcal{S}^D$ have been defined through their densities over the coordinates with respect to an orthonormal basis and their main algebraic properties have been studied. From a probabilistic point of view, those laws of probability are identical to the lognormal and to the additive logistic normal distribution defined using the Lebesgue measure and the standard methodology based on transformations. Nevertheless, some differences are obtained in the moments and in the shape of the density function. In particular, the expected value with respect to the new measure differs from what would be obtained with the Lebesgue measure for the lognormal and additive logistic normal distributions, but leads to the parameters that are used for these models. It thus yields directly a suitable characterization of these models. In the normal in $\mathbb{R}_+$ case, a consistent estimator and confidence intervals for the mean are easily obtained directly from the log-transformed data, while in the lognormal case, i.e., keeping the Lebesgue measure and therefore aiming at the corresponding, common expected value, a bias correction is necessary. In the normal in $\mathcal{S}^D$ case we show important differences in the shape of the density. The normal in $\mathcal{S}^D$ always appears unimodal, whereas bimodal and trimodal densities could be obtained using the standard approach.

Acknowledgements

The authors thank two anonymous referees for their suggestions which greatly improved the article. This work has been supported by the Spanish Ministry of Education and Science under project MTM2009-13272 and by the Agència de Gestió d'Ajuts Universitaris i de Recerca of the Generalitat de Catalunya under project Ref: 2009SGR424.

Appendix

This appendix contains the proofs of properties contained in Section 3.1 and Section 4.2. They use the expected value, the covariance matrix, the linear transformation property of the multivariate normal distribution and some matrix relationships among vectors of coordinates and among log-ratio transformations.

Proof of property 3.1. The coordinates of the random variable X^* are obtained from the coordinates of the variable X as $\ln(X^*) = \ln(a) + b\ln(X)$. The density function of $\ln(X)$ is the classical normal density on the real line; thus, the linear transformation property can be used to obtain the density function of the $\ln(X^*)$ random variable. Therefore, $X^* \sim \mathcal{N}_+(\ln a + b\mu, b^2\sigma^2)$.

Proof of property 3.2. From property 3.1 we know that $a \oplus X = a \cdot X \sim \mathcal{N}_+(\ln a + \mu, \sigma^2)$. From (8) we get

$$f^+_{a\oplus X}(a \oplus x) = \frac{1}{\sqrt{2\pi}\sigma} \exp\left(-\frac{1}{2}\frac{(\ln(ax) - (\ln a + \mu))^2}{\sigma^2}\right) = f^+_X(x).$$

Proof of property 3.3. From (6) we know that $\mathrm{E}^E[X] = \exp(\mathrm{E}[\ln X])$ because the orthonormal coordinates on the positive real line are obtained with the logarithmic transformation. Given some coordinate, the exponential function provides the element on $\mathbb{R}_+$. The density function of $\ln X$ is the normal distribution, as stated in Definition 3.1. Thus, the expected value is the μ parameter and, consequently, $\mathrm{E}^E[X] = \exp(\mu)$. The same result is obtained for the median and the mode, as the normal distribution is symmetric around its expected value μ.

Proof of property 3.4. From (7) we know that the variance can be understood as the expected value of the squared distance around its expected value, i.e. $\mathrm{Var}^+[X] = \mathrm{E}[\mathrm{d}^2_+(X, \mathrm{E}^+[X])]$. Working on coordinates and using the density function of $\ln X$ we obtain $\mathrm{Var}^+[X] = \mathrm{E}[\mathrm{d}^2(\ln X, \mathrm{E}[\ln X])] = \mathrm{Var}[\ln X] = \sigma^2$.

Proof of property 4.1. The orthonormal coordinates of the random composition $\mathbf{X}^*$ are obtained from the orthonormal coordinates of the composition $\mathbf{X}$ via $h(\mathbf{X}^*) = h(\mathbf{a}) + bh(\mathbf{X})$. The density function of $h(\mathbf{X})$ is the classical normal density in real space; thus, the linear transformation property can be used to obtain the density function of $h(\mathbf{X}^*)$. Therefore, $\mathbf{X}^* \sim \mathcal{N}^D_S(h(\mathbf{a}) + b\mu, b^2\mathbf{\Sigma})$.

Proof of property 4.2. Using property 4.1, $\mathbf{a} \oplus \mathbf{X} \sim \mathcal{N}^D_S(h(\mathbf{a}) + \boldsymbol{\mu}, \mathbf{\Sigma})$. We know that $h(\mathbf{a} \oplus \mathbf{x}) = h(\mathbf{a}) + h(\mathbf{x})$, therefore,

$$\begin{aligned} f_{\mathbf{a}\oplus\mathbf{X}}(\mathbf{a} \oplus \mathbf{x}) &= (2\pi)^{-(D-1)/2} \mid \mathbf{\Sigma} \mid^{-1/2} \\ &\quad \times \exp\left[-\frac{1}{2}\left(h(\mathbf{a} \oplus \mathbf{x}) - (h(\mathbf{a}) + \boldsymbol{\mu})\right)^{\mathsf{T}} \mathbf{\Sigma}^{-1} \left(h(\mathbf{a} \oplus \mathbf{x}) - (h(\mathbf{a}) + \boldsymbol{\mu})\right)\right] \\ &= (2\pi)^{-(D-1)/2} \mid \mathbf{\Sigma} \mid^{-1/2} \exp\left[-\frac{1}{2}\left(h(\mathbf{x}) - \boldsymbol{\mu}\right)^{\mathsf{T}} \mathbf{\Sigma}^{-1} \left(h(\mathbf{x}) - \boldsymbol{\mu}\right)\right] \\ &= f_{\mathbf{X}}(\mathbf{x}). \end{aligned}$$

Proof of property 4.3. For a centered log-ratio transformed vector it is straightforward to see that $\mathrm{clr}(\mathbf{X}_P) = \mathbf{P}\mathrm{clr}(\mathbf{X})$ (Aitchison, 1986, p. 94). Using the matrix relationship between the centered and the isometric log-ratio vectors (Egozcue et al., 2003) we conclude that $h(\mathbf{X}_P) = (\mathbf{U}^\mathsf{T}\mathbf{P}\mathbf{U})h(\mathbf{X})$. Given the density of $h(\mathbf{X})$, and applying the linear transformation property of the normal distribution in real space, a $\mathcal{N}_S^D(\boldsymbol{\mu}_P, \boldsymbol{\Sigma}_P)$ distribution is obtained for the random composition $\mathbf{X}_P$.

Proof of property 4.4. (Aitchison, 1986, p. 119) gives the matrix relationship between $\mathrm{alr}(\mathbf{X}_S)$ and $\mathrm{alr}(\mathbf{X})$. Using the matrix relationships between the additive, centered and isometric log-ratio vectors (Egozcue et al., 2003), we conclude that $h(\mathbf{X}_S) = (\mathbf{U}^{*\prime}\,\mathbf{S}\mathbf{U})h(\mathbf{X})$. Given the density of $h(\mathbf{X})$, and applying the linear transformation property of the normal distribution in real space, the density of $h(\mathbf{X}_S)$ is obtained as that of the $\mathcal{N}_S^C(\boldsymbol{\mu}_S, \boldsymbol{\Sigma}_S)$ distribution.

Proof of property 4.5. From (6) we know that $\mathrm{E}^a[\mathbf{X}] = h^{-1}(\mathrm{E}[h(\mathbf{X})])$, and from (14) we know that the density function of $h(\mathbf{X})$ is the multivariate normal distribution; thus $\mathrm{E}[h(\mathbf{X})] = \boldsymbol{\mu}$. Finally, the composition $\mathrm{E}_a[\mathbf{X}]$ is obtained applying h^{-1}.

Proof of property 4.6. From (7) we know that the variance can be understood as the expected value of the squared distance around its expected value, i.e. $\mathrm{Var}^a[\mathbf{X}] = \mathrm{E}[d_a^2(\mathbf{X}, \mathrm{E}_a[\mathbf{X}])]$. Working on coordinates and using the density function of $h(\mathbf{X})$ we obtain $\mathrm{Var}^a[\mathbf{X}] = \mathrm{trace}(\boldsymbol{\Sigma})$.

References

Ahrens, L. (1954). The lognormal distribution of the elements. *Geochimica et Cosmochimica Acta*, 5, 49–73.

Aitchison, J. (1982). The statistical analysis of compositional data (with discussion). *Journal of the Royal Statistical Society: Series B*, 44(2), 139–177.

Aitchison, J. (1986). *The Statistical Analysis of Compositional Data*. Monographs on Statistics and Applied Probability. Chapman & Hall Ltd., London, 1986. (Reprinted in 2003 with additional material by The Blackburn Press).

Aitchison, J. (1997). The one-hour course in compositional data analysis or compositional data analysis is simple. *Proceedings of IAMG'97, the third annual conference of the International Association for Mathematical Geology*, V. Pawlowsky-Glahn ed.,International Center for Numerical Methods in Engineering (CIMNE), Barcelona, vol. 1, 3–35.

Aitchison, J. and Brown, J. A. C. (1957). *The Lognormal Distribution*. Cambridge University Press, Cambridge.

Aitchison, J., Mateu-Figueras, G. and Ng, K. (2003). Characterization of distributional forms for compositional data and associated distributional tests. *Mathematical Geology*, 35(6), 667–680.

Barceló-Vidal, C. (1996). *Mixturas de Datos Composicionales*. Ph.D. Diss., Universitat Politècnica de Catalunya.

Billheimer, D., Guttorp, P. and Fagan, W. (2001). Statistical interpretation of species composition. *Journal of the American Statistical Association*, 96(456), 1205–1214.

Clark, I. and Harper, W. V. (2000). *Practical Geostatistics 2000*. Ecosse North America Llc., Columbus Ohio.

Crow, E. L. and Shimizu, K. (1988). *Lognormal Distributions. Theory and Applications*. Marcel Dekker, Inc. New York.

Davis, J. C. (1986). *Statistics and Data Analysis in Geology. 2nd ed*. John Wiley & Sons. New York.

Eaton, M. L. (1983). *Multivariate Statistics. A Vector Space Approach*. John Wiley & Sons.

Egozcue, J. J., Barceló-Vidal, C., Martín-Fernández, J. A., Jarauta-Bragulat, E., Díaz-Barrero, J. L. and Mateu-Figueras, G. (2011). Elements of Simplicial Linear Algebra and Geometry. *Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn & A. Buccianti eds., John Wiley & Sons, Chichester, 141–157.

Egozcue, J. J. and Pawlowsky-Glahn, V. (2005). Groups of parts and their balances in compositional data analysis. *Mathematical Geology*, 37(7), 795–828.

Egozcue, J. J. Pawlowsky-Glahn, V., Mateu-Figueras, G. and Barceló-Vidal, C. (2003). Isometric logratio transformations for compositional data analysis. *Mathematical Geology*, 35(3), 279–300.

Galton, F. (1879). The geometric mean, in vital and social statistics. *Proceedings of the Royal Society of London*, 29, 365–366.

Herdan, G. (1960). *Small Particle Statistics*. Butterwoths, London.

Kocherlakota, S. and Kocherlakota, K. (1982). *Multinormal Distribution. Encyclopedia of statistical sciences*. S. Kotz and N.L. Johnson eds., John Wiley & Sons, New York, vol. 5, 668–677.

Krige, D. G. (1981). *Lognormal-de Wijsian Geostatistics for Ore Evaluation*. South African Inmstitute of Mining and Metallurgy, Johannesburg.

Martín-Fernández, J. A., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1998). A critical approach to non-parametric classification of compositional data. *Advances in Data Science and Classification, Proceedings of the 6th Conference of the International Federation of Classification Societies, IFCS'98*, A. Rizzi, M. Vichi & H. H. Bock eds., Springer-Verlag, Berlin, 49–56.

Martín-Fernández, J. A., Bren, M., Barceló-Vidal, C. and Pawlowsky-Glahn, V. (1999). A measure of difference for compositional data based on measures of divergence. *Proceedings of IAMG'99, the fifth annual conference of the International Association for Mathematical Geology*, S. J. Lippard, A. Næss & R. Sinding-Larsen eds., Tapir, Trondheim, 211–216.

Mateu-Figueras, G. (2003). *Models de distribució sobre el símplex*. Ph.D. Diss., Universitat Politècnica de Catalunya.

Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2007). The skew-normal distribution on the simplex. *Communications in Statistics-Theory and Methods, Special Issue Skew-elliptical Distributions and their Application*, 36(9), 1787–1802.

Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2008). A critical approach to probability laws in geochemistry. *Mathematical Geosciences*, 40(5), 489–502.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Barceló-Vidal, C. (2005). The additive logistic skew-normal distribution on the simplex.*Stochastic Environmental Research and Risk Assessment*, 19, 205-214.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Egozcue, J. J. (2011). The principle of working on coordinates. *Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn & A. Buccianti eds., John Wiley & Sons, Chichester, 31–42.

Mateu-Figueras, G., Pawlowsky-Glahn, V. and Martín-Fernández, J. A. (2002). Normal in R+ vs lognormal in R. *Terra Nostra*, 3, 305–310.

McAlister, D. (1879). The law of geometric mean. Proceedings of the Royal Society of London, 29, 367–376.

Monti, G. S., Mateu-Figueras, G. and Pawlowsky-Glahn, V. (2011). Notes on the Scaled Dirichlet Distribution. *Compositional Data Analysis: Theory and Applications*, V. Pawlowsky-Glahn & A. Buccianti eds., John Wiley & Sons, Chichester, 128–138.

Pawlowsky-Glahn, V. (2003). Statistical modelling on coordinates. *Compositional Data Analysis Workshop – CoDaWork'03 Proceedings*, S. Thió-Henestrosa & J. A. Martín-Fernández eds., Universitat de Girona.

Pawlowsky-Glahn, V. and Egozcue, J. J. (2001). Geometric approach to statistical analysis on the simplex. *Stochastic Environmental Research and Risk Assessment*, 15(5), 384–398.

Pearson, K. (1897). Mathematical contributions to the theory of evolution. On a form of spurious correlation which may arise when indices are used in the measurement of organs. *Proceedings of the Royal Society of London*, LX, 489–502.

Stevens, J. (1986). *Applied Multivariate Statistics for the Social Sciences*. Lawrence Erlbaum Associates, Hillsdale, USA.

Tolosana-Delgado, R. (2005). *Geostatistics for Constrained Data: Positive Data, Compositions and Probabilities. Application to Environmental Hazard Monitoring.* Ph.D. Diss., Universitat de Girona.

Tolosana-Delgado, R. and Pawlowsky-Glahn, V. (2007). Kriging regionalized positive variables revisited: sample space and scale considerations. *Mathematical Geology*, 39, 529-558.

von Eynatten, H., Pawlowsky-Glahn, V. and Egozcue, J. J. (2002). Undestanding perturbation on the simplex: a simple method to better visualise and interpret compositional data in ternary diagrams. *Mathematical Geology*, 34, 249–257.

SORT 37 (1) January-June 2013, 57-78

New insights into evaluation of regression models through a decomposition of the prediction errors: application to near-infrared spectral data

María Isabel Sánchez-Rodríguez[1,*], Elena Sánchez-López[2], José M[a] Caridad[1], Alberto Marinas[2], Jose M[a] Marinas[2] and Francisco José Urbano[2]

Abstract

This paper analyzes the performance of linear regression models taking into account usual criteria such as the number of principal components or latent factors, the goodness of fit or the predictive capability. Other comparison criteria, more common in an economic context, are also considered: the degree of multicollinearity and a decomposition of the mean squared error of the prediction which determines the nature, systematic or random, of the prediction errors. The applications use real data of extra-virgin oil obtained by near-infrared spectroscopy. The high dimensionality of the data is reduced by applying principal component analysis and partial least squares analysis. A possible improvement of these methods by using cluster analysis or the information of the relative maxima of the spectrum is investigated. Finally, obtained results are generalized via cross-validation and bootstrapping.

MSC: 62H25, 62J05, 62Q99.

Keywords: Principal components, partial least squares, multivariate calibration, near-infrared spectroscopy.

1. Introduction

Principal component analysis (PCA) and *partial least squares* (PLS) are widely used in linear modelling when the number of explanatory variables greatly exceeds the number of observations. PCA and PLS calculate, from the explanatory variables, a reduced

* Corresponding author e-mail: td1sarom@uco.es
[1] Dep. Estadística, Econometría, I.O., Org. Empresas y Ec. Aplicada. Avda. Puerta Nueva, s/n. 14071. University of Córdoba. Córdoba. Spain.
[2] Dep. Química Orgánica. Campus de Excelencia Agroalimentario ceiA3. Edificio C-3 (Marie Curie-Anexo). Campus de Rabanales. 14014. University of Córdoba. Córdoba. Spain.
Received: April 2012
Accepted: October 2012

number of components or latent factors orthogonal among themselves. These components or factors are obtained as linear combinations of the explanatory variables, for PCA explaining the variability among these variables and, for PLS maximizing the covariance between each explanatory variable and the response one. Both methodologies reduce the dimensionality of the space of explanatory variables as the information provided by these variables is summarized in only a few ones.

PCA and PLS have been used in the last decades in some chemometric areas such as, for example, in pattern recognition (in this context, PCA or PLS linear discriminant analyses establish classification models based on experimental data in order to assign unknown samples to a sample class) and in multivariate calibration, in which PCA or PLS regression models predict a numeric variable as a function of several explanatory ones. Although papers comparing the goodness between PCA and PLS are well-known, most of them even considering PLS preferable to PCA for both regression and discrimination (see, for example, Frank and Friedman (1993) or Barker and Rayens (2003)), the fact is that PCA (besides PLS) is still widely used nowadays in chemometrics. Papers such as Gurdeniz and Ozen (2009), López-Negrete de la Fuente, García-Muñoz and Blegler (2010), Mevik and Cederkvist (2004), Nelson, MacGregor and Taylor (2006) and Yamamoto et al. (2009) can be cited as examples of using PCA in discrimination and calibration. For this reason, this paper revisits the comparison between PCA and PLS regressions in new terms. Firstly, the possible improvement of the regression models incorporating causal additional information of data is analyzed. Secondly, a proposed decomposition of the prediction errors makes it possible to determine the nature of these errors and evaluate their predictive capacity.

In this paper, the described methodology is applied to data obtained by near-infrared (NIR) spectroscopy. The NIR methods are used in food chemistry providing fast, reliable and cost-effective analytical procedures which, contrary to some others – such as gas chromatography – require no or little sample manipulation. Even though the data acquisition process is relatively easy for all spectral techniques, interpretation of spectra can be difficult. Separation techniques, such as gas chromatography, lead to discrete information including several usually well-defined, separated peaks from which, on proper integration, the content of various chemical components in the sample can be determined. On the contrary, spectroscopy generates continuous information, rich in both isolated and overlapping bands attributed to vibration of chemical bonds in molecules, which leads to the availability of multivariate data matrices. In this context, the use of mathematical and statistical procedures allows us to extract the maximum useful information from data (Berrueta, Alonso-Salces and Héberger, 2007).

There are many chemometric papers establishing comparison criteria of models. Thus, for example, Gowen et al. (2010) or Li, Morris and Martin (2002) propose some measures to determine the optimal number of latent factors in PLS regression models; Anderson (2009) compares diverse models of PLS regression as a function of their stability; Andersen and Bro (2010) or Reinaldo, Martins and Ferreira (2008) propose several selection criteria for variables in multiple calibration models; and Mevik and

Cerderkvist (2004) provide estimators of the mean squared error of prediction (MSEP) in PCA and PLS regression models. The aim of this paper is to compare PCA and PLS regression models on the basis of some criteria such as the number of latent factors or components, the goodness of fit and the predictive capability. However, this study goes a step further, incorporating an approach usually associated with an economic context. The degree of multicollinearity (absent when the regressors of the model are uncorrelated among themselves) is considered. Moreover, a decomposition of MSEP is proposed in order to point out the nature, systematic or random, of the prediction errors. As a final conclusion, the development of the study highlights the potential of the PLS regression.

There are several examples in the literature on the application of PCA and PLS regression models to near-infrared spectral data from oils and fats. For instance, Dupuy et al. (1996), Gurdeniz and Ozen (2009), Kasemsumran et al. (2005) and Öztürk, Yalçın and Özdemir (2010) use these multivariate calibration models to predict the content of some olive oil compounds in order to detect possible adulteration with some other vegetable oil. In the present study, the application is carried out by using NIR spectral data of extra-virgin olive oil and estimates the capability of the models to predict the oleic acid content. However, our approach could be used to estimate some other chemicals or features of importance in food chemistry from spectral data (see Mevik and Cederkvist (2004)). Firstly, the regression models are fitted by applying PCA and PLS from all the variables associated to different wavelengths of the spectrum (considering the matrix of data as a black box). Later on, models incorporating information provided by the relative maxima of the curve are estimated, because the principal components and the factors are obtained, in an independent manner, in each spectral peak. Then, PCA and PLS regressions are applied in combination with cluster analysis, a multivariate statistical technique that uses a measure of distance or similarity to classify a set of variables or cases in clusters of variables or cases, respectively, similar among themselves; in this case, components and factors are obtained independently in each cluster of wavelengths. The above-mentioned criteria are calculated for each model in order to evaluate their performance. For models in which PCA or PLS are carried out in an independent manner in different parts of the spectrum and so the resulting components or factors are not orthogonal among themselves, the degree of multicollinearity is also considered. Finally, techniques of cross-validation and bootstrapping are incorporated to extend the previous results to more general applications.

2. Review of selection criteria in regression models

2.1. Common comparison criteria

a) ***Goodness of fit.*** Let s_Y^2 and $s_{\widehat{Y}}^2$ be the respective variances of the observations, $y_1, y_2, \ldots, y_n$, of the dependent variable Y, and the corresponding predictions,

$\widehat{y}_1, \widehat{y}_2, \ldots, \widehat{y}_n$, in a regression model. The coefficient of determination, $R^2 = s^2_{\widehat{Y}}/s^2_Y$, ranges in the interval $[0,1]$ by definition and, expressed as a %, indicates the percentage of variability of the dependent variable explained by the regression model. Obviously, a model is better as the coefficient of determination approaches 1. The adjusted coefficient of determination, $\overline{R}^2$, is calculated from R^2, taking into account the number of observations and the number of the regressors in the regression, in such a way that the goodness of fit is not overestimated.

The mean squared error of calibration, MSEC $= \sum_{i=1}^{n} (y_i - \widehat{y}_i)^2 / n$, takes values nearer to 0 for a good fit, but it is non-dimensionless, that is, it depends on the units of measure of the variable.

There are other measures of the goodness of fit, that are not contemplated in this study, based on the likelihood criterion (see Burnham and Anderson (2004)).

b) ***Predictive capability.*** Given the predictions for the future t observations, $\widehat{y}_{n+1}$, $\widehat{y}_{n+2}, \ldots, \widehat{y}_{n+t}$, of a certain regression model, the mean squared error of the prediction, MSEP$= \sum_{j=1}^{t} (y_{n+j} - \widehat{y}_{n+j})^2 / t$, evaluates the predictive capability of a regression model. The predictive capability of a model is obviously better as MSEP approaches 0, taking into account that it also depends on the measurement units.

As is indicated by Berrueta et al. (2007), the ideal situation is when there are enough data available to create separate test set completely independent from the model building process (this validation procedure is known as external validation). When an independent test set is not available (e.g., because cost or time constraints make it difficult to increase the sample size), MSEP has to be estimated from the learning data, that is, the data used to train the regression. In this context, Mevik and Cederkvist (2004) present several estimators for MSEP, based on cross-validation or bootstrapping: Let $\mathbf{X} = [X_1|X_2|...|X_p]$ be the matrix containing the explanatory variables in a regression model and let Y be the dependent variable. For a set of n observations, it is assumed that $L = \{(\mathbf{x}_i, y_i) : i = 1, \ldots, n_L\}$ is a learning data set (of n_L observations) and $T = \{(\mathbf{x}_{n_L+i}, y_{n_L+i}) : i = 1, \ldots, n_T\}$ is a test data set (of size n_T). Besides, f_L is a predictor trained on L. When L is divided randomly into K segments, $L_1, L_2, \ldots, L_K$, of roughly equal size $(n_1, n_2, \ldots, n_K)$, f_k is a predictor trained on $L \backslash L_k$. Finally, R bootstrap samples are drawn in L, $L_1^*, L_2^*, \ldots, L_R^*$, and f_r^* is a predictor trained on L_r^*. In the described context, Mevik and Cederkvist (2004) present the MSEP estimators shown in Table 1.

c) ***Number of regressors.*** Attending to the parsimony principle, if some regression models present similar characteristics in terms of goodness of fit, predictive capacity and multicollinearity, the simplest among them, i.e. the one with the smallest number of regressors, is considered the best.

Table 1: MSEP *estimators adopted from Mevik and Cederkvist (2004).*

MSEP Estimator	**Definition**
Test set estimate	$MSEP_{\text{test}} = \frac{1}{n_T}\sum_{i=1}^{n_T}\left(f_L\left(\boldsymbol{x}_{n_L+i}\right)-y_{n_L+i}\right)^2 (=\text{MSEP})$
Apparent *MSEP*	$MSEP_{\text{app}} = \frac{1}{n_L}\sum_{i=1}^{n_L}\left(f_L\left(\boldsymbol{x}_i\right)-y_i\right)^2 (=\text{MSEC})$
Cross-validation	$MSEP_{\text{cv}.K} = \frac{1}{n_L}\sum_{k=1}^{K}\sum_{i\in L_k}\left(f_k\left(\boldsymbol{x}_i\right)-y_i\right)^2$
Adjusted cross-validation	$MSEP_{\text{adj.cv}.K} = MSEP_{\text{cv}.K} + MSEP_{\text{adj}}$, where $MSEP_{\text{adj}} = MSEP_{\text{app}} - \frac{1}{n_L}\sum_{k=1}^{K}\frac{n_k}{n_L}\sum_{i\notin L_k}\left(f_k\left(\boldsymbol{x}_i\right)-y_i\right)^2$
Naive bootstrap estimate	$MSEP_{\text{naive}} = \frac{1}{R}\sum_{r=1}^{R}\frac{1}{n_L}\sum_{i=1}^{n_L}\left(f_r^*\left(\boldsymbol{x}_i\right)-y_i\right)^2$
Ordinary bootstrap estimate	$MSEP_{\text{boot}} = MSEP_{\text{app}} + Bias_{\text{app}}$, where $Bias_{\text{app}} = \frac{1}{R}\sum_{r=1}^{R}\left(\frac{1}{n_L}\sum_{i=1}^{n_L}\left(f_r^*\left(\boldsymbol{x}_i\right)-y_i\right)^2 - \frac{1}{n_L}\sum_{i=1}^{n_L}\left(f_r^*\left(\boldsymbol{x}_i^r\right)-y_i^r\right)^2\right)$, where $\left(\boldsymbol{x}_i^r, y_i^r\right)$ is the ith observation of the rth bootstrap sample
Bootstrap smoothed cross-validation	$MSEP_{\text{BCV}} = \frac{1}{n_L}\sum_{i=1}^{n_L}\frac{1}{R_{-i}}\sum_{r:i\notin L_r^*}\left(f_r^*\left(\boldsymbol{x}_i\right)-y_i\right)^2$, where R_{-i} is the number of bootstrap samples excluding observation i
The 0.632 estimate	$MSEP_{0.632} = 0.632\cdot MSEP_{\text{BCV}} + (1-0.632)\cdot MSEP_{\text{app}}$, where $0.632 \approx 1-e^{-1}$ is approximately the average fraction of distinct observations in each bootstrap data set

In PCA, the Kaiser criterion is the default in SPSS and most statistical software (but many authors do not recommend to use it as the only cut-off criterion as it tends to extract too many factors): Let $X_1^*, X_2^*, \ldots, X_p^*$ be the standardized variables of the explanatory variables, $X_1, X_2, \ldots, X_p$. When a random sample of dimension n is considered, $\mathbf{X}^* = [X_1^* | X_2^* | \cdots | X_p^*]$ is a matrix of dimension $n \times p$. Then, $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$ is a square $p \times p$ matrix and has p eigenvalues, $\lambda_1, \lambda_2, \ldots, \lambda_p$. The eigenvalue λ_k represents the variance of the k-th principal component (or factor), $k = 1, \ldots, p$. The Kaiser criterion suggests that those factors with eigenvalues equal or higher than 1

should be retained (taking into account that the variables are standardized and so the average of the eigenvalues is precisely 1).

In PLS analysis, the criterion of the first increase of the mean squared error of prediction is considered: the number of latent factors taken into account is $h^* = min\{h > 1 : MSEP(h+1) - MSEP(h) > 0\}$, where $MSEP(h)$ the mean squared error of prediction of the regression model with h factors.

Gowen et al. (2010) show that the over-fitting in a regression model entails some additional problems, such as the introduction of noise in the regression coefficients. More specifically, their paper presents some measures for preventing the over-fitting in PLS calibration models of NIR spectroscopy data, investigating the use of both model bias and variance simultaneously in selecting the number of latent factors to include in the model. Initially, the authors consider the Durbin-Watson statistic:

$$DW = \frac{\sum_{i=1}^{p} (b_i - b_{i-1})^2}{\sum_{i=0}^{p} b_i^2},$$

being p the number of the regressors and $b_0, b_1, \ldots, b_p$ the coefficients of the multiple regression model. The named regression vector measure, RVM, is calculated by rescaling DW from 0 to 1. Then, a bias measure, BM, is obtained once the root of the mean squared error of calibration, RMSEC, is rescaled from 0 to 1. Gowen et al. (2010) propose to obtain the measures RVM_j and BM_j for models with j latent factors or components, varying j. Finally, the optimal number of latent factor to consider in a PLS regression model is j^* if the minimum of the sum $RVM_j + BM_j$ is obtained for $j = j^*$.

2.2. Other comparison criteria

In this section, other comparison criteria, more frequent in economics research, are proposed. Thus, for example, the decomposition of MSEP provided in d) below is developed in EViews, a program of econometric analysis. Similarly, Essi, Chukuigwe and Ojekudo (2011), Greenberg and Parks (1997), Mynbaev (2011), Spanos and McGuirk (2002) and Yamagata (2006) deal with the multicollinearity under different hypotheses in an economic context. These new criteria establish additional arguments to the ones proposed in a)-c) and can assist in selecting the most adequate model.

d) ***Decomposition of MSEP.*** In Section 2.1.b, MSEP has been established as a criterion for evaluating the predictive capability of a model which, in general terms, is better as MSEP approaches 0. But this issue can be dealt more in depth, trying to determine the causes of the prediction errors.

Around 1920, Fisher introduced analysis of variance (ANOVA), a collection of statistical procedures in which the observed variance in a particular variable is partitioned into components attributable to different sources of variation. Diverse authors, e.g. Climaco-Pinto et al. (2009), Mark (1986), Mark and Workman (1986), Zwanenburg et al. (2011) have used the ANOVA in a chemometric context. We use this technique to decompose MSEP into three components, with the aim of investigating if there is any systematic cause that produces the prediction errors or if they are randomly distributed.

Given the predictions for the future t observations, $\widehat{y}_{n+1}, \widehat{y}_{n+2}, \ldots, \widehat{y}_{n+t}$, of a certain regression model, $\overline{y}$ and $\overline{\widehat{y}}$ are the means of the observations and the predictions, respectively, s_Y and $s_{\widehat{Y}}$ are the corresponding standard deviations and $s_{Y\widehat{Y}}$ is the covariance. Then, the MSEP,

$$MSEP = \frac{1}{t}\sum_{j=1}^{t}\left(y_{n+j} - \widehat{y}_{n+j}\right)^2 = \frac{1}{t}\sum_{j=1}^{t} y_{n+j}^2 + \frac{1}{t}\sum_{j=1}^{t}\widehat{y}_{n+j}^2 - \frac{2}{t}\sum_{j=1}^{t} y_{n+j}\widehat{y}_{n+j},$$

can be decomposed, once the terms $\left(\overline{y} - \overline{\widehat{y}}\right)^2$ and $2s_{Y\widehat{Y}}$ are added and subtracted, in the following way

$$MSEP = \left(\overline{y} - \overline{\widehat{y}}\right)^2 + \left(s_Y - s_{\widehat{Y}}\right)^2 + 2\left(s_Y s_{\widehat{Y}} - s_{Y\widehat{Y}}\right) = E_B + E_V + E_R,$$

or, equivalently, by the identity

$$1 = \frac{E_B}{MSEP} + \frac{E_V}{MSEP} + \frac{E_R}{MSEP} = U_B + U_V + U_R,$$

where U_B is the part of MSEP corresponding to the bias, representing systematic errors in the prediction; U_V indicates the difference between the variability of the real values and the variability of the observed ones; finally, U_R shows the random variability in the prediction errors.

The decomposition of MSEP evidences that the predictions are affected by systematic and random errors. Random errors are, in general, low in absolute value, resulting from the additive effect of many insignificant events (detected with difficulty) and so inherent to a process. This kind of error can only be reduced with the increasing of the sample size, and fluctuate around a constant value, being distributed as a *white noise*. However, systematic errors are usually associated with an identifiable cause, such as an interference in the observation process or a defect of calibration in the instrument of measurement. They usually originate in a great fluctuation in the evolution of a process and must be detected and eliminated (for

example, this is the objective of statistical quality control or the aim of papers such as Guldberg et al. (2005) or Vasquez and Whiting (2006)).

A model is obviously better as MSEP approaches 0 (taking into account that MSEP is not upperly bounded and depends on the unit of measurement). But, using the proposed decomposition, if MSEP shows a great percentage attributable to systematic errors, this aspect indicates that there is some detectable cause originating these deviations in the predictions. This cause must be detected in order to eliminate systematic errors. Thus, a great percentage of MSEP attributable to systematic prediction errors indicates that the fit model can be improved in some sense. Nevertheless, this improvement is difficult if the predictions generated by a model have a random nature.

However, the study of the statistical general linear model (in particular, the multivariate linear regression model) assumes the random nature of its perturbations (which must be, by hypotheses, centered, homoscedastic, uncorrelated and normally distributed random variables). And so the presence of systematic errors in the predictions (represented by a high U_B ratio) or the discrepancy between the variability of the real and the observed values (represented by a high U_V ratio) prevent the validation of the fitted model, since these facts point out the absence of the hypotheses of randomness and homoscedasticity.

Definitively, the ideal situation for evaluating the predictive capability of a model is presented when MSEP has a value nearer to 0 and besides $U_B = 0$, that is, systematic errors do not exist in the prediction; $U_V = 0$, which indicates that the variability of the real values is the same as that of the predictions; and $U_R = 1$, which corresponds to random prediction errors.

e) ***Possible existence of multicollinearity***[1]. In the fit of a regression model, it is frequent the appearance of a certain linear relationship among the regressors, which can be even exact (for example, when the number of cases is lower than the number of explanatory variables). The presence of multicollinearity in the regression makes that the least squares estimators obtained are not, in general, very precise. Although these estimators are still linear, unbiased and efficient (Gauss-Markov theorem), the multicollinearity complicates the precise quantification of the effect of each regressor on the dependent variable, because the variances of the estimators are high.

1. In PCA and PLS regressions, the orthogonal character of the components or factors guarantees the absence of multicollinearity in the model. In this paper, multicollinearity is evaluated in models whose components or factors are obtained applying PCA or PLS to different parts of the spectrum, in an independent manner. Thus, these components or factors are uncorrelated only in the corresponding spectral part.

In a multiple linear regression model, the estimator of the variance of a certain coefficient, $\widehat{\beta}_j$, is given by the expression

$$\widehat{Var\left(\widehat{\beta}_j\right)} = \frac{\widehat{\sigma}^2}{p\left(1-R_j^2\right)s_j^2}, \qquad j = 1,\ldots,p,$$

where $\widehat{\sigma}^2$ is the estimation of the disturbance variance, assumed to be constant by the hypothesis of homoscedasticity; p is the number of explanatory variables in the model; R_j^2 is the coefficient of determination of the regression of the variable X_j on the rest of the explanatory variables; and s_j^2 is the observed variance of X_j.

The ***variance inflation factor***, VIF, is defined as the ratio between the observed variance and the variance existing when X_j is uncorrelated to the rest of the regressors of the model (and, then, $R_j^2 = 0$). Some authors consider that there is a grave multicollinearity when $VIF\left(\widehat{\beta}_j\right) > 10$ for any $j = 1,\ldots,p$, that is, when $R_j^2 > 0.90$. Some computational programs (SPSS, for example) define the term "tolerance" as $T_j = 1 - R_j^2$; in this case, a serious multicollinearity is identified when $T_j < 0.10$ for any $j = 1,\ldots,p$.

Then, let $\mathbf{X}^{*\mathsf{T}}\mathbf{X}^*$ be the matrix defined in Section 2.1.c ($\mathbf{X}^*$ contains the standardized observations). As indicated in that section, it is a square matrix of dimension p and, therefore, has p eigenvalues. In this case, its condition number, κ, is defined as the root of the ratio between the highest eigenvalue (λ_{max}) and the lowest one (λ_{min}). The condition number measures the sensitivity of the least-squares estimates to small changes in the data. The multicollinearity can be considered as serious if κ (which is not affected by the measurement units because it is calculated, as stated above, from standardized variables) ranges between 20 and 30; if κ is greater than 30, the multicollinearity is very serious.

3. Materials and methods

3.1. Acquisition of spectral data

This work is based on data obtained from olive oil from different olive varieties (mainly 'Zaity', 'Arbequina', 'Frantoio', 'Picual' and 'Hojiblanca') harvested in the 2005/06, 2006/07, 2007/08 and 2008/09 seasons. Samples correspond to Andalusian olive oils principally, though some others from Tarragona and Edleb (Syria) have also been included. There are 302 cases in total. Olive oil was either extracted by the producers through a two-phase centrifugation system or by the staff of the Agronomy Department of University of Córdoba via the Abencor System. This system reproduces the industrial

process on the laboratory scale and follows the same stages of grinding, beating, centrifugation and decantation.

^{1}H-NMR analyses were carried out at the NMR Service of the University of Sevilla on a Bruker Avance spectrometer (Kahlsruhe, Germany), by using a resonance frequency of 500.2MHz and a direct-detection 5mm QNP 1H/15N/13C/31P probe. Determination of oleic acid content was carried out following the method suggested by Guillén and Ruiz (2003). NIR spectra were obtained by the staff of the Organic Chemistry Department of the University of Córdoba within 15 days after reception of the samples, which where kept in the fridge so that properties were not modified (Baeten et al., 2003). The instrument employed for spectra collection was available at the Central Service of Analyses (SCAI) at the University of Córdoba. It consisted of a Spectrum One NTS FT-NIR spectrophotometer (Perkin Elmer LLC, Shelton, USA) equipped with an integrating sphere module. Samples were analyzed by transflectance by using a glass petri dish and a hexagonal reflector with a total transflectance pathlength of approximately 0.5 mm. A diffuse reflecting stainless steel surface placed at the bottom of the cup reflected the radiation back through the sample to the reflectance detector. The spectra were collected by using Spectrum Software 5.0.1 (Perkin Elmer LLC, Shelton, USA). The reflectance (log 1/R) spectra were collected with two different reflectors. Data correspond to the average of results with both reflectors, thus ruling out the influence of them on variability of the obtained results. Moreover, spectra were subsequently smoothed using the Savitzky-Golay technique (that performs a local polynomial least squares regression in order to reduce the random noise of the instrumental signal). Once pre-treated, NIR data of 1237 measurements for each case (representing energy absorbed by olive oil sample at 1237 different wavelengths, from 800.62 to 2499.64 nm) were supplied to the Department of Statistics (University of Córdoba) in order to be analyzed.

3.2. Calibration models

As stated above, the aim of this study is to compare PCA and PLS regression models following the criteria described in Section 2. In this application, the regression models predict the content in monounsaturated acids (fundamentally, oleic acid, fatty acid of the omega 9 series with beneficiary cardiovascular and hepatic effects) of extra-virgin olive oil by using NIR spectral data. For each statistical case, that is, for each oil sample – $n = 302$, in total – the observations corresponding to $p = 1237$ wavelengths of the spectrum – varying from 800.62 to 2499.64 nm – are available. Therefore, a statistical approach considers a matrix of data, $\mathbf{X}$, of dimensions $n = 302 \times p = 1237$, whose rows are referred to the cases studied and the columns are associated to the different explanatory variables in the regression. The dependent variable, Y, is given by the content in oleic acid of olive oil, in percentage, observed by using proton nuclear magnetic resonance (^{1}H-NMR). The information provided by the potential explanatory variables (1237 in total, corresponding to the different wavelengths), will be summarized

in a reduced number of uncorrelated factors in order to avoid multicollinearity, due to the high dimensionality of the space of the explanatory variables. The factors are obtained by using the procedures described as follows:

Method 1. *Extraction of latent factors from the whole space of explanatory variables*

Firstly, a small number of latent factors or components are determined from the whole space of 1237 explanatory variables. The factors are obtained as linear combinations of the explanatory variables and summarize the information provided by these variables. The components are extracted by PCA and, later on, by using PLS. In PCA, the factors initially considered are associated to the eigenvalues of the correlation matrix of the explanatory variables greater than 1 (Kaiser criterion), resulting 6 components (as λ_6=1.706 and λ_7=0.941). In PLS analysis, the criterion of the first increase of MSEP (see Section 2.1.c) is considered; as shown in Table 2, $h^* = 9$ in this case. Then, the number of factors is increased to 15, number of components closer to the ones considered by next Methods 2 and 3. For subsequent comparisons, the results for 6, 9 and 15 latent factors in PCA and PLS are considered. The percentage of the explanatory variables explained, in each case, by the extracted factors is greater than 99%.

Table 2: *Optimal number of factors in PLS analysis.*

Nr. components	1	2	3	4	5	6	7	8	9	10
$MSEP(h)$	20.68	20.14	13.87	9.22	8.41	6.49	2.07	1.42	0.79	0.89
$MSEP(h+1)-MSEP(h)$	−0.54	−6.27	−4.65	−0.81	−1.92	−4.42	−0.65	−0.63	0.10	

Table 3: *Optimal number of factors (according to criterion by Gowen* et al. *(2010)).*

Model	No. factors (j)	DW_j	RVM_j	$RMSEC_j$	BM_j	$RVM_j + BM_j$
1.1.1 (6 PCA)	6	1.004	0.326	3.673	1	1.326
1.1.2 (9 PCA)	9	1.001	0	3.323	0.806	**0.806**
1.1.3 (15 PCA)	15	1.010	1	1.868	0	1
2.1.1 (6 PLS)	6	1	1	2.363	1	2
2.1.2 (9 PLS)	9	0.999	0.568	1.252	0.329	0.897
2.1.3 (15 PLS)	15	0.998	0	0.707	0	**0**

Once the components summarizing the sample information have been obtained, PCA regression models (Models 1.1.1, 1.1.2 and 1.1.3 with 6, 9 and 15 factors, respectively) and PLS regression models (Models 2.1.1, 2.1.2 and 2.1.3 with 6, 9 and 15 factors, respectively) are proposed. These models consider the content in oleic acid by ^{1}H-NMR spectroscopy as explained variable (Y) and the previously obtained factors as regressors. The last column of Table 3 shows that, based on the criterion presented in Gowen et al.

(2010) (see the measures defined in Section 2.1.c), the optima models among PCA and PLS regression ones are those with 9 and 15 components, respectively.

Method 2. *Extraction of latent factors from the different spectral peaks*

NIR spectroscopy yields spectra presenting both isolated and overlapping bands assigned to vibrations of one or more chemical bonds in molecules. For this reason, the explanatory variables associated to wavelengths corresponding to NIR spectral peaks could contain valuable information to predict the content in oleic acid of olive oil. Thus, wavelength intervals associated to spectral peaks are determined (Figure 1 shows six regions corresponding to wavelengths 800.62-936.74 nm, 1142.99-1280.49 nm, 1349.24-1486.74 nm, 1658.62-1899.24 nm, 2105.49-2208.62 nm, 2242.99-2499.64 nm, approximately). Therefore, if $\mathbf{X}$ is the matrix containing the 1237 explanatory variables, $\mathbf{X}$ can be divided into six boxes, $\mathbf{X}^p_{(1)},\mathbf{X}^p_{(2)},\ldots,\mathbf{X}^p_{(6)}$, each one containing the explanatory variables associated to the corresponding region and a seventh box, with residual character, $\mathbf{X}^p_{(r)}$, containing the remaining explanatory variables: $\mathbf{X} = \left[\mathbf{X}^p_{(1)}|\mathbf{X}^p_{(2)}|\cdots|\mathbf{X}^p_{(6)}|\mathbf{X}^p_{(r)}\right]$.

Then, PC and PLS analyses are applied to each of the seven boxes previously considered, in an independent manner, with the aim of determining factors summarizing the information associated to each region of the spectrum (Table 4). Afterwards, a PCA regression model (Model 1.2, Peaks PCA) and a PLS regression model (Model 2.2, Peaks PLS) are proposed to predict the content in oleic acid of olive oil, Y, considering the above-mentioned factors as regressors. The regressors (principal components or factors) in these last models are not uncorrelated among themselves; they are only orthogonal for each of the defined boxes: $\mathbf{X}^p_{(1)},\mathbf{X}^p_{(2)},\ldots,\mathbf{X}^p_{(6)},\mathbf{X}^p_{(r)}$. This fact introduces any degree of multicollinearity in the models.

Method 3. *Extraction of latent factors from the different clusters of spectral wavelengths*

Cluster analysis is applied to determine ten groups of similar explanatory variables, in terms of the squared Euclidean distance, in order to predict the composition in oleic acid of the olive oil. Therefore, the matrix of the explanatory variables, $\mathbf{X}$, is expressed as $\mathbf{X} = \left[\mathbf{X}^c_{(1)}|\mathbf{X}^c_{(2)}|\cdots|\mathbf{X}^c_{(10)}\right]$, where $\mathbf{X}^c_{(i)}$ contains the explanatory variables classified in the i-th cluster, $i = 1,\ldots,n$, after the application of the procedure. The graphical and analytical results obtained, in this case, are shown in Figure 2 and Table 5, respectively.

As in Method 2, PCA and PLS are applied to summarize in a reduced number of components or factors the information of the explanatory variables associated to each cluster, in an independent manner (which also introduces a certain degree of multicollinearity among components or factors). Subsequently, a PCA regression model (Model 1.3, Clusters PCA) and a PLS regression model (Model 2.3, Clusters PLS) are proposed considering the estimated factors as explanatory variables and the content in oleic acid, as determined by ^{1}H-NMR, as dependent variable.

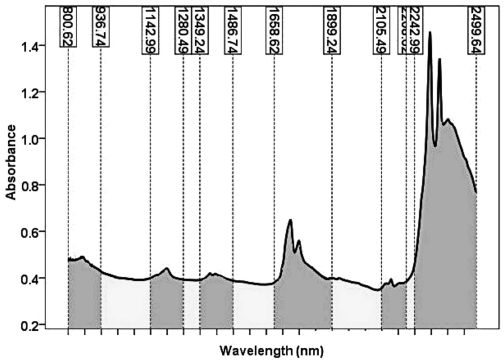


Figure 1: *Wavelength intervals associated to spectral peaks.*

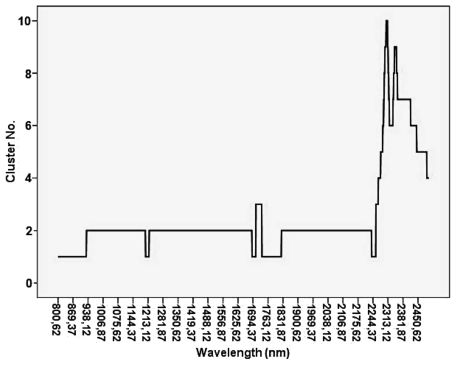


Figure 2: *Clusters of wavelength.*

Table 4: *Factors in wavelength intervals associated to spectral peaks.*

Wavelength interval	**Nr. int. var.**	**X Box**	*% Y* **var.**$^{(a)}$	**Nr. fac.**$^{(b)}$	*%* **int. var.**$^{(c)}$ **(PCA)**	*%* **int. var.**$^{(c)}$ **(PLS)**
800.62-936.74	100	$\mathbf{X}^p_{(1)}$	50.8	2	98.49	98.2
1142.99-1280.49	101	$\mathbf{X}^p_{(2)}$	52.1	1	99.30	99.2
1349.24-1486.74	101	$\mathbf{X}^p_{(3)}$	35.0	1	99.03	98.8
1658.62-1899.24	176	$\mathbf{X}^p_{(4)}$	91.4	3	99.61	99.4
2105.49-2208.62	76	$\mathbf{X}^p_{(5)}$	81.7	1	99.32	99.4
2242.99-2499.64	188	$\mathbf{X}^p_{(6)}$	95.3	2	99.32	97.9
Rest of wavelenghts	495	$\mathbf{X}^p_{(e)}$	82.2	3	99.49	99.4

(a) Percentage of Y variance explained by $\mathbf{X}^p_{(i)}$

(b) Number of factors according to Kaiser criterion in PCA

(c) Percentage of $\mathbf{X}^p_{(i)}$ variance explained by interval factors

4. Results and discussions[2]

Taking into account the results shown in Table 6 and Table 7, the comparison among the values $\overline{R}^2$, *MSEP* and κ allows us to conclude that all the PLS regression models clearly provide better results in terms of goodness of fit, predictive capability and multicollinearity than the corresponding to PCA regressions with the same number of latent factors.

2. The chemometric applications can be developed using different software. Some packages of statistical or mathematical analysis have implemented the principal techniques usual in chemometrics, such as PASW Statistics – formerly SPSS, currently belonging to IBM, UNSCRAMBLER from CAMO, the PLS Toolbox of MatLab from MathWorks, or the free package "pls" in R.

Table 5: *Factors in clusters of NIR spectrum.*

Cluster	**Nr. clus. var.**	**X Box**	*% Y* **var.**$^{(a)}$	**Nr. fac.**$^{(b)}$	**% clus. var.**$^{(c)}$ **(PCA)**	**% clus. var.**$^{(c)}$ **(PLS)**
1	119	$\mathbf{X}^c_{(1)}$	93.4	4	99.02	98.2
2	191	$\mathbf{X}^c_{(2)}$	95.2	4	99.65	99.5
3	12	$\mathbf{X}^c_{(3)}$	86.0	1	98.33	98.0
4	13	$\mathbf{X}^c_{(4)}$	44.4	1	98.92	98.5
5	41	$\mathbf{X}^c_{(5)}$	88.3	1	99.42	99.3
6	35	$\mathbf{X}^c_{(6)}$	80.3	1	99.50	98.7
7	50	$\mathbf{X}^c_{(7)}$	85.4	1	99.81	99.8
8	10	$\mathbf{X}^c_{(8)}$	72.2	1	99.60	99.6
9	13	$\mathbf{X}^c_{(9)}$	49.4	1	99.59	99.6
10	5	$\mathbf{X}^c_{(10)}$	5.0	1	99.84	99.8

(a) Percentage of Y variance explained by $\mathbf{X}^c_{(i)}$

(b) Number of factors according to Kaiser criterion in PCA

(c) Percentage of $\mathbf{X}^c_{(i)}$ variance explained by cluster factors

Table 6: *Comparison of models.*

Model	**Nr. fac.**	R^2	$\overline{R}^2$	*MSEP*	κ
1.1.1 (6 PCA)	$6^{(a)}$	0.023	-0.004	19.094	–$^{(c)}$
1.1.2 (9 PCA)	$9^{(b)}$	0.200	0.166	13.770	–$^{(c)}$
1.1.3 (15 PCA)	15	0.748	0.729	1.839	–$^{(c)}$
1.2 (Peaks PCA)	13	0.349	0.308	7.662	195.698
1.3 (Clusters PCA)	16	0.619	0.591	4.156	301.477
2.1.1 (6 PLS)	$6^{(a)}$	0.596	0.584	6.490	–$^{(c)}$
2.1.2 (9 PLS)	$9^{(b)}$	0.887	0.882	0.792	–$^{(c)}$
2.1.3 (15 PLS)	15	0.964	0.961	0.307	–$^{(c)}$
2.2 (Peaks PLS)	13	0.692	0.672	2.557	183.837
2.3 (Clusters PLS)	16	0.859	0.847	0.382	370.059

(a) Number of factors according to Kaiser criterion in PCA

(b) Number of factors according to the first increase of the MSEP in PLS regression

(c) Orthogonal factors

Focusing on PCA regression, the model with 15 latent factors calculated from the explanatory variables directly, neither extracting the components in each interval of wavelengths associated to spectral peaks nor applying cluster analysis, is the one that provides the best results in fit and prediction. This model is named Model 1.1.3 (15 PCA) and has associated values $\overline{R}^2_{113}$=0.729 and $MSEP_{113}$=1.840. Besides, the orthogonal

Table 7: *Decomposition of* MSEP.

Model	$\bar{\hat{y}}$	$s_{\hat{Y}}$	$s_{Y\hat{Y}}$	*MSEP*	E_B	E_V	E_R	U_B	U_V	U_R
1.1.1 (6 PCA)	80.902	0.493	0.557	19.094	0.834	15.061	3.200	0.044	0.789	0.168
1.1.2 (9 PCA)	81.189	1.811	4.516	13.770	0.392	6.570	6.808	0.028	0.477	0.494
1.1.3 (15 PCA)	81.268	3.569	15.164	1.840	0.299	0.648	0.892	0.163	0.352	0.485
1.2 (Peaks PCA)	81.115	2.387	8.830	7.662	0.490	3.946	3.227	0.064	0.515	0.421
1.3 (Clusters PCA)	80.860	3.251	13.229	4.156	0.912	1.261	1.983	0.220	0.303	0.477
2.1.1 (6 PLS)	80.775	3.015	11.406	6.490	1.083	1.848	3.560	0.167	0.285	0.549
2.1.2 (9 PLS)	81.228	3.996	17.328	0.792	0.346	0.143	0.303	0.436	0.180	0.383
2.1.3 (15 PLS)	81.625	4.397	19.098	0.307	0.036	0.001	0.270	0.118	0.002	0.881
2.2 (Peaks PLS)	81.407	3.537	14.627	2.556	0.167	0.700	1.689	0.065	0.274	0.661
2.3 (Clusters PLS)	81.746	4.039	17.533	0.382	0.005	0.112	0.265	0.013	0.295	0.693

Note: $\bar{y} = 81.8153$, $s_Y = 4.3740$

character of the components guarantees the absence of multicollinearity in the model. Finally, the decomposition of MSEP according to expression given in Section 2.2.d (see Table 7) points out that the last term, $U_{R,113}$=0.485, is the highest one, thus indicating that the prediction errors are random, ideal situation for the predictions of a model.

As regards PLS regression, the model in which the sample information is summarized directly from the explanatory variables in 15 PLS components (Model 2.1.3, 15 PLS) shows the best results: $\overline{R}^2_{213}$=0.961, $MSEP_{213} = 0.307$ and absence of multicollinearity because of the uncorrelated character of the latent factors. Likewise, this model has the highest value for the term U_R in the decomposition of *MSEP* ($U_{R,213}$=0.881 in Table 7); which again confirms the random nature of the prediction errors.

Taking into account the two previous conclusions, neither the distinction of the information associated to the spectral peaks nor the previous application of cluster analysis improve the results of the regression on the PC or PLS latent factors obtained directly (Method 1). In fact, the results are worse because of the appearance of multicollinearity, as the values of κ contained in Table 6 evidence.

Finally, in view of the above-mentioned considerations, Model 2.1.3 (15 PLS) is the best among all the fit models, presenting optimal characteristics regarding number of latent factors (Table 6 and Table 7), goodness of fit, predictive capability (and causes of prediction errors) and obviously absence of multicollinearity.

4.1. Cross-validation and bootstrapping

In this section, the attention is focused on the PLS regression model with 15 latent factor (Model 2.1.3, 15 PLS) as it has been considered the best among all the models studied above. This model will be compared to the PCA regression model with 15 components (Model 1.1.3, 15 PCA) in terms of cross-validation and bootstrapping with the aim of generalizing the previously obtained results.

Table 8: MSEP *estimations.*

MSEP **Estimation**	**PCA regression (Model 1.1.3)**	**PLS regression (Model 2.1.3)**
$MSEP_{\text{test}}$	1.838	0.308
$MSEP_{\text{app}}$	3.486	0.500
$MSEP_{\text{cv}.K}$	20.119	2.360
$MSEP_{\text{adj.cv}.K}$	22.539	5.609
$MSEP_{\text{naive}}$	3.672	0.492
$MSEP_{\text{boot}}$	4.224	0.569
$MSEP_{\text{BCV}}$	3.273	0.480
$MSEP_{0.632}$	3.352	0.487

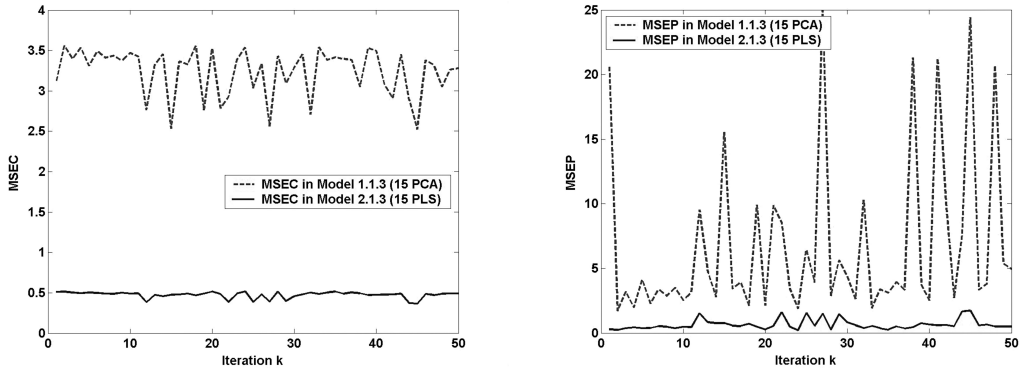


Figure 3: *$MSEC_i$ and $MSEP_i$ for Model 1.1.3 (15 PCA) and Model 2.1.3 (15 PLS) –$i = 1,2,\ldots,50$–.*

Firstly, Table 8 shows the estimations of MSEP obtained, in this case, from the different estimators considered in Mevik and Cederkvist (2004) and presented in Table 1 (see Section 2.1.b). The corresponding algorithms divide the learning data set, L, into $K = 6$ segments, $L_1, L_2, \ldots, L_6$, of equal size ($n_k = 39$) for cross-validation; so, 6 regression models, $f_1, f_2, \ldots, f_6$, are fit (where each model f_k uses the observations not contained in L_k). Regarding the bootstrapping, $R = 50$ bootstrap samples (of size 25), $L_1^*, L_2^*, \ldots, L_{50}^*$, are drawn from the learning data set, L. For $r = 1,2,\ldots,50$, f_r^* is the regression model trained on L_r^*. Table 8 shows that all the estimations of *MSEP* obtained by using the different algorithms described are greater for Model 1.1.3 (15 PCA) than for Model 2.1.3 (15 PLS). Again, this fact points out that the predictive capability is higher for the PLS model than for the PCA one.

Afterwards, also in the context of bootstrapping, another algorithm is programmed to compare MSEP in both models. Now, the objective is to change, in each iteration i of the algorithm ($i = 1,2,\ldots,50$), the learning data set, L_i, and the test data set, T_i.

Then, $MSEC_i (= MSEP_{app,i})$ and $MSEP_i (= MSEP_{test,i})$ are calculated and compared for each iteration i. Besides, $MSEP_i$ is decomposed (Section 2.2.d) in the components $U_{B,i}$, $U_{V,i}$ and $U_{R,i}$, in order to determine the nature of the prediction errors, investigating if they are randomly distributed or they respond to a systematical cause. In this context, Figure 3 shows that $MSEC_i$ and $MSEP_i$ – calculated for Model 1.1.3 (15 PCA) and Model 2.1.3 (15 PLS) – are clearly higher in PCA regression than in PLS regression, for each iteration $i = 1, 2, \ldots, 50$. Furthermore, the variability of both goodness of fit and predictive capability is higher in PCA regression, appearing for PLS regression as a *white noise*. Regarding the decomposition of MSEP, Figures 4 and 5 depict that, although the component $U_{R,i}$ is the highest in both PCA and PLS models for $i = 1, 2, \ldots, 50$, in PLS one $U_{R,i}$ represents a percentage of the variability of the prediction errors higher than in PCA one.

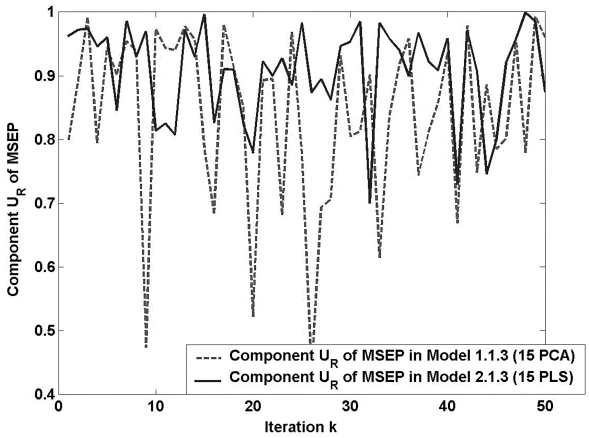


Figure 4: *Component $U_{R,i}$ of $MSEP_i$, –$i = 1, 2, \ldots, 50$–, for Model 1.1.3 (15 PCA) and Model 2.1.3 (15 PLS).*

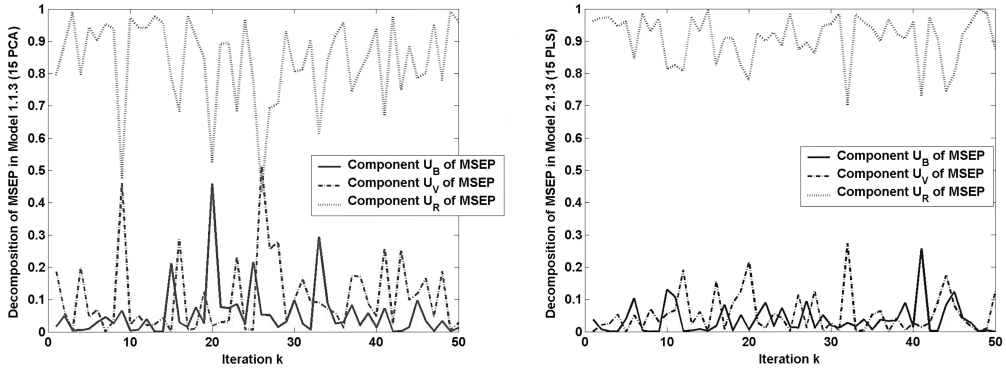


Figure 5: *Decomposition of $MSEP_i$ –$i = 1, 2, \ldots, 50$–, for Model 1.1.3 (15 PCA) and Model 2.1.3 (15 PLS).*

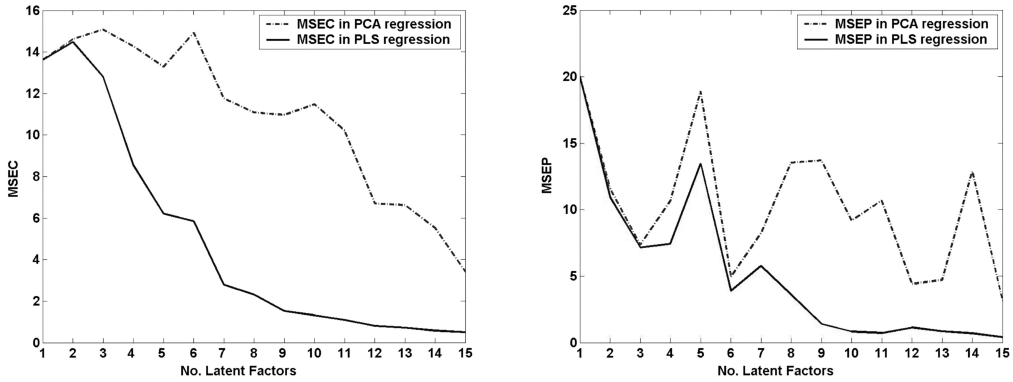


Figure 6: MSEC *and* MSEP *for PCA and PLS regression as a function of the number of latent factors.*

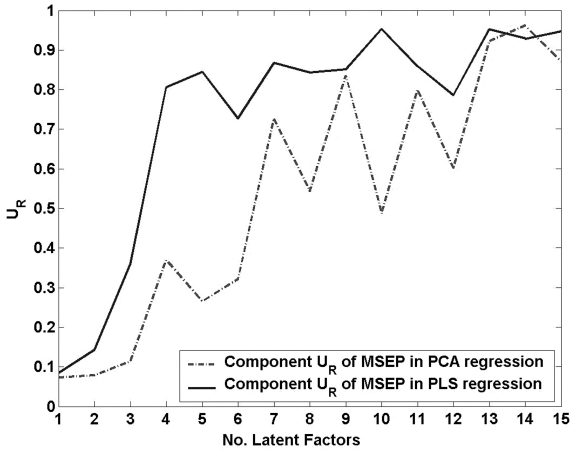


Figure 7: *Component* U_R *of* MSEP *in PCA and PLS regression as a function of the number of latent factors.*

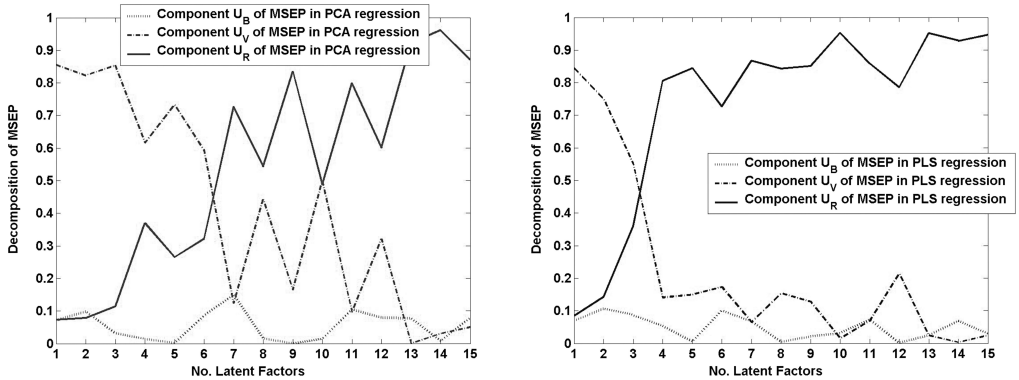


Figure 8: *Decomposition of* MSEP *in PCA and PLS regression as a function of the number of latent factors.*

Finally, the evolution of MSEC and MSEP (and their components U_B, U_V, U_R) is studied as a function of the number of latent factors or components (from 1 to 15) in PCA and PLS regression models, the learning and the test data set being changed for each number of latent factors. Figure 6 shows that obviously MSEC and MSEP tend to decrease with the inclusion of latent factors in the model. However, the decrease is more pronounced for PLS regression. Figure 7 illustrates that the component U_R, which is associated to the random variability in the prediction errors, increases with the number of latent factors. Nevertheless, U_R represents a percentage or MSEP higher in PLS regression than in PCA regression. Figure 8 shows that, on average, the component U_B of MSEP – which represents the systematic prediction errors – stay invariant with the inclusion of new latent factors in the regression models; U_V is higher when there are few components in the model and the random component U_R increases with the inclusion of latent factors in the model. The improvement is clearly higher for PLS regression than for PCA regression as evidenced by these figures.

5. Conclusions

This paper presents linear regression models explaining the oleic acid chemical composition of olive oil through factors extracted by principal components and partial least squares regression analyses from NIR spectral data. Relative maxima of the spectrum and cluster analysis are used to previously classify the explanatory variables. The different proposed models are compared on the basis of several criteria such as the number of latent factors or components, the goodness of fit and the mean squared error of prediction. The comparison among the models is improved by the consideration of some issues more commonly used in an economic context. More specifically, an exhaustive study about the multicollinearity is developed and a decomposition of MSEP is set up in order to analyze the nature of the prediction errors. In conclusion, the PLS regression model (of 15 latent factors) directly obtained from the data matrix (considered as a black box), applying neither additional information about the spectral peaks nor cluster analysis, is the best among all the considered models and exhibits optimal features on the basis of the diverse comparison criteria previously established. Besides, the decomposition of MSEP of this model points out the absence of systematic causes in the predictive errors, that are randomly distributed. Finally, cross-validation and bootstrapping allow us to confirm and generalize the previously obtained results, highlighting the potential of the PLS regression.

Acknowledgements

The authors thank the financial support by 'Junta de Andalucía' (Project P08-FQM-03931) and FEDER funds. Cooperativa Hojiblanca and Andalusian Protected Designations of

Origin (PDOs) are also gratefully acknowledged for providing the olive oil samples whereas the access to samples from germplasm bank of Córdoba is also thanked to IFAPA. Finally, the authors wish to thank Drs Rallo and Moalem for their kind scientific assistance.

References

Andersen, C. M. and Bro, R. (2010). Variable selection in regression-a tutorial. *Journal of Chemometrics*, 24, 728–737.

Anderson, M. (2009). A comparison of nine PLS1 algorithms. *Journal of Chemometrics*, bf 23, 518–529.

Baeten, V., Aparicio, R., Marigheto, N. and Wilson, R. (2003). *Manual del aceite de oliva*. AMV ediciones, Mundi-Prensa.

Barker, M. and Rayens, W. (2003). Partial least squares for discrimination. *Journal of Chemometrics*, 17, 166–173.

Berrueta, L. A., Alonso-Salces, R. M. and Héberger, K. (2007). Supervised pattern recognition in food analysis. *Journal of Chromatography A*, 1158, 196–214.

Burnham, K. P. and Anderson, D. R. (2004). Multimodel inference: understanding AIC and BIC in model selection. *Sociological Methods & Research* , 33, 261–304.

Climaco-Pinto, R., Barros, A. S., Locquet, N., Schmidtke, L. and Rutledge, D. N. (2009). Improving the detection of significant factors using ANOVA-PCA by selective reduction of residual variability. *Analytica Chimica Acta*, 653, 131–142.

Dupuy, N., Duponchel, L., Huvenne, J. P., Sombret, B. and Legrand, P. (1996). Classification of edible fats and oils by principal component analysis of Fourier transform infrared spectra. *Food Chemistry*, 57(2), 245–251.

Essi, I. D., Chukuigwe, E. C. and Ojekudo, N. A. (2011). On multicollinearity in nonlinear econometric models with mis-specified error terms in large samples. *Journal of Economics and International Finance* , 3(2), 116–120.

Frank, I. E. and Friedman, J. H. (1993). A statistical view of some chemometrics regression tools. *Technometrics*, 35(2), 109–135.

Gowen, A. A., Downewy, G., Esquerre, C. and O'Donnell, C. P. (2010). Preventing over-fitting in PLS calibration models of near-infrared (NIR) spectroscopy data using regression coefficients. *Journal of Chemometrics*, 25, 375–381.

Greenberg, E. and Parks, R. P. (1997). A predictive approach to model selection and multicollinearity. *Journal of Applied Econometrics*, 12, 67–75.

Guillén, M. D. and Ruiz, A. (2003). Rapid simultaneous determination by proton NMR of unsaturation and composition of acyl groups in vegetable oils. *European Journal of Lipid Science and Technology*, 105(11), 688–696.

Guldberg, A., Kaas, E., Déqué, M., Yang, S. and Vester Thorsen, S. (2005). Reduction of systematic errors by empirical model correction: impact on seasonal prediction skill. *Tellus*, 57(A), 575–588.

Gurdeniz, G. and Ozen, B. (2009). Detection of adulteration of extra-virgin oil by chemometric analysis of mid-infrared spectral data. *Food Chemistry*, 116, 519–525.

Kasemsumran, S., Kang, N., Christy, A. and Ozaki, Y. (2005). Partial least squares processing of near-infrared spectra for discrimination and quantification of adulterated olive oils. *Spectroscopy Letters*, 38(6), 839–851.

Li, B., Morris, J. and Martin, E. B. (2002). Model selection for partial least squares regression. *Chemometrics and Intelligent Laboratory Systems*, 64, 79–89.

López-Negrete de la Fuente, R., García-Muñoz, S. and Blegler, L. T. (2010). An efficient nonlinear programming strategy for PCA models with incomplete data sets. *Journal of Chemometrics*, 24, 301–311.

Mark, H. (1986). Comparative study of calibration methods for near-infrared reflectance analysis using a nested experimental design. *Analytical Chemistry*, 58, 2814–2819.

Mark, H. and Workman, J. (1986). Effect of repack on calibrations produced for near-infrared reflectance analysis. *Analytical Chemistry*, 58, 1454–1459.

Mevik, B. H. and Cerderkvist, H. R. (2004). Mean squared error of prediction (MSEP) estimates for principal component regression (PCR) and partial least squares regression (PLSR). *Journal of Chemometrics*, 18(9), 422–429.

Mynbaev, K. T. (2011). Regressions with asymptotically collinear regressors. *Econometrics Journal*, 14, 304–320.

Nelson, P. R. C., MacGregor, J. F. and Taylor, P. A. (2006). The impact of missing measurements on PCA and PLS prediction and monitoring applications. *Chemometrics and Intelligent Laboratory Systems*, 80, 1–12.

Öztürk, B., Yalçin, A. and Özdemir, D. (2010). Determination of olive oil adulteration with vegetable oils by near infrared spectroscopy coupled with multivariate calibration. *Journal of Near Infrared Spectroscopy*, 18, 191–201.

Reinaldo, F. T., Martins, J. P. A. and Ferreira, M. M. C. (2008). Sorting variables using informative vectors as a strategy for feature selection in multivariate regression. *Journal of Chemometrics*, 23, 32–48.

Spanos, A. and McGuirk, A. (2002). The problem of near-multicollinearity revisited: erratic vs systematic volatility. *Journal of Econometrics*, 108, 365–393.

Vasquez, V. R. and Whiting, W. B. (2006). Accounting for both random errors and systematic errors in uncertainty propagation analysis of computer models involving experimental measurements with Monte Carlo methods. *Risk Analysis*, 25(6), 1669–1680.

Yamagata, T. (2006). The small sample performance of the Wald test in the sample selection model under the multicollinearity problem. *Economics Letters*, 93, 75–81.

Yamamoto, H., Yamaji, H., Abe, Y., Harada, K., Waluyo, D., Fukusaki, E., Kondo, A., Ohno, H. and Fukuda, H. (2009). Dimensionality reduction for metabolome data using PCA, PLS, OPLS, and RFDA with differential penalties to latent variables. *Chemometrics and Intelligent Laboratory Systems*, 98, 136–142.

Zwanenburg, G., Hoefsloot, H. C. J., Westerhuis, J. A., Jansen, J. J. and Smilde, A. K. (2011). ANOVA-principal component analysis and ANOVA-simultaneous component analysis: a comparison. *Journal of Chemometrics*, 25(10), 561–567.

SORT 37 (1) January-June 2013, 81-94

Flexible quantile regression models: application to the study of the purple sea urchin

Isabel Martínez-Silva[1], Javier Roca-Pardiñas[2], Vicente Lustres-Pérez[3]
Altea Lorenzo-Arribas[1], Carmen Cadarso-Suárez[1]

Abstract

In many applications, it is often of interest to assess the possible relationships between covariates and quantiles of a response variable through a regression model. In some instances, the effects of continuous covariates on the outcome are highly nonlinear. Consequently, appropriate modelling has to take such flexible smooth effects into account. In this work, various flexible quantile regression techniques were reviewed and compared by simulation. Finally, all the techniques were used to construct the overall zone specific reference curves of morphologic measures of sea urchin *Paracentrotus lividus* (Lamarck, 1816) located in NW Spain.

MSC: 62G08, 62J02, 62P10.

Keywords: Boosting, quantile regression, reference curves, smoothness.

1. Introduction

Quantile regression is a statistical technique which allows, among other applications, to calculate growth curves and reference values, and is extremely useful in various fields of application, such as Ecology, Economy and Medicine, examples of which can be seen in Brian (2003), Koenker (2001), González-Barcala (2008), respectively. In the applied field, the need arises to extend the classic parametric approach by using smoothing techniques in regression to capture all the variations that occur in population quantile curves in response to a set of covariates.

[1] Departamento de Estatística e I. O., Universidade de Santiago de Compostela (USC), Spain.
isabelmaria.martinez@usc.es
[2] Departamento de Estatística e I. O., Universidade de Vigo (UVigo), Spain.
[3] Departamento de Zoología y Antropología Física, USC, Spain.
Received: December 2011
Accepted: November 2012

Figure 1: *Shown at left is a colony of P. lividus sea urchins from the intertidal site. Shown at right in orange are the gonads (the part that is marketed).*

Quantile regression is used in cases where a study seeks to estimate the different percentiles (e.g., the median) of a population of interest. One advantage of using quantile regression to estimate the median rather than using ordinary least squares regression (to estimate the mean), is that the former is less sensitive to the presence of atypical values. When it comes to using different measures of central trend and dispersion, quantile regression can be regarded as a natural analogue in regression analysis for ensuring a more complete and robust data analysis. A further advantage of this type of regression lies in the possibility of estimating any quantile and thus being able to ascertain what occurs in the case of extreme population values.

In practice there are different methodologies – with freeware implementations developed by the R Development Core Team (2011) – which address quantile regression. To our knowledge, while no general comparative analysis has targeted all of these methodologies, one such analysis has reportedly been conducted by Fenske (2011) on two of them.

Our principal aim was to conduct a comparative study, using simulation and application to real data, to carry out a brief review of a number of currently used flexible quantile regression techniques implemented in R software. Specifically, the following were reviewed: i) Koenker and Basset's methodology in Koenker (1978), using the `quantreg` package; ii) Cole (1988)'s least means squares (LMS) method, represented here in the form of a vector generalised additive model as proposed by (Yee (1996)), using the `VGAM` package; iii) the method based on generalised additive models for location, scale and shape proposed by Rigby (2001) and implemented in the `gamlss` package; and, iv) a new approach to quantile regression using the boosting process described by Buehlmann (2007), with the `mboost` package.

This study is structured as follows: Section 2.1 takes classic quantile regression and extends it to the non-parametric case; Section 2.2 outlines four current methods of non-parametric quantile regression; and Section 3 then makes a comparative study of the different techniques reviewed. The simulation study envisages a non-parametric

scenario that allows for the respective results yielded by the above-mentioned quantile regression techniques to be compared. Lastly, section 4 takes two of the four.

2. Quantile regression

2.1. Overview

Let $(x_1, y_1), \ldots, (x_n, y_n)$ be a random sample with variable response y and covariate x. The problem of parametric quantile regression is thus defined as

$$y_i = \beta_{0\tau} + \beta_{1\tau}\, x_i + \varepsilon_{i,\tau} \qquad \forall i \in \{1, \ldots, n\} \tag{1}$$

with $\beta_{0\tau}, \beta_{1\tau} \in R$ and $\varepsilon_{i,\tau} \sim H_\tau$ verifying $H_\tau(0) = \tau$. The estimated $\widehat{\beta}_{0\tau}$ and $\widehat{\beta}_{1\tau}$ are obtained by solving

$$(\widehat{\beta}_{0\tau}, \widehat{\beta}_{1\tau}) = \arg \min_{(\beta_{0\tau}, \beta_{1\tau}) \in R^2} \left\{ \sum_{y_i \geq A} \tau \, |y_i - \beta_{0\tau} - \beta_{1\tau}\, x_i| + \sum_{y_i < A} (1-\tau) \, |y_i - \beta_{0\tau} - \beta_{1\tau}\, x_i| \right\} \tag{2}$$

Due to the assumption of linearity in the covariate, the above model can be very restrictive in some instances. This constraint can be avoided by replacing the linear index $\beta_{0\tau} + \beta_{1\tau} \cdot x_i$ with a non-parametric structure. Accordingly, a generalisation of the model in (1) is given by

$$y_i = f_\tau(x_i) + \varepsilon_{i,\tau} \qquad \forall i \in \{1, \ldots, n\} \tag{3}$$

with f_τ being an unknown smooth function and $\tau \in (0,1)$. Moreover, the τ-th quantile of the error ε conditional on the covariate x is assumed to be zero, namely, $Q_\tau(\varepsilon_{i,\tau}|x) = 0$. Given the sample $(x_1, y_1), \ldots, (x_n, y_n)$ the estimation of f_τ is obtained by using some smoother of the form

$$\widehat{f}_\tau(x) = \sum_{i=1}^{n} \omega_{\lambda,\tau}(x_i)\, y_i \tag{4}$$

where λ is the smoothing parameter and $\omega_{\lambda,\tau}$ is the function of weights (kernel type, splines, etc.). Some of these methods are now reviewed below.

2.2. Methods reviewed

A number of techniques for calculating population growth curves are described in the current literature. Four techniques displaying different approaches and implemented in R software developed by the R Development Core Team (2011), are further discussed below.

2.2.1. Linear-programming-based technique

As its starting point, the linear-programming-based (LP-based) approach to the calculation of the quantile reference curve τ deems estimations of penalised quantile regression splines to be solutions to the minimisation of:

$$\sum_{i=1}^{n} \rho_\tau \{y_i - g(x_i)\} + \lambda \int \{g''(x)\}^2 dx \tag{5}$$

where $\rho_\tau(u) = u\{\tau - I(u < 0)\}$ is the function *check* proposed by Koenker (1978) and λ is the smoothing parameter of the resulting cubic spline, which generalises the classic approach of least squares smoothing splines pioneered by Wahba (1990). Since the minimisation problem posed entails a high computational cost, in expression (5) $\{g''(x)\}^2$ is usually replaced by $|g''(x)|$ (Koenker, 1994). Indeed, this is the approach used in the `quantreg` package. In our study, the `rqss` function was used to estimate the quantile curves, with smoothing being added in the non-parametric case via the `qss` function. No specifications were laid down as to the monotonicity of the data. This was due to the fact that, since the work scenarios encountered by us are not always monotonic, we felt this was something that should be borne in mind when it came to fitting the model.

2.2.2. Cole's least means squares method

In this case, the percentile reference curves are calculated on the basis of the distribution of the data. Hence, based on the LMS technique described in Cole (1988), the calculation of the τ-th percentile uses Box-Cox family power transformations λ to obtain the pertinent estimates for the mean and standard deviation. In this procedure, one obtains the τ-th percentile curve given by the equation

$$Q_{y_i}(\tau|x_i) = M(x_i)[1 + L(x_i)S(x_i)z_\tau]^{1/L(x_i)} \tag{6}$$

with z_τ being the normal equivalent deviate for tail area τ and $L(x), M(x)$ and $S(x)$ being functions that, as shown in (Cole, 1988), relate to the parameters λ, μ and σ of the distribution of the original simple data. These functions are estimated using vector generalised additive models (VGAM) proposed by Yee (1996) and based on smoothing

splines (Hastie, 1990). To implement this method, we used the VGAM-library `vgam` function.

2.2.3. Methodology of generalised linear models for location, scale and shape

The generalised linear models for location, scale and shape (GAMLSS) methodology proposed by Rigby (2005) assumes the structure

$$Q_{y_i}(\tau|x_i) = f_\tau(x_i) + \exp(g_\tau(x_i))z_\tau = \mu(x_i) + \sigma(x_i)z_\tau \tag{7}$$

with z_τ as being defined previously and where smoothing is introduced into the estimation of the data-distribution parameters, $\mu(x)$ and $\sigma(x)$, via the functions $f_{\tau j}$ and $g_{\tau j}$ using regression B-splines described in Boor (1978). Computational implementation was performed using `gamlss` belonging to the package of the same name. The resulting estimations, $\hat{\mu}$ and $\hat{\sigma}$, are based on B-Spline regression.

2.2.4. Boosting algorithms for quantile regression

Calculation of percentile curves based on boosting algorithms (BOOSTING) for quantile regression evolved from boosting algorithms for classification, the best known of which is the AdaBoost described in Freund (1997). Over the following two years, this algorithm was propounded by Breiman (1998, 1999), as a backward stepwise algorithm, known as the functional gradient descent FGD algorithm. Friedman, Hastie and Tibshirani (2000) and Friedman (2001) then carried out statistical developments which enabled the FGD algorithm to be applied to estimating functions, including regression. Subsequently Buehlmann (2007) developed boosting methods for estimation in quantile regression, and more recently, Fenske (2009) propounded the functional gradient boosting algorithm for additive quantile regression. In this approach, the τ-th percentile is given by

$$Q_{y_i}(\tau|x_i) = f_\tau(x_i) \tag{8}$$

where the non-linear term of equation (8) introduces smoothing function, f_τ, for continuous non-linear covariate x. In this paper, we fitted this model by means of smoothing P-splines with B-spline bases, using the `mboost` package `gamboost` function for the purpose.

3. Simulations

A simulation study was conducted to compare the behaviour of the different quantile regression techniques reviewed. To this end, samples were generated in accordance with the model

$$y = 2 + 1.5\log(x) + 0.5x\epsilon \tag{9}$$

with errors ϵ_i independently and identically distributed, and the covariate x was generated following a uniform distribution $U(0,3)$. One hundred $(m = 100)$ independent samples $\{(x_i, y_i)\}_{i=1}^{n}$ of size $n = 400$ were generated from the model (9) with independent random variables distributed following these different scenarios: **Scenario A** Normal standard distribution; **Scenario B** Student t distribution; and **Scenario C** Gamma distribution.

The mean squared error (MSE) and the mean absolute deviation error (MADE) were calculated for the quantile curves corresponding to $\tau \sim 0.3, 0.5, 0.7$. These errors are given by the equations (10) and (11) respectively,

$$MSE = \frac{1}{100}\sum_{j=1}^{100}(\widehat{Q}_{\tau}^{(j)}(x) - Q_{\tau}^{(j)}(x))^2 \tag{10}$$

$$MADE = \frac{1}{100}\sum_{j=1}^{100}|\widehat{Q}_{\tau}^{(j)}(x) - Q_{\tau}^{(j)}(x)| \tag{11}$$

where $\widehat{Q}_{\tau}^{(j)}(x)$ is the estimation of the $\tau-$th percentile for x_i, $Q_{\tau}^{(j)}(x)$ is the real value of the $\tau-$th percentile for x.

Table 1: *This table shows the mean (standard deviation) of the MSE and the MADE for the different methodologies and scenarios in the simulation sample.*

Scenario	τ		LP-based	LMS	GAMLSS	BOOSTING
A	0.3	MSE	0.115(0.133)	0.146(0.028)	0.143(0.017)	0.126(0.041)
		MADE	0.254(0.172)	0.204(0.067)	0.201(0.060)	0.241(0.068)
	0.5	MSE	0.110(0.092)	0.151(0.027)	0.171(0.028)	0.119(0.030)
		MADE	0.259(0.165)	0.184(0.061)	0.192(0.055)	0.224(0.056)
	0.7	MSE	0.134(0.115)	0.174(0.028)	0.171(0.031)	0.137(0.048)
		MADE	0.284(0.182)	0.188(0.060)	0.182(0.064)	0.240(0.062)
B	0.3	MSE	0.230(0.064)	0.359(0.043)	0.264(0.049)	0.188(0.069)
		MADE	0.296(0.078)	0.267(0.055)	0.276(0.065)	0.298(0.066)
	0.5	MSE	0.295(0.052)	0.169(0.038)	0.202(0.040)	0.144(0.057)
		MADE	0.256(0.064)	0.219(0.057)	0.235(0.050)	0.248(0.057)
	0.7	MSE	0.376(0.086)	0.227((0.050)	0.226(0.057)	0.193(0.078)
		MADE	0.288(0.077)	0.251(0.050)	0.239(0.064)	0.293(0.064)
C	0.3	MSE	0.191(0.036)	0.515(0.077)	0.771(0.069)	0.120(0.044)
		MADE	0.195(0.045)	0.581(0.080)	0.624(0.076)	0.222(0.060)
	0.5	MSE	0.390(0.086)	0.645(0.071)	0.458(0.084)	0.171(0.067)
		MADE	0.309(0.077)	0.588(0.088)	0.540(0.090)	0.285(0.066)
	0.7	MSE	0.524(0.120)	0.611(0.169)	0.690(0.157)	0.231(0.089)
		MADE	0.404(0.094)	0.981(0.139)	0.942(0.085)	0.340(0.081)

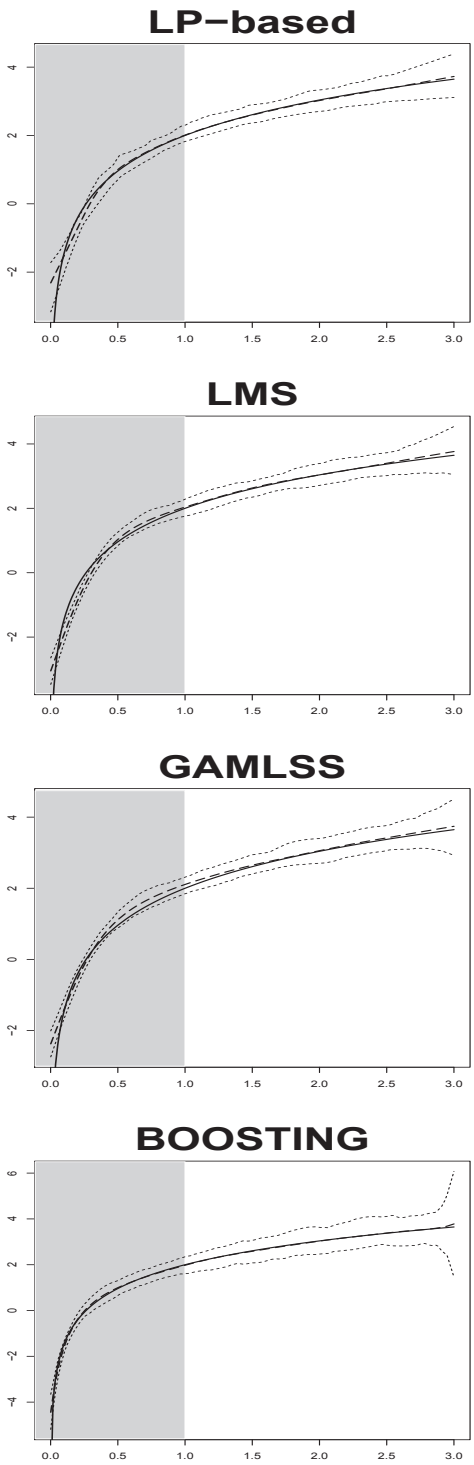


Figure 2: *The solid line shows the theoretical median curve and the dashed line shows the 95% simulation bands for the different techniques in the Scenario A.*

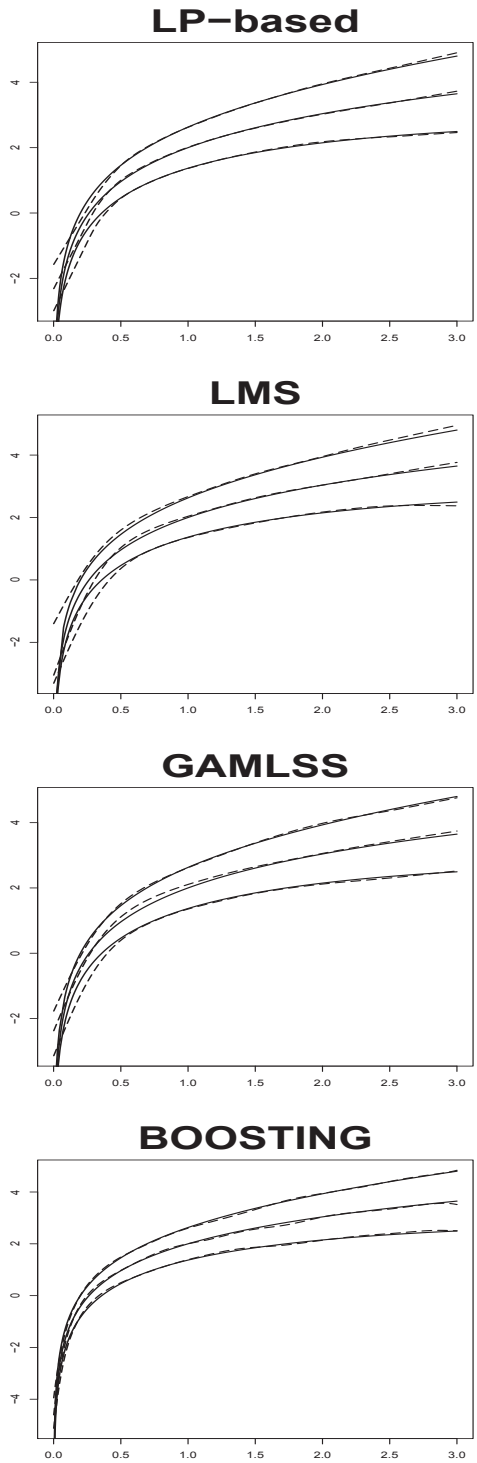


Figure 3: *The solid line shows the theoretical curve and the dashed line shows the fit for the quantiles* $\tau \sim 0.3, 0.5, 0.7$ *using the respective techniques in the Scenario A. Results are shown for the* 100^{th} *simulation.*

The results of this study are shown in Table 1 and similar results can be appreciated for all techniques. As can be seen with the MSE criterion: under Scenario A, the LP-based technique presents the lowest mean values and the BOOSTING technique shows the second lowest; for the Scenarios B and C, the BOOSTING technique presents the lowest mean values and the LP-based technique shows the highest mean values in Scenario B. Following the MADE criterion: in Scenarios A and B, the LP-based and BOOSTING techniques present the highest mean values; although under Scenario C the BOOSTING technique shows the lowest mean values in 5^{th} and 7^{th} percentile and the LP-based technique presents the second lowest, being the opposite in the 3^{rd} percentile. When comparing standard deviation for the MSE and the MADE values, can be appreciated that the LMS and GAMLSS technique shows the lowest values in Scenarios A and B although the LP-based and BOOSTING techniques present the lowest values in Scenario C.

As mentioned above, we have not seen a clear winner in Table 1. But when graphing this, a clear change has been noticed and we can see the improvement of working with the boosting methodology. In the graphical presentations, the 95% simulation bands for the median and the quantile curves corresponding to different values of τ are shown in Figures 2 and 3.

As can be seen from Figure 2, the inability of the LP-based, LMS and GAMLSS techniques to capture the variability of the data completely gave rise to problems in the simulation bands, and in the initial values of the covariate in particular.

When boosting algorithms were used, however, an improvement in the fit was observed across the entire scenario, with this being especially evident in the initial values referred to above. These characteristics can likewise be discerned in the calculation of the percentiles corresponding to $\tau \sim 0.3, 0.5, 0.7$ shown in Figure 3.

4. Application to the exploitation of marine resources

The study was undertaken at the following two sites along Galicia's Atlantic seaboard (NW Spain): Punta Area das Vacas (42°06′54″ N; 008°54′30″ W) (intertidal 1) situated on the Vigo estuary (*Ría de Vigo*); and Lago (42°19′25″ N; 008°49′37″ W) (intertidal 2) located on Aldán Bay (*Ensenada de Aldán*), at the southern edge of the Pontevedra estuary (*Ría de Pontevedra*). Both sites are representative of populations with a great abundance of *P. lividus* on the Galician coast.

Samples were collected from January 2002 to February 2003 along the lower intertidal zone of both sites (intertidal 1 and intertidal 2), and in the sublittoral area of Lago (site 2−sublittoral). The samples were randomly collected, with each comprising a total of 25 specimens of *P. lividus*. A total of 725 specimens were finally studied.The specimens were weighed and measured while fresh. The parameters considered for study purposes were the following two continuous variables: fresh weight, which is a good indicator of the commercial potential of sea urchins and was taken into account by

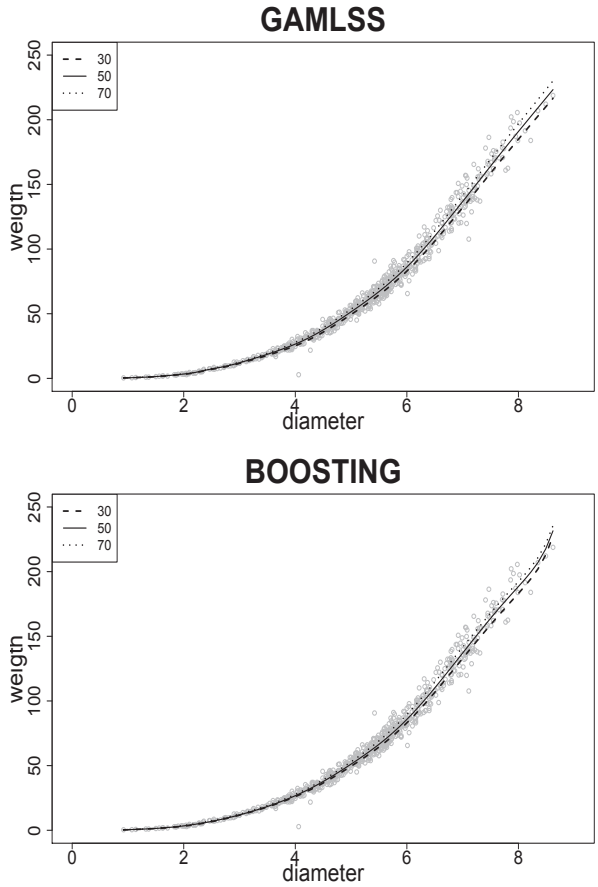


Figure 4: *Global population: depiction of the fits for the τ^{th} percentiles ($\tau \in \{30, 50, 70\}$) with the GAMLSS and BOOSTING techniques.*

Table 2: *This table shows the values of the estimates obtained at the global sample and at the various sites for different diameter (Diam.) values for the median ($\tau = 0.5$). These estimates were computed using the GAMLSS(G-T) and BOOSTING(B-T) techniques.*

	Global		intertidal 2		intertidal 2		sublittoral	
Diam.	G-T	B-T	G-T	B-T	G-T	B-T	G-T	B-T
2.0	3.42	3.46	3.49	3.40	2.90	2.66	3.83	3.68
2.5	7.48	7.05	7.24	7.03	7.03	6.14	7.60	6.74
4.0	26.55	26.75	27.17	27.00	27.38	26.32	32.94	28.10
5.5	67.10	66.47	65.93	66.84	66.74	66.04	71.76	69.11
6.0	86.13	87.75	81.52	85.79	83.22	83.94	91.36	87.55
8.0	191.96	189.43	—	—	—	—	182.15	192.37

being treated as a variable of interest; and diameter which, according to (Lustres-Pérez (2006)), is an indicator of size and strongly correlated with age, and was deemed to be a covariate in the model fitted.

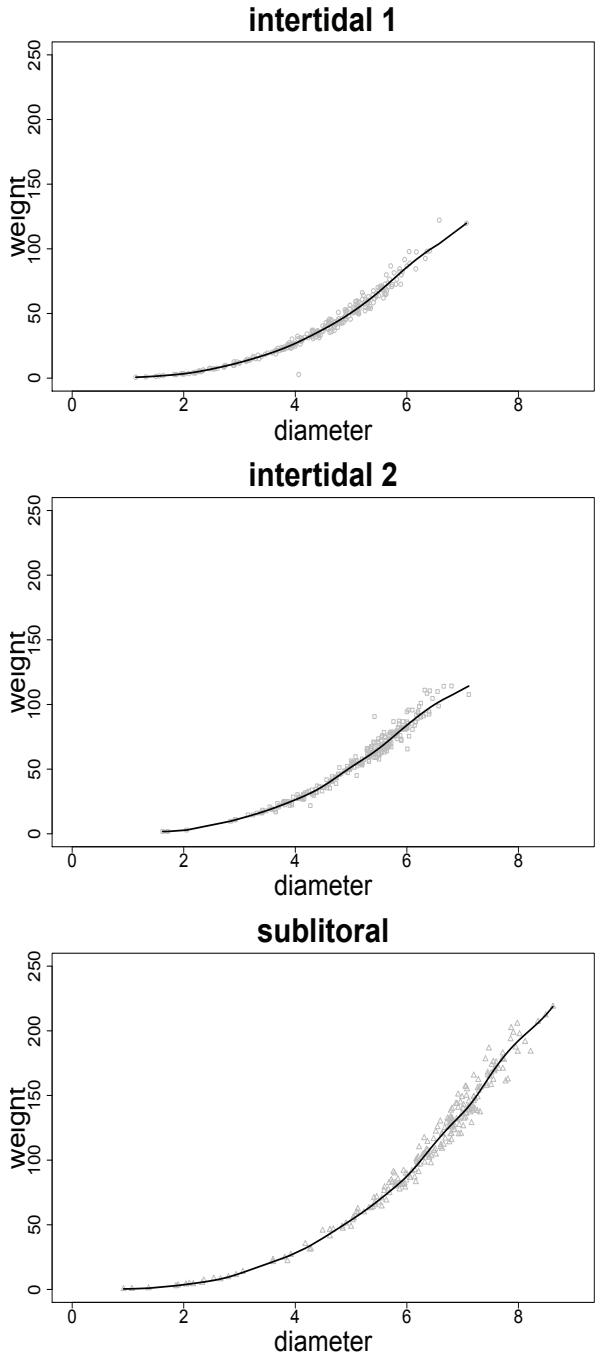


Figure 5: *Suitable zone by reference to resource exploitation: depiction of the fits for the $\tau^t h$ percentiles ($\tau \in \{30, 50, 70\}$) with the BOOSTING technique.*

In order to show the growth of the urchin population at different percentiles, two of the techniques applied to the global population studied are considered. Figure 4 shows

the weight change versus diameter for the τ^{th} percentile, ($\tau \sim 0.3, 0.5, 0.7$), to GAMLSS and BOOSTING techniques. The results are similar in both cases but present slight differences as can be seen in Table 2.

Since the sample was collected in three separate locations, the behaviour of previous percentiles in each of the zones has also been studied. In this case only one of the techniques studied, the boosting technique, has been used.

As can be seen in Figure 5 and in Table 3, our results showed that specimens of the sublittoral population displayed important differences with respect to those collected from the two intertidal populations. For any given size, sublittoral sea urchins were thus observed to register higher weights than those that inhabited the intertidal strip, across all the population quantile curves. These divergences increased from the point at which *P. lividus* attained the stipulated commercial size (diameter 5.5 cm). Furthermore the existence of a greater number of larger-sized specimens in the sublittoral population was also in evidence.

Table 3: *This table shows the values of the estimates obtained at the various sites for different diameter (Diam.) values and for three different percentiles (τ). These estimates were computed using the BOOSTING technique.*

	intertidal 1			intertidal 2			sublittoral		
Diam.	τ =0.3	τ =0.5	τ =0.7	τ =0.3	τ =0.5	τ =0.7	τ =0.3	τ =0.5	τ =0.7
2.0	3.15	3.40	3.67	2.67	2.66	2.68	3.30	3.68	3.80
2.5	6.68	7.03	7.43	6.09	6.14	6.08	6.48	6.74	8.15
4.0	26.37	27.00	27.80	25.18	26.32	27.61	26.16	28.10	29.45
5.5	64.45	66.84	68.88	64.08	66.04	68.28	67.48	69.11	73.01
6.0	82.12	85.79	87.06	81.46	83.94	86.60	84.79	87.55	90.86
8.0	–	–	–	–	–	–	186.64	192.37	196.73

5. Discussion

The results yielded by the simulation process suggest that the methods are competitive for fitting quantile regression models. Estimation of parameters and selection of variables cannot be made at a single stage of the estimation, nor can the degree of smoothing be selected automatically with the LP-based technique. The LMS method is likewise unable to select the smoothing parameter automatically, is very sensitive to data-dispersion and displays problems when it comes to working with negative-value responses. This last-mentioned aspect makes it necessary for translations to be made before and after fitting the model, to ensure that the results obtained can be properly assessed. As with the two previously described techniques, the GAMLSS methodology requires selection of the degree of smoothing. The boosting-based method is the one which (1) estimates the parameters, (2) selects the variables at a single stage of the estimation, and (3) implements automatic selection of the degree of smoothing. Furthermore, in the light of

the results shown in Figures 2 and 3, among the four methods discussed, the boosting-based method is the one for which the data best fits both small and large values of the covariate. The drawback of this last-mentioned methodology arises due to the fact that the percentile curves are calculated separately, and this leads to problems with the cross-tabulation of quantiles. With respect to application to real data, as Figure 5 and Table 3 show, there is a clear difference between the populations considered. The study confirms that sublittoral populations display conditions better suited to exploitation of *P. lividus*, due to:

- the existence of a greater number of commercial specimens; data corroborated in earlier studies undertaken on the Galician coast, such as those by Fernández-Pulpeiro (1999) and Lustres-Pérez (2006). In the latter case, a study of 206 intertidal and 63 sublittoral sites showed that the percentage of commercial sea urchins was 7% at the intertidal site and exceeded 50% at the sublittoral site; and,
- the greater development of sublittoral versus intertidal sea urchins, i.e., higher weights for any given diameter. This in turn means that during the harvesting periods on the Galician coast (from October to April), the quantity of gonads extracted from each sea urchin (the substance that is marketed) is appreciably higher.

Accordingly, we feel that it would be advisable for exploitation of *P. lividus* to be basically undertaken in the sublittoral area and always in a controlled manner. This would prevent the harvesting of a sizeable quantity of specimens with low commercial yields. Inappropriate extraction leads to a greater depletion of specimens, which limits the regeneration of populations of this echinoderm and, in turn, brings about a greater alteration in coastal ecosystems, bearing in mind the fundamental role that this species plays in the equilibrium of the habitats in which it lives (e.g. Benedetti-Cecchi, 1995; Kitching, 1961 and Ruitton, 2000).

Acknowledgements

The authors would like to express their gratitude for support received in the form of National Research Projects MTM2008-01603 and MTM2010-09213-E from the Spanish Ministry of Science and Innovation and the Galician Regional Authority Research Project INCITE08PXIB208113PR.

References

Benedetti-Cecchi, L. and Cinelli, F. (1995). Habitat heterogeneity, sea urchin grazing and the distribution of algae in litoral rock pools on the west coast of italy (western mediterranean). *Marine Ecology Progress Series*, 126, 203–212.

Breiman, L. (1998). Arcing classifiers (with discussion). *Annals of Statistics*, 26, 801–849.

Breiman, L. (1999). Prediction games and arcing algorithms. *Neural Computation*, 11, 1493–1517.

Brian, S., Cade, B. and Noon, R. (2003). A gentle introduction to quantile regression for ecologists. *Frontiers in Ecology and the Environment*, 1, 412–420.

Buehlmann, P. and Hothorn, T. (2007). Boosting algorithms: regularization, prediction and model fitting. *Statistical Science*, 22, 477–505.

Cole, T. J. (1988). Using the lms method to measure skewness in the nchs and dutch national height standards. *Annals of Human Biology*, 16, 407–419.

de Boor, C. (1978). *A Practical Guide to Splines*. Springer.

Fenske, N., Kneib, T. and Hothorn, T. (2009). *Identifying risk factors for severe childhood malnutrition by boosting additive quantile regression*. Technical report, Department of Statistics University of Munich.

Fenske, N., Kneib, T. and Hothorn, T. (2011). Childhood malnutrition by boosting additive quantile regression. *Journal of the American Statistical Association*, 106, 494–510.

Fernández Pulpeiro, E., César Aldariz, J., Lustres-Pérez, V. and Ojea Bouzo, C. (1999). *Ordenación integral del espacio marítimo-terrestre de Galicia. Fauna asociada a sustratos rocosos*. Informe final, 1996-1999. Technical report, Consellería de Pesca, Marisqueo e Acuicultura-Xunta de Galicia.

Freund, Y. and Schapire, R. (1997). A decision-theoretic generalization of on-line learning and an application to boosting. *Journal of Computer and System Sciences*, 55, 119–139.

Friedman, J. (2001). Greedy function approximation: a gradient boosting machine. *Annals of Statistics*, 29, 1189–1232.

Friedman, J., Hastie, T. and Tibshirani, R. (2000). Additive logistic regression: a statistical view of boosting (with discussion). *Annals of Statistics*, 28, 379–407.

González Barcala, F. J., Cadarso-Suárez, C., Valdés Cuadrado, L., Leis Trabazo, R., Cabanas, R. and Tojo Sierra, R. (2008). Lung function reference values in children and adolescents aged 6 to 18 years in Galicia. *Arch Bronconeumology*, 44, 295–302.

Hastie, T. J. and Tibshirani, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, London.

Kitching, J. and Ebling, F. (1961). The ecology os lough ine. xi. the control of algae by paracentrotus lividus (echinoidea). *Journal of Animal Ecology*, 30, 373–383.

Klein, B. and Korsholm, L., editor (2001). *The GAMLSS Project: a Flexible Approach to Statistical Modelling*.

Koenker, R. W. and D'Orey, V. (1994). Computing regression quantiles. *Applied Statistics*, 43, 410–414.

Koenker, R. and Hallock, K. F. (2001). Quantile regression. *Journal of Economic Perspectives*, 15, 143–156.

Koenker, R. W. and Bassett, G. W. (1978). Regression quantiles. *Econometrica*, 46, 33–50.

Lustres-Pérez, V. (2006). *El erizo de mar:* Paracentrotus lividus *(Lamarck, 1816) en las costas de Galicia*. PhD thesis. Universidad de Santiago de Compostela.

R Development Core Team (2011). *R: A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna, Austria. ISBN 3-900051-07-0.

Rigby, R. A. and Stasinopoulos, D. M. (2005). Generalized additive models for location, scale and shape, (with discussion). *Applied Statistics*, 54, 507–554.

Ruitton, S., Francour, P. and Boudouresque, C. (2000). Relationships between algae, benthic herbivorous invertebrates and fishes in rocky sublittoral communities of a temperature sea (mediterranean). *Estuarine, coastal and Shelf Science*, 50, 217–230.

Wahba, G. (1990). *Spline Models for Observational Data*. SIAM, Philadelphia.

Yee, T. W. and Wild, C. J. (1996). Vector generalized additive models. *Journal of Royal Statistical Society-Series B*, 58(3), 481–493.

SORT 37 (1) January-June 2013, 95-106

Modelling "calçots" (*Allium cepa* L.) growth by the Gompertz function

Joan Simó, Marçal Plans, Francesc Casañas and Jose Sabaté*

Abstract

"Calçots" are the second-year resprouts of the "Ceba Blanca Tardana de Lleida" landrace of onions. The evolution of three "calçots" populations has been modeled to help farmers to plan the optimal time to harvest. Four different models that essentially differ in the type of distribution of the fitting Gompertz function parameters (lag time, maximum growth rate and the maximum attainable number of commercial size "calçots") have been tested. The model that considers a multinomial distribution of the fitting parameters showed the best agreement with the experimental data.

MSC: MSC2000 classification of the American Mathematical Society: 92D25

Keywords: "Calçots", onion, crop modeling, Gompertz, simulation.

1. Introduction

"Calçots" are the second-year resprouts of the "Ceba Blanca Tardana de Lleida" landrace of onions. In the production of "calçots", all the resprouts from an onion are harvested at the same time, when ≥ 50 % reach commercial size (1.7 cm–2.5 cm in diameter and 20 cm in length, as specified in the regulations for the "Calçot de Valls" (Protected Geographical Indication). Each onion yields between 1 and 20 "calçots", but their thickness is negatively correlated with the number of "calçots" per onion, so in the most productive onions many "calçots" never fulfill the commercial requirements for size. Production lasts from mid-November to the end of April, and a more or less constant release of marketable product is needed during this period. Farmers exploit genetic variability in earliness, using combinations of genotypes and/or sowing dates to adjust the production to consumer demand, but these combinations are haphazard and

* Dept. d'Enginyeria Agroalimentària i Biotecnologia, UPC, ESAB, Campus del Baix Llobregat, Carrer Esteve Terradas 8, 08860 Castelldefels, Barcelona. joan.simo@upc.edu
Received: March 2012
Accepted: October 2012

inefficient. Thus, it would be interesting to develop a methodology that would enable farmers to predict crop evolution and help them plan the optimal time to harvest.

The data recorded in our experimental crops suggest that the evolution of the number of commercial "calçots" through the growing season can be described by a sigmoidal pattern: an initial period where no commercial "calçots" are observed is followed by a second stage of rapid expansion and a final phase where the number of "calçots" asymptotically tends to a maximum value. Sigmoidal curves have been widely reported in biology, particularly in the growth of microorganisms under specific physical and chemical conditions (Rodriguez-Gonzalez et al., 2011; Zwietering et al., 1990), in the microorganisms inactivation (Gil et al., 2011), in the seasonal growth of fish (Singhi, 2011), and in the growth and development of field crops (Barker et al., 2010; Tei, Aikman and Scaife, 1996).

One of the most popular models to explain sigmoidal curves is based on the Gompertz function, which can be expressed in several forms, such as the following one, which involves three biologically meaningful parameters.

$$N = N_m e^{-e^{\frac{\mu_m e}{N_m}((\lambda - t)+1)}}, \tag{1}$$

where N and t are the measured number of individuals and time, respectively. N_m is the maximum N that can be reached at infinite time, μ_m is the maximum growth rate, and λ is the lag time.

The Gompertz model for microorganisms' growth has been used to predict the safety and shelf life of foods (Gil et al., 2011; Rodriguez-Gonzalez et al., 2011; Zwietering et al., 1990). The Gompertz function has been fitted to lettuce growth, although onions and red beets have expolinear growth (Tei et al., 1996). The Gompertz model has also been fitted to herbage mass and herbage accumulation (Barker et al., 2010); growth of tobacco leaves, stems, and roots (Ismail, Khamis and Jaafar, 2003); total biomass, leaf area index, number of plants per meter, and productivity of sugarcane (Simoes, Rocha and Lamparelli, 2005); and dry matter production and cob weight in maize cultivars (Ramachandra Prasad, Krishnamurthy and Kailasam, 1992).

We aim to i) use parameters based on the Gompertz function to discriminate among different populations and ii) use the same parameters to model the growth of "calçots" to enable the evolution of the crop to be predicted and the optimal harvest time to be planned.

2. Materials and methods

In a single location, we cultivated 100 onions from three populations ($P1$, $P2$, and $P3$) corresponding to three different genotypes. We scored the number of commercial "calçots" (N) in each viable plant every two weeks over a seven-month period.

First, the three Gompertz function parameters ($\lambda_{i,j}$, $\mu_{i,j}$ and $N_{m,i,j}$) were estimated for any plant i of the population j, by nonlinear least squares using the Gauss-Newton

algorithm. The goodness of fit was tested for each plant through the R^2, the root mean square error (RMSE) and the residuals distribution. All these data were used to compare the behaviour of the three populations and to develop a simulation process aimed at predicting the evolution of the three crops.

2.1. Comparison of the three populations

Multivariate ANOVA of the Gompertz parameters was performed following $\mathbf{Y} = \mathbf{XB} + \mathbf{E}$ where $\mathbf{Y}$ is the parameters matrix distributed as a $N_p \sim (\boldsymbol{\mu}, \boldsymbol{\Sigma})$, $\mathbf{X}$ is the design matrix, $\mathbf{B}$ is the unknown parameters matrix (μ_{ip}, α_{ip}), and $\mathbf{E}$ is the error matrix. The distribution of error matrix was supposed $N_p \sim (0, \boldsymbol{\Sigma}\varepsilon)$, Wilks' statistic was used to test the significance of MANOVA. Comparisons of several multivariate means were analyzed using simultaneous confidence intervals (*CI*) with Bonferroni correction (Johnson and Witchern, 2007; Chung et al., 2009).

2.2. Simulations

Simulations were performed independently for each genotype. Viable plants of each genotype (n_j) were randomly split in 2 groups: the "calibration" and the "validation" groups, composed of n_c, j and $n_{v,j}$ plants respectively, being $n_{c,j} \approx 2n_{v,j}$ and $n_j = n_{c,j} + n_{v,j}$. Gompertz parameters achieved from the calibration set were used to generate a simulated set of $n_{v,j}$ plants.

The simulated set was generated according to four different models that essentially differ in the type of distribution of the fitting parameters of the Gompertz function. Model 1 only considers the average value of λ, μ_m and N_m. Then all the simulated plants evolved with the same rate for this model. Model 2 takes the average value of λ, μ_m and N_m and their variance-covariance matrix into account considering a normal multivariate distribution (Ripley, 1987) following the Equation (2):

$$f_{\mathbf{x}}(x_1, \ldots, x_k) = \frac{1}{(2\pi)^{\frac{k}{2}} |\boldsymbol{\Sigma}|^{\frac{1}{2}}} \Big(-\frac{1}{2} (\mathbf{x} - \boldsymbol{\mu})^{\mathsf{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu}) \Big), \tag{2}$$

where k is the numbers of random variables, $\boldsymbol{\Sigma}$ is the Variance-Covariance matrix between variables and $\boldsymbol{\mu}$ is the mean vector of these variables. Model 3 is similar to Model 2 but uses the transformed parameters obtained from the Box-Cox method (Box and Cox, 1964; Ripley, 1987). Model 4 considers a univariate Weibull distribution of λ, μ_m and N_m independently (Johnson et al., 1994). The probability density function of a random variable x is described by Equation (3):

$$f(x; k, \lambda) = \begin{cases} \frac{k}{\lambda} \left(\frac{x}{\lambda} \right)^{k-1} e^{-(\frac{x}{\lambda})^k} & \text{if } x \geq 0, \\ 0 & \text{if } x > 0, \end{cases} \tag{3}$$

where k is the shape parameter and λ is the scale parameter. Then, the shape and scale parameters were obtained by maximum likelihood estimation and were subsequently used to generate the simulated set (Johnson et al., 1994).

For each model and each genotype, the simulation was repeated 100 times from the first step (i.e., from the random selection of the calibration and validation sets). The suitability of the simulations for predicting the evolution of the crop was evaluated by comparing each simulated set with its corresponding validation set. Comparisons were carried out in three ways: first, by applying the chi-square test on the total number of commercial "calçots" of any population, N_t along the 14 scoring dates, second with a parametric model survival analysis in which the target success was the time (t_X) when a plant produces a given fraction (X) of "calçots" meeting the commercial specifications, and third using a one-way ANOVA performed on the maximum number of commercial "calçots", N_m in which scores at the latest time (count 14) were taken for the N_m of validation plants.

All calculations were done with the R-program (`www.R-project.org`), using packages `agricolae` (Mendiburu, 2010), `doBy` (Højsgaard and Halekoh, 2011), `fitdistrplus` (Delignette-Muller et al., 2010), `CAR` (Fox and Weisberg, 2011), `MASS` (Venables and Ripley, 2002) and `survival` (Therneau and Lumley, 2011).

3. Results

The evolution of commercial "calçots" number for a typical plant of any population is shown in Figure 1. The experimental vs. fitted values of $N_{i,j}$ for each plant with the modified Gompertz equation yielded average R^2 values of 0.901, 0.915, and 0.906, with their corresponding standard deviations of 0.078, 0.051, and 0.066 for the populations $P1$, $P2$ and $P3$, respectively. The average values of RMSE were 0.41, 0.39 and 0.22, which can be considered low compared with the number of "calçots" expected to

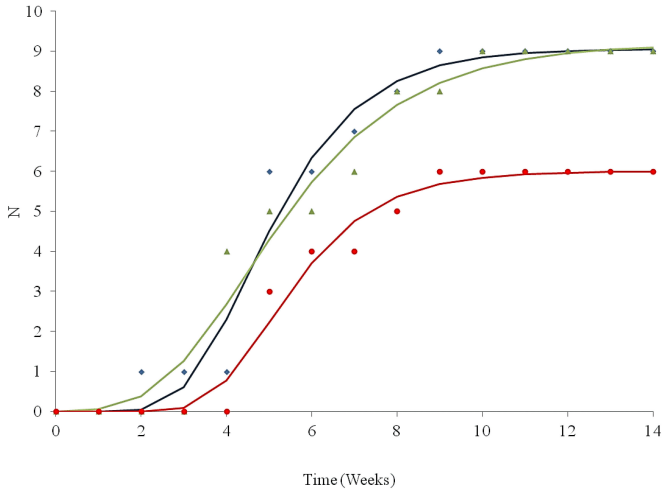


Figure 1: *Evolution of commercial "calçots" number for some typical plants:* ◇ *P1,* △ *P2 and* ○ *P3.*

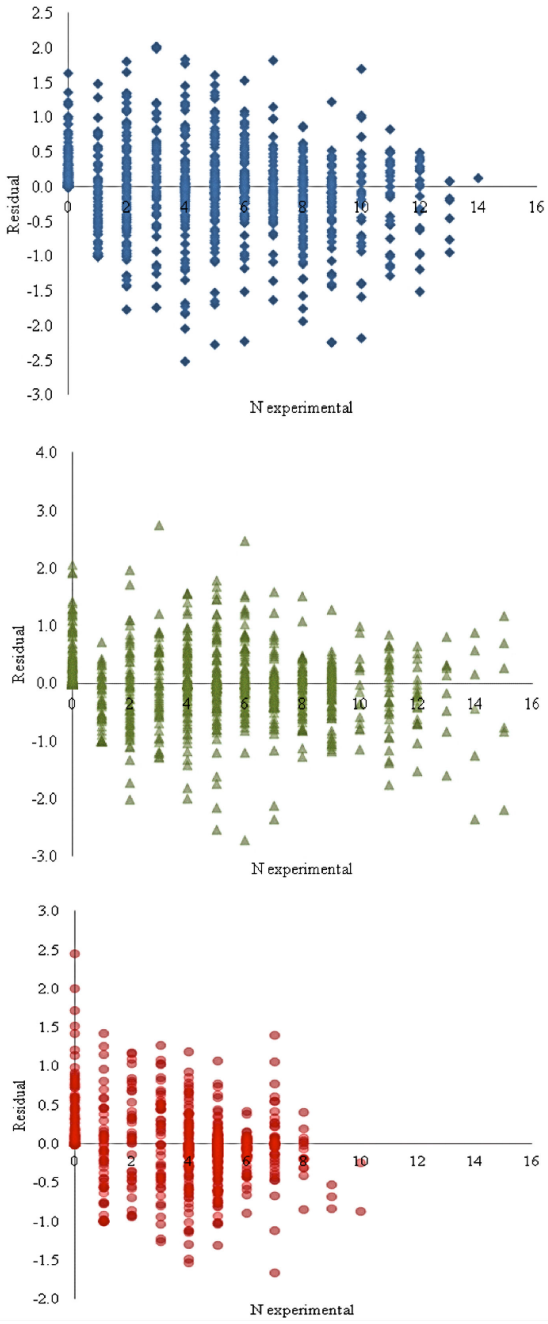


Figure 2: *Residuals vs experimental number of commercial "calçots". a)* ◇ *P1, b)* △ *P2 and c)* ○ *P3.*

harvest, 6 to 9. The residuals for of all the plants of any population, plotted in Figure 2, are nearly symmetrically distributed around the X axis for all the values of N. The only exception to this trend is observed for $N = 0$, as the predicted value of N should always

be positive according to Equation (1). All these data together support the suitability of the fitting.

3.1. Comparison of the three populations

First of all, the Shapiro-Wilks test shows that the set of estimates values of λ, μ_m and N_m of the individual plants for each population do not mainly follow a normal distribution, so they were transformed using the Box-Cox method and checked again (Box and Cox, 1964; Royston, 1982). Although conversion by the Box-Cox method resulted in transformed λ and μ_m values with a normal distribution, the transformed N_m ($P = 0.033$) still does not pass the test (Table 1). The Box-Cox parameter values for λ, μ_m and N_m were 0.7, -0.2, and -0.4, respectively. The p-values using Wilks tests were < 0.001 for both raw and transformed parameters. MANOVA with Bonferroni simultaneous *CI* test for multiple means comparisons (Table 2) indicates that the λ_2 mean value is smaller than λ_1 and λ_3 and μ_{m3} is smaller than μ_{m1} and μ_{m2} for both raw and transformed data. When the raw parameters are compared, the N_{m2} mean value is larger than N_{m1} and N_{m3}; however, when the transformed parameters are compared the N_m mean values of the three populations show significant differences.

3.2. Simulations

Mean parameters values and their coefficients of variation for simulation and variation sets can be seen in Table 3. Although a rigorous comparison was not performed, it can be seen that, generally, mean parameter values for the simulation sets are close to those corresponding to the validation sets. Further, the coefficients of variation indicate that fitting parameters are more scattered for simulation sets than for validation samples.

Table 1: *P-values for Shapiro-Wilks test of normality.*

	Raw parameters	Transformed parameters
λ	6.4×10^{-3}	0.215
μ_m	6.89×10^{-16}	0.146
N_m	2.75×10^{-13}	0.033

Table 2: *Comparison of parameters mean values with simultaneous CI for treatments' difference using Bonferroni correction in multiple comparison.*

	λ		μ_m		N_m	
	Raw	Transformed	Raw	Transformed	Raw	Transformed
$P1$	6.25 a	6.09 a	2.78 a	2.07 a	7.89 b	7.19 b
$P2$	4.84 b	4.66 b	2.51 a	1.93 a	9.40 a	8.25 a
$P3$	5.99 a	5.75 a	2.21 b	1.74 b	6.91 b	6.40 c

Table 3: *Mean values and coefficients of variation (CV) for simulation and validation sets.*

Set	Factor	Population	Model 1		Model 2		Model 3		Model 4	
			Mean	CV (%)	Mean	CV (%)	Mean	CV (%)	Mean	CV (%)
Simulation	λ	$P1$	6.51	9.37	6.48	8.90	6.43	8.58	6.42	9.92
		$P2$	5.13	9.17	5.17	8.71	4.69	10.56	4.78	10.17
		$P3$	6.21	7.22	6.28	8.48	5.90	10.87	5.91	11.37
	μ_m	$P1$	3.20	14.75	3.16	13.99	2.89	20.56	2.90	14.06
		$P2$	2.69	10.56	2.69	10.17	2.89	19.27	2.60	13.18
		$P3$	2.37	12.53	2.46	12.93	2.25	16.54	2.24	14.19
	N_m	$P1$	8.48	7.92	8.44	6.93	6.94	5.57	8.42	7.77
		$P2$	10.19	7.14	9.99	7.76	7.35	5.68	10.32	8.00
		$P3$	7.47	6.25	7.52	6.27	6.53	5.83	7.53	7.04
Validation	λ	$P1$	6.25	7.28	6.21	8.23	6.18	6.14	6.18	7.06
		$P2$	4.83	6.40	4.80	6.88	4.88	7.38	4.80	7.51
		$P3$	5.96	8.33	5.98	6.94	6.05	7.20	5.95	7.75
	μ_m	$P1$	2.80	13.76	2.76	13.33	2.79	13.97	2.72	13.76
		$P2$	2.55	10.97	2.49	10.66	2.55	11.12	2.48	11.48
		$P3$	2.24	11.79	2.21	11.64	2.20	12.82	2.21	10.68
	N_m	$P1$	7.87	6.79	7.89	5.69	7.97	6.73	7.95	5.89
		$P2$	9.38	6.55	9.52	6.60	9.37	6.75	9.39	7.68
		$P3$	6.94	6.34	6.84	5.39	6.86	6.01	6.90	6.15

That can be understood because fitting parameters of validation set are estimated in one step. Nevertheless, the fitting parameters of simulations sets are estimated after two previous processes, the parameters estimation of calibration samples and the generation of a validation set with a limited number of samples. Additionally, for simulation sets, the parameters of Model 3 tend to be slightly lower than those of other models. Although the mean parameter values of the validation sets will not be used in future computations, comparing them to the corresponding simulation sets gives a first rough view of the goodness of the simulation.

Table 4: *Percentage of simulations that pass the chi-square test without reaching significance.*

	Model 1	Model 2	Model 3	Model 4
$P1$	0	28	32	22
$P2$	0	29	23	16
$P3$	0	27	27	18
All	0	5	10	10

First, the chi-square test was performed on the total number of commercial "calçots" in a given population. The percentage of simulations that accomplish the chi-square without significance ($p \leq 0.01$) is very low: about 20% for Model 4 and about 30% for Models 2 and 3 (Table 4). No simulations of Model 1 pass the test, which indicates that a

given distribution of λ, μ_m and N_m, and not only their mean values, has to be considered in order to make reasonable predictions of the evolution of the crop.

To analyze the reason for the failures of the chi-square test, the residuals between the predicted and experimental values of N vs. time were plotted (Figure 3). As expected, Model 1 gave the highest residual values throughout the growing season. Models 2 and 3 show a similar tendency for any population and two smaller peaks are observed at counts 4 and 9. Model 4 follows a particular trend for any population, peaks are observed at count 4 for population 1 but peaks at counts 4 and 8 appear for population 3.

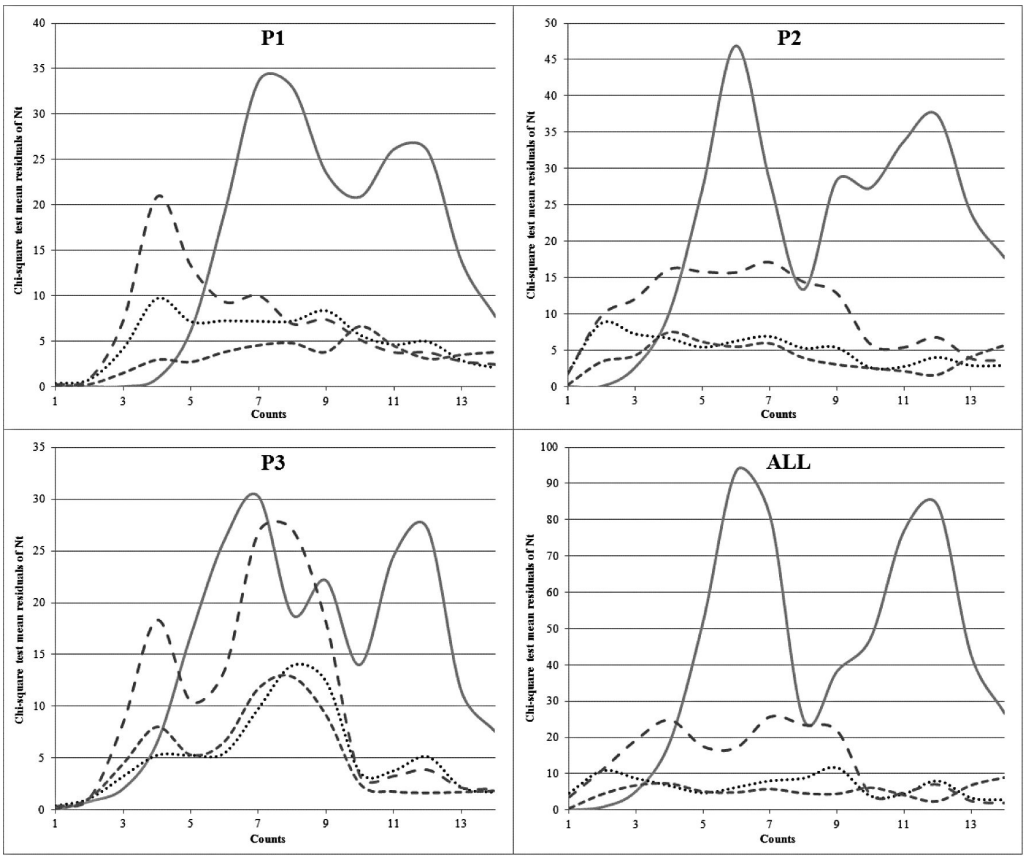


Figure 3: *Chi-square test mean residuals of N_t between validation and simulation sets for 100 simulations.*

As a second method to check the suitability of the simulations, a survival analysis was performed at the four times when 25%, 50%, 75%, and 90% of the "calçots" of one plant achieved the commercial size. In agreement with the results of the chi-square test, Model 1 fails for the three populations at all times. The other models behave differently depending on the percentage considered. When 25% of the "calçots" of a plant attained the commercial size (t_{25}), the number of successful simulations was visibly lower than

those corresponding to t_{50}, t_{75}, and t_{90}, for which more than 90% of the simulations were successful (Table 5).

Table 5: *Percentage of simulations that fulfill the survival analysis without reaching significance ($p \leq 0.01$).*

	t_{25}				t_{50}				t_{75}				t_{90}			
	*P*1	*P*2	*P*3	ALL	*P*1	*P*2	*P*3	ALL	*P*1	*P*2	*P*3	ALL	*P*1	*P*2	*P*3	ALL
Model 1	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0	0
Model 2	54	86	69	92	90	94	96	98	100	95	98	100	83	97	94	98
Model 3	73	86	87	95	99	94	97	100	99	97	97	99	99	98	100	100
Model 4	70	84	74	96	94	93	97	98	96	98	98	100	76	93	93	99

The third way to evaluate the simulation was a one-way ANOVA performed on the maximum number of commercial "calçots", N_m. Again, the results showed that model 1 did not work (Table 6). Models 2 and 4 lead to more than 90% of simulations with the ANOVA test non-significant, whereas the suitability of Model 3 is clearly lower and varies greatly depending on the population.

Table 6: *Percentage of simulations that fulfill the ANOVA for N_m.*

	Model 1	Model 2	Model 3	Model 4
*P*1	4	99	32	100
*P*2	55	100	86	98
*P*3	17	94	65	99
All	55	100	93	97

A global comparison of the different models can be seen in Figure 4, where the total number of commercial "calçots" of any population, N, is represented. Points for the hundred simulations of any count are included. As stated above, the points of Model 1 are farthest from the target line, where experimental and calculated values of N would match. Models 2 and 4 tend to overestimate the production of "calçots". For the last stages of the culture, when N_t approaches its maximum value, the points of Model 3 move away from the bisector, in agreement with the ANOVA test for N_m. Globally, the best predictions were achieved by Model 2.

4. Conclusions

The modified Gompertz function allowed us to compare several populations with different genotypes throughout the growing season instead of making the comparison at peak times when scoring is carried out.

The multinomial distribution of fitting parameters of the Gompertz function used in Model 2 was the best distribution to model the growth of "calçots", predict the evolution of the crop, and decide the optimal harvest time.

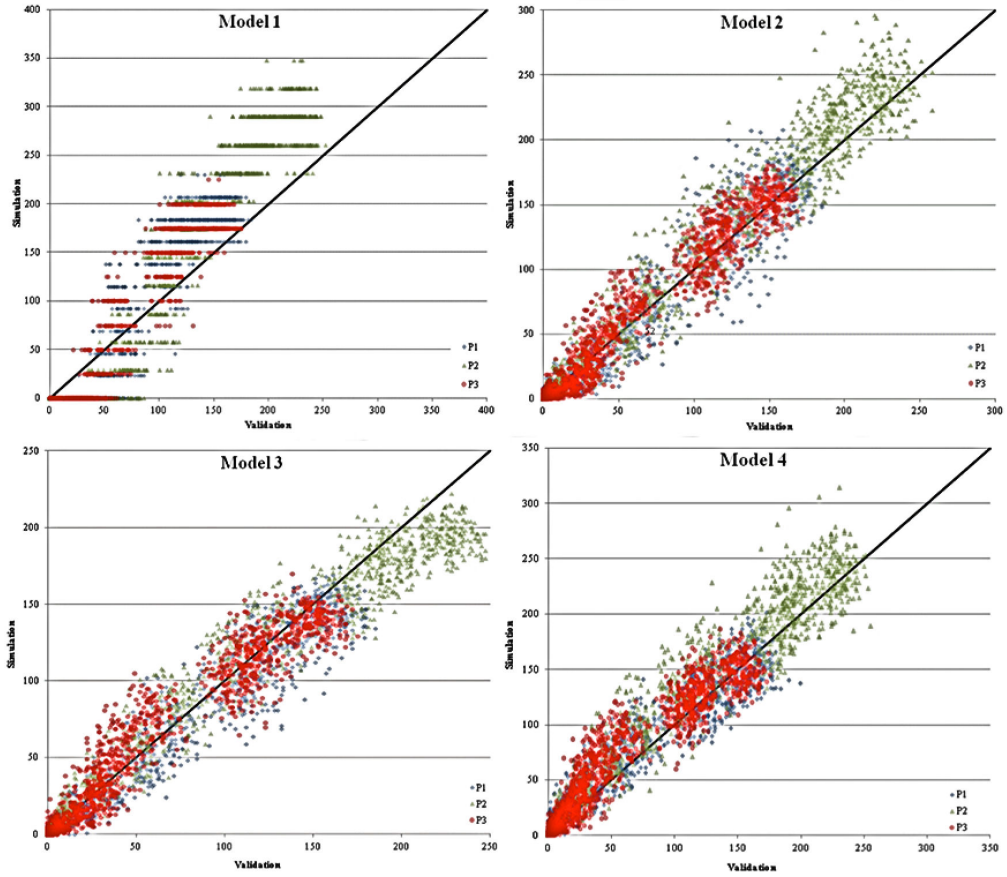


Figure 4: *Simulation vs. validation of total number of commercial "calçots". Symbols:* ◇ *P1,* △ *P2 and* ○ *P3.*

Model 1 yielded the worst results for all the tests used. Thus, models that consider a given distribution of the Gompertz fitting parameters (λ, μ_m and N_m) are much more suitable to explain the growth of "calçots" than those that consider only mean values.

Studies of the effects of the environment and genotypes on crop growth are needed to understand the different behaviour of each population so that better models can be constructed for the entire growing season.

References

Barker, D. J., Ferraro, F. P., Nave, R. L., Sulc, R. M., Lopes, F. and Albrecht, K. A. (2010). Analysis of herbage mass and herbage accumulation rate using Gompertz equations. *Agronomy Journal*, 102, 849–857.

Box, G. E. P. and Cox, D. R. (1964) An analysis of transformations. *Journal of the Royal Statistical Society, Series B*, 26, 211–252.

Chung, C., Du, H., Su, Y. and Zhang, X. (2009). *The R project for Comparisons of Several Multivariate Means*. In http://www.divms.uiowa.edu/ kcowles/s166_2009/ ChungS166projectreport.pdf (September 2012).

Delignette-Muller, M. L., Pouillot, R., Denis J.-B. and Dutang C. (2010). fitdistrplus: help to fit of a parametric distribution to non-censored or censored data. http://www.R-project.org

Fox, J., and Weisberg, S. (2011). *An R Companion to Applied Regression*. *SAGE Publications*, Thousand Oaks, Calif.

Gil, M. M., Miller, F. A., Brandao, T. R. S. and Silva, C. L. M. (2011). On the use of the Gompertz model to predict microbial thermal inactivation under isothermal and non-isothermal conditions. *Food Engineering Reviews*, 3, 17–25.

Højsgaard, S. and Halekoh, U. (2011). doBy: groupwise summary statistics, general linear contrasts, LSMEANS (least-squares-means), and other utilities. http://www.R-project.org

Ismail, Z., Khamis, A. and Jaafar, Y. (2003). Fitting nonlinear Gompertz curve to tobacco growth data. *Pakistan Journal of Agronomy*, 2, 223–236.

Johnson, R. A. and Wichern, D. W. (2007). *Applied Multivariate Statistical Analysis (6th Ed.)*. Englewood Cliffs, New Jersey: Prentice-Hall. Chapter 6.

Johnson, N. L., Kotz, S. and Balakrishnan, N. (1994). *Continuous Univariate Distributions*, volume 1, chapter 21. Wiley, New York.

Mendiburu, F. (2010). agricolæ: statistical procedures for agricultural research. http://www.R-project.org

Royston, P. (1982). An extension of Shapiro and Wilk's W test for normality to large samples. *Applied Statistics*, 31, 115–124.

Ramachandra Prasad, T. V., Krishnamurthy, K. and Kailasam, C. (1992). Functional crop and Cob growth models of maize (*Zea mays L.*) cultivars. *Journal of Agronomy and Crop Science*, 168, 208–212.

Ripley, B. D. (1987). *Stochastic Simulation.* Wiley. Page 98.

Rodriguez-Gonzalez, O., Walkling-Ribeiro, M., Jayaram, S. and Griffiths, M. W. (2011). Cross-protective effects of temperature, pH, and osmotic and starvation stresses in *Escherichia coli* O157:H7 subjected to pulsed electric fields in milk. *International Dairy Journal*, 21, 953–962.

Simoes, M. d. S., Rocha, J. V. and Lamparelli, R. A. C. (2005). Growth indices and productivity in sugarcane. *Scientia Agricola*, 62, 23–30.

Singhi, N. O., Paul, A. K., Singh, N. G., Singh, P. and Alam, W. (2011). Modeling seasonal growth of fish using modified Gompertz model with sine wave function. *Indian Journal of Animal Sciences*, 81(6), 648–650.

Tei, F., Aikman, D. P. and Scaife, A. (1996). Growth of lettuce, onion and red beet. 2. Growth modelling. *Annals of Botany*, 78, 645–652.

Therneau, T. and Lumley, T. (2011). survival: survival analysis, including penalised likelihood. http://www.R-project.org

Venables, W. N. and Ripley, B. D. (2002). *Modern Applied Statistics with S.* Springer, New York.

Zwietering, M. H., Jongenburger, I., Rombouts, F. M. and Vantriet, K. (1990). Modeling of the bacterial-growth curve. *Applied and Environmental Microbiology*, 56, 1875–1881.