Exploratory Data Analysis techniques for detection of possible outlying subjects in designs of bioavailability studies

Cornelia Enăchescu[*] and Denis Enăchescu[**]
[*]Institute for Mathematical Statistics and Applied Mathematics, [**]University of Bucharest

ABSTRACT

One of the problems commonly encountered in bioavailability studies is that the data set may contain some extremely large or small (i.e. outlying) observations. These observations may have an influence on the conclusion of the bioequivalence.

This contribution is structured into three parts:

- the first section presents some new Exploratory Data Analysis (EDA) techniques, such Principal Component Analysis (PCA) and Projection Pursuit (PP), for detection of possible outlying subjects in designs of bioavailability studies;
- in the second section, a discordance test for one or more outlying observations for an individual subject, based on sample Kurtosis, is discussed;
- in section three, the AUC (Area Under the Curve) data (both raw data and log-transformed data) of the two erythromycin formulation in Clayton and Leslie's study are used to illustrate the procedures presented above.

Although computer intensive, the above EDA techniques are efficient in detecting outlying subjects and observations.

INTRODUCTION

FUNDAMENTAL BIOEQUIVALENCE ASSUMPTION

When two drug products are equivalent in the rate and extent to which the active drug ingredient or therapeutic moiety is absorbed and becomes available at the site of drug action, it is assumed that they will be therapeutically equivalent.

THE DEFINITION OF OUTLYING OBSERVATIONS

One of the problems commonly encountered in bioavailability studies is that the data set may contain some extremely large or small (i.e., outlying) observations. These observations may have an influence on the conclusion of bioequivalence. Basically, there are three different kinds of outliers:

1. Unexpected observations in the blood or plasma concentration-time curve
2. Extremely large or small observations within a given formulation
3. Unusual subjects who exhibit extremely high or low bioavailability relative to the reference formulation.

For the first kind of outlier, Chow and Liu (2000) indicated that unexpected observations in the plasma concentration-time curve usually have little effect on calculation of AUC and, consequently, have little effect on the comparison of bioavailability.

EXPLORATORY DATA ANALYSIS TECHNICS

ANDREWS CURVES

Andrews curves [Andrews, 1972] were developed as a method for visualizing multi-dimensional data by mapping each observation onto a function. This function is defined as

$$f_x(t) = x_1/\sqrt{2} + x_2 \sin t + x_3 \cos t + x_4 \sin 2t + x_5 \cos 2t + \dots ,$$

where the range of t is given by $-n < t < n$. Each observation is projected onto a set of orthogonal basis functions represented by *sines* and *cosines* and then plotted. Thus, each sample point is now represented by a curve.

It has been shown that because of the mathematical properties of the trigonometric functions, the Andrews curves preserve means, distance (up to a constant) and variances. One consequence of this is that Andrews curves showing functions close together suggest that the corresponding data points will also be close together. Thus, one use of Andrews curves is to look for clustering of the data points.

Andrews curves are dependent on the order of the variables. Lower frequency terms exert more influence on the shape of the curves, so re-ordering the variables and viewing the resulting plot might provide insights about the data. By lower frequency terms, we mean those that are first in the sum given in the above equation.
It has been suggest that the data be rescaled so they are centered at the origin and have covariance equal to the identity matrix. Andrews curves can be extended by using orthogonal bases other than sines and cosines (for example Legendre polynomials and Chebychev polynomials).

PRINCIPAL COMPONENTS ANALYSIS
The Andrews curves are attempts to visualize all of the data points and all of the dimensions at once; the curves accomplish this by mapping a data point to a curve. Another option is to tackle the problem of visualizing multi-dimensional data by reducing the data to a smaller dimension via a suitable projection. These methods reduce the data to 1-D or 2-D by projecting onto a line or a plane and then displaying each point in some suitable graphic, such as a scatterplot. Once the data are reduced to something that can be easily viewed, then exploring the data for patterns or interesting structure is possible.
One well-known method for reducing dimensionality is principal component analysis (PCA) [Enachescu D, 2003]. This method uses the eigenvector decomposition of the covariance (or the correlation) matrix. The data are then projected onto the eigenvector corresponding to the maximum eigenvalue (sometimes known as the first principal axis) to reduce the data to one dimension. In this case, the eigenvector is one that follows the direction of the maximum variation in the data. Therefore, if we project onto the first principal axis, then we will be using the direction that accounts for the maximum amount of variation using only one dimension.
We could project onto two dimensions using the eigenvectors corresponding to the largest and second largest eigenvalues. This would project onto the plane (named the principal factorial plane) spanned by these eigenvectors. As we see shortly, PCA can be thought of in terms of projection pursuit, where the interesting structure is the variance of the projected data.
For normally distributed observations projected onto the principal factorial plane, the square distance to 0 is a weighted sum of *d* independent χ_1^2 variables. Its expectation is $\sum_{i=1}^{d}\lambda_i = d$ and its variance is $2\sum_{i=1}^{d}\lambda_i^2$. Observations with a square distance greater than $d + 2\sqrt{2\sum_{i=1}^{d}\lambda_i^2}$ may be considered as *outliers*.[Enachescu and Enachescu, 2000]

PROJECTION PURSUIT
There are an infinite number of planes that we can use to reduce the dimensionality of our data. As we just mentioned, the first two principal axes in PCA span one such plane, providing a projection such that the variation in the projected data is maximized over all possible 2-D projections. However, this might not be the best plane for highlighting interesting and informative structure in the data. Structure is defined to be departure from normality and includes such things as clusters, linear structures, holes, outliers, etc. Thus, the objective is to find a projection plane that provides a 2-D view of our data such that the structure (or departure from normality) is maximized over all possible 2-D projections.
In the literature, the projection pursuit is described as a way of searching for and exploring nonlinear structure in multi-dimensional data by examining many 2-D projections. The idea is that 2-D orthogonal projections of the data should reveal structure that is in the original data. The projection pursuit technique can also be used to obtain 1-D projections, but we look only at the 2-D case. In our presentation of projection pursuit exploratory data analysis, we follow the method of Posse [1995a, 1995b].
Projection pursuit exploratory data analysis (PPEDA) is accomplished by visiting many projections to find an interesting one, where interesting is measured by an index. In most cases, our interest is in non-normality, so the projection pursuit index usually measures the departure from normality. The index we use is known as the chi-square index.
PPEDA consists of two parts:
1) a projection pursuit index that measures the degree of the structure (or departure from normality), and
2) a method for finding the projection that yields the highest value for the index.
Posse [1995a, 1995b] uses a random search to locate the global optimum of the projection index and combines it with the structure removal of Freidman (1987) to get a sequence of interesting 2-D

projections. Each projection found shows a structure that is less important (in terms of the projection index) than the previous one.

Projection Pursuit Index

Posse developed an index based on the chi-square. The plane is first divided into 48 regions or boxes B_k that are distributed in rings. See Figure for an illustration of how the plane is partitioned. All regions have the same angular width of 45 degrees and the inner regions have the same radial width of $(2\log 6)^{1/2}/5$. This choice for the radial width provides regions with approximately the same probability for the standard bivariate normal distribution. The regions are constructed in this way to account for the radial symmetry of the bivariate normal distribution.

Posse provides the population version of the projection index. We present only the empirical version here, because that is the one that must be implemented on the computer. The projection index is given by

$$PI_{\chi^2}(\alpha,\beta) = \frac{1}{9}\sum_{j=1}^{8}\sum_{k=1}^{48}\frac{1}{c_k}\left[\frac{1}{n}\sum_{i=1}^{n} I_{B_k}(z_i^{\alpha(\eta_j)}, z_i^{\beta(\eta_j)}) - c_k\right]^2$$

where

X is an *n x d* matrix, where each row $(\mathbf{X}_i)$ corresponds to a *d*-dimensional observation and *n* is the sample size. **Z** is the studentized version (i.e. 0 mean and 1 variance) of **X**.

$\hat{\mu}$ is the *1 x d* sample mean:

$$\hat{\mu} = \sum \mathbf{X}_i/n.$$

$\hat{\Sigma}$ is the sample covariance matrix:

$$\hat{\Sigma}_{ij} = \frac{1}{n-1}\sum(\mathbf{X}_i - \hat{\mu})(\mathbf{X}_j - \hat{\mu})^T.$$

α, β are orthonormal *d*-dimensional vectors that span the projection plane.

$P(\alpha,\beta)$ is the projection plane spanned by α and β

$z_i^{\alpha}, z_i^{\beta}$ are the studentized observations projected onto the vectors α and β

$$z_i^{\alpha} = z_i^T\alpha$$

$$z_i^{\beta} = z_i^T\beta$$

(α^*, β^*) denotes the plane where the index is maximum.

$PI_{\chi^2}(\alpha,\beta)$ denotes the chi-square projection index evaluated using the data projected onto the plane spanned by α and β

ϕ_2 is the standard bivariate normal density.

c_k is the probability evaluated over the *k*-th region using the standard bivariate normal,

$$c_k = \int\int_{B_k}\phi_2 dz_1 dz_2.$$

B_k is a box in the projection plane, I_{B_k} is the indicator function for region B_k

$\eta_j = \pi j/36, j = 0, \ldots, 8$ is the angle by which the data are rotated in the plane before being assigned to regions B_k

$\alpha(\eta_j)$ and $\beta(\eta_j)$ are given by

$$\alpha(\eta_j) = \alpha\cos\eta_j - \beta\sin\eta_j$$

$$\beta(\eta_j) = \alpha\sin\eta_j + \beta\cos\eta_j$$

The chi-square projection index is not affected by the presence of outliers. It is sensitive to distributions that have a hole in the core, and it will also yield projections that contain clusters. The chi-square projection pursuit index is fast and easy to compute, making it appropriate for large sample sizes.

Using a similar argument as in the PCA case, for normally distributed observations projected onto the $P(\alpha,\beta)$ plane, the points which fall outside circle 2 can be considered as outlying values.

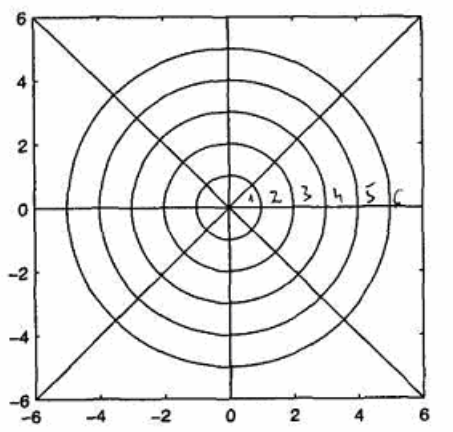


Figure 1 The layout of the regions B_k for the chi-square projection index

DISCORDANCY TESTS OF ONE OR MORE OUTLIERS (IRRESPECTIVE OF THEIR DIRECTIONS) IN A NORMAL SAMPLE WITH UNKNOWN MEAN AND VARIANCE

Test statistic [Barnett and Lewis, 1994]:

$$T = \text{sample kurtosis} = \frac{n\sum_{j=1}^{n}(x_j - \bar{x})^4}{\left[\sum_{j=1}^{n}(x_j - \bar{x})^2\right]^2}.$$

T is the locally best-unbiased invariant test of given size against a location-slippage alternative in which k of the n observations arises from separate normal distributions $N(\mu + a_1, \sigma^2), ., N(\mu + a_k, \sigma^2)$, where $a_1, ., a_k$ differ from zero but are otherwise arbitrary, provided that the contamination proportion *k/n* under the alternative hypothesis is less than 0.21. T is also the locally best invariant test of given size against a dispersion-slippage alternative in which k of the observations arises from separate normal distributions $N(\mu, b_1\sigma^2), ., N(\mu, b_k\sigma^2)$, $b_1 > 1, ., b_k > 1$ irrespective of the proportion *k/n*. Its power is nearly as good as that of two-sided discordancy test of an extreme outlier in a normal sample with unknown mean and variance against slippage in location for a single observation by medium or large amounts. Against slippage in location by two observations it is superior to the above mentioned test in power, greatly so when the sample size is less than, say, 20.

T has the advantage of being robust against possible masking effects. It is suitable for *consecutive* use in the possible presence of more than one outlier.

CONSECUTIVE ('RECURSIVE') TEST OF UP TO K OUTLIERS (IRRESPECTIVE OF DIRECTIONS) IN A NORMAL SAMPLE WITH μ AND σ^2 UNKNOWN

Test procedure: Suppose T is a test statistic for a (two-sided) discordancy test of a single outlier in a normal sample. Let $x_1, x_2, \ldots, x_k$ (for prescribed k) be the observations yielding the maximum value of T in subsamples $s_1, s_2, \ldots, s_k$ where s_j $(j = 1, 2, \ldots, k)$ is the set of observations excluding $x_1, x_2, \ldots, x_{j-1}$. (That is, x_1 produces the maximum value of T in the complete sample; x_2 produces the next largest value of T calculated for the sample of n - 1 observations on omission of x_1, and so on.) Suppose the successive values of T so obtained are $T_1, T_2, \ldots, T_k$. We determine $\lambda_i(\beta)$ where $P[T_i > \lambda_i(\beta)] = \beta$ for $i = 1, 2, \ldots, k$ and

$$P\left\{\bigcup_{i}^{k}[T_i > \lambda_i(\beta)]\right\} = \alpha$$

Then a level-α test operates as follows. If $T_k > \lambda_k(\beta)$ then $x_1, x_2, \ldots, x_k$ are discordant. Otherwise we proceed by examining T_l for $l = k - 1, k - 2, \ldots, 1$ until $T_l > \lambda_l(\beta)$, at which stage $x_1, x_2, \ldots, x_l$ are adjudged discordant at level α. (If $T_l \leq \lambda_l(\beta)$ for all $l = 1, 2, \ldots, k$, we conclude, of course, that there are no discordant outliers).

Properties of test: This procedure embody the estimation of the *number* of contaminants in the sample.

The policy of examining samples of successively reduced size in reverse order protects against masking effects in the more usual forms of consecutive test. The inconvenience is having to recalculate summary statistics at each of the *k* stages.

APPLICATION AND COMMENTS

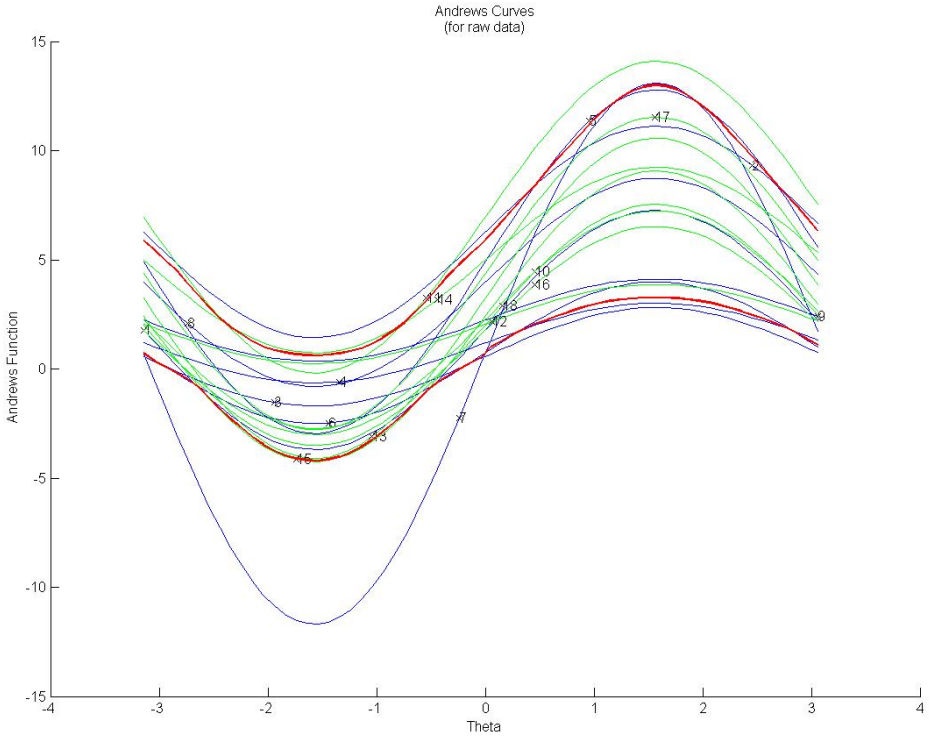
We will use AUC data (both raw data and log-transformed data) of the two erythromycin formulations in Clayton and Leslie's study (see Chow and Liu, 2000) to illustrate the above presented methodology. Note that this data set has been analyzed by many researchers because of its possible violation of the normality assumption in raw data and the existence of potential outliers.

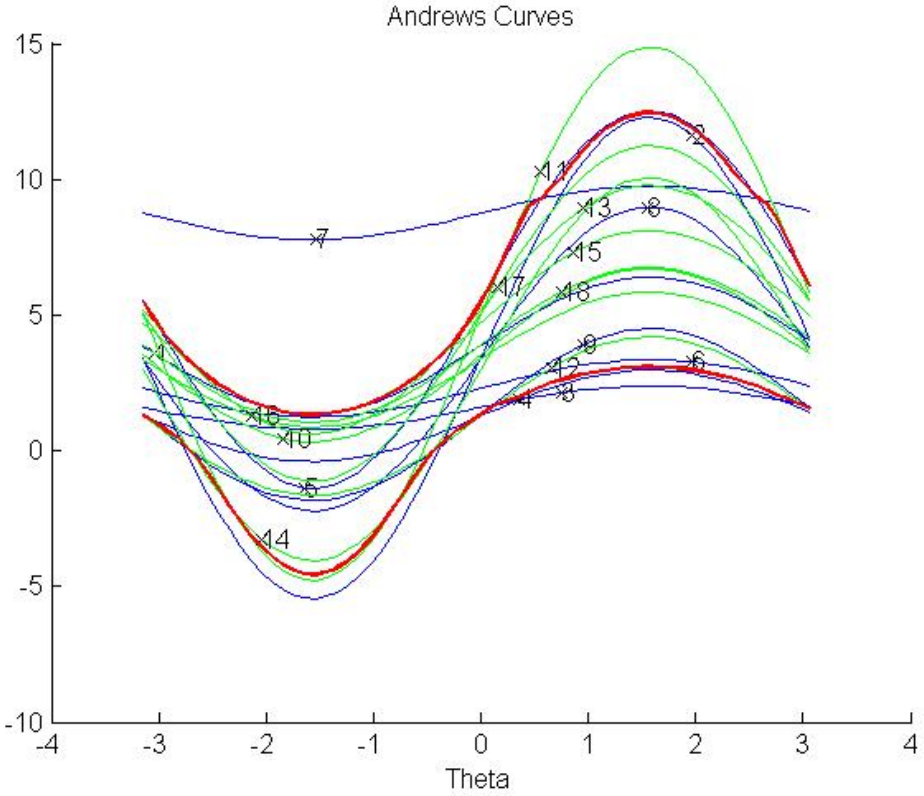
Table 1 AUCs for Two Erythomycin Formulations

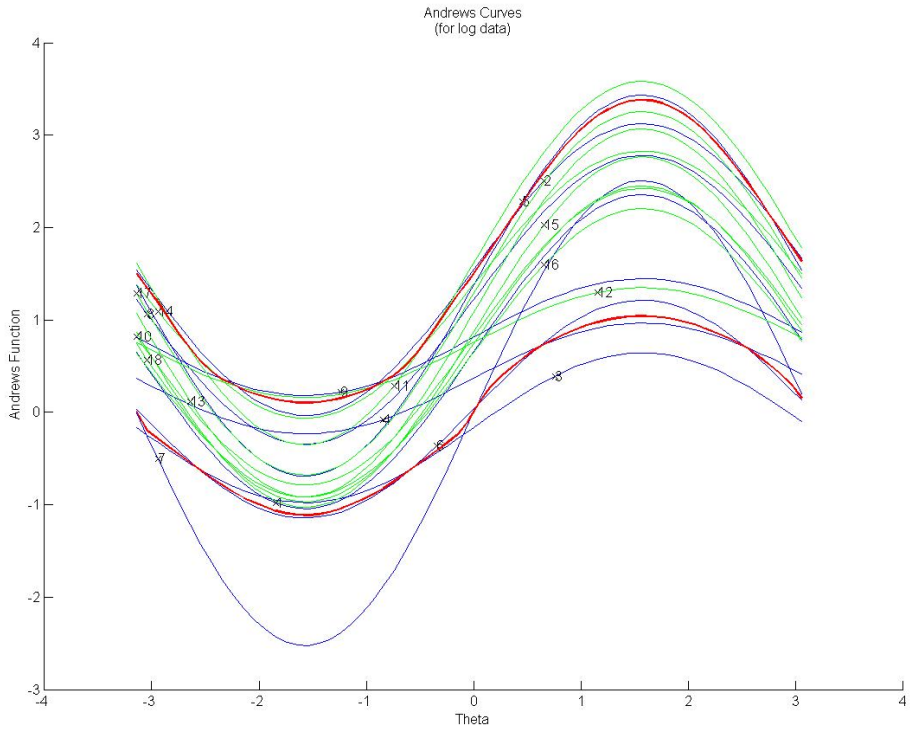
		C (Stearate)		D (Base)		Ratio C/D	
Subject	Sequence	Raw	Log (raw)	Raw	Log (raw)	Raw	Log (raw)
1	CD	2.52	0.9243	5.47	1.6993	0.4607	−0.7750
2	CD	8.87	2.1827	4.84	1.5769	1.8326	0.6058
3	CD	0.79	−0.2357	2.25	0.8109	0.3511	−1.0467
4	CD	1.68	0.5188	1.82	0.5988	0.9231	−0.0800
5	CD	6.95	1.9387	7.87	2.0631	0.8831	−0.1243
6	CD	1.05	0.0488	3.25	1.1787	0.3231	−1.1299
7	CD	0.99	−0.0101	12.39	2.5169	0.0800	−2.5269
8	CD	5.60	1.7228	4.77	1.5624	1.1740	0.1604
9	CD	3.16	1.1506	1.88	0.6313	1.6809	0.5193
10	DC	3.19	1.1600	4.98	1.6054	0.6406	−0.4454
11	DC	9.83	2.2854	7.14	1.9657	1.3768	0.3197
12	DC	2.91	1.0682	1.81	0.5933	1.6077	0.4748
13	DC	4.58	1.5217	7.34	1.9933	0.6240	−0.4716
14	DC	7.05	1.9530	4.25	1.4469	1.6588	0.5061
15	DC	3.41	1.2267	6.66	1.8961	0.5120	−0.6694
16	DC	2.49	0.9123	4.76	1.5603	0.5231	−0.6480
17	DC	6.18	1.8213	7.16	1.9685	0.8631	−0.1472
18	DC	2.85	1.0473	5.52	1.7084	0.5163	−0.6611

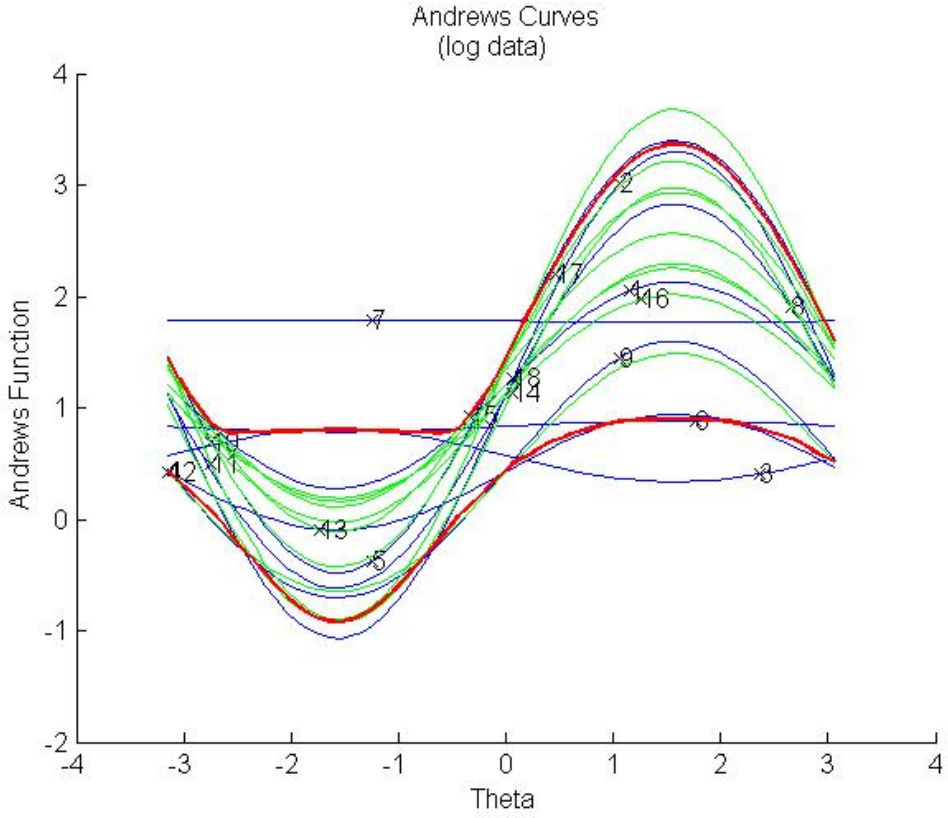
Source: Clayton and Leslie (1981).

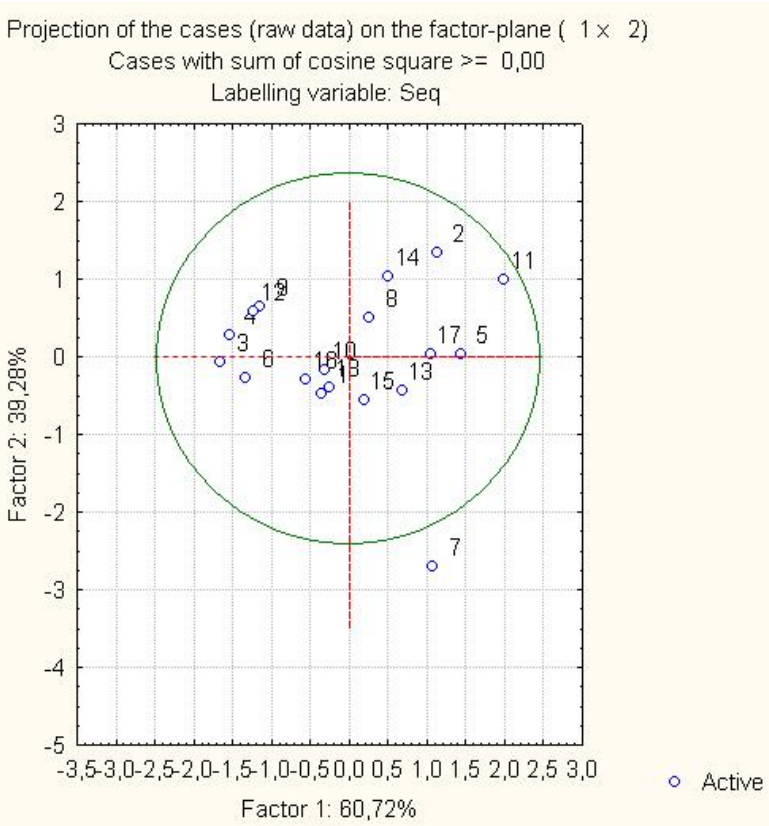
From a visual inspection of the table it can be seen that, in raw data, subjects 2 and 14 are possible upper outliers (i.e. extremly large observations) as subject 7 is a possible lower outlier.
Using the EDA technics to represent the data we obtain the following plots:

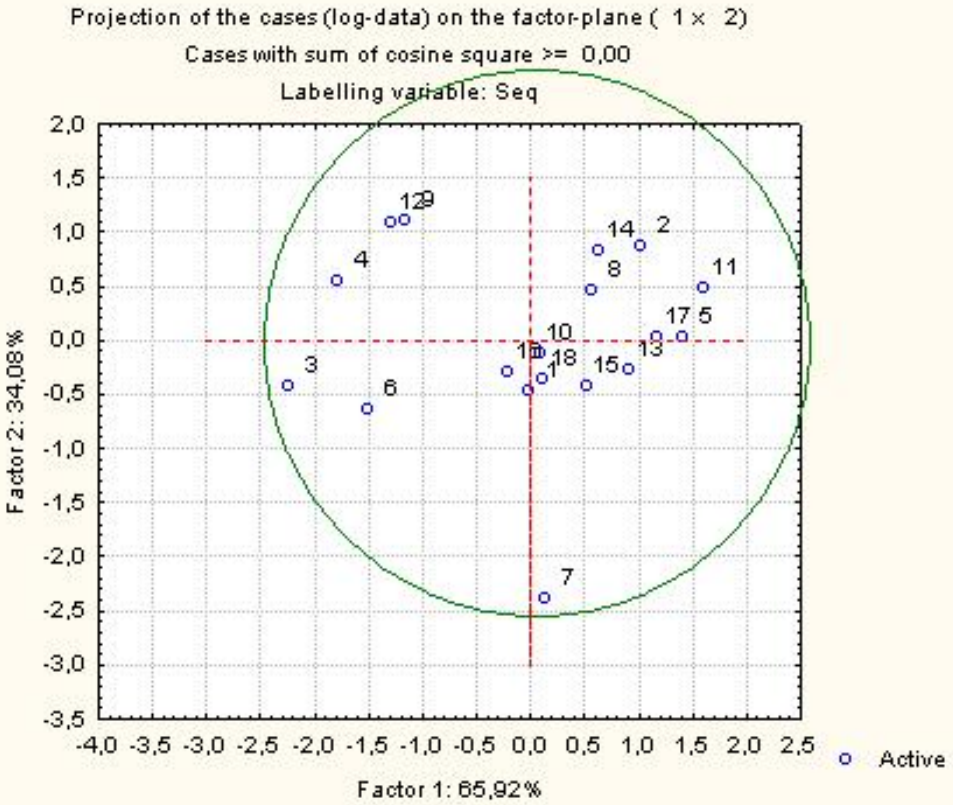

Andrews Curves
(for raw data)
Andrews Function
Theta
15
10
5
0
-5
-10
-15
-4
-3
-2
-1
0
1
2
3
4

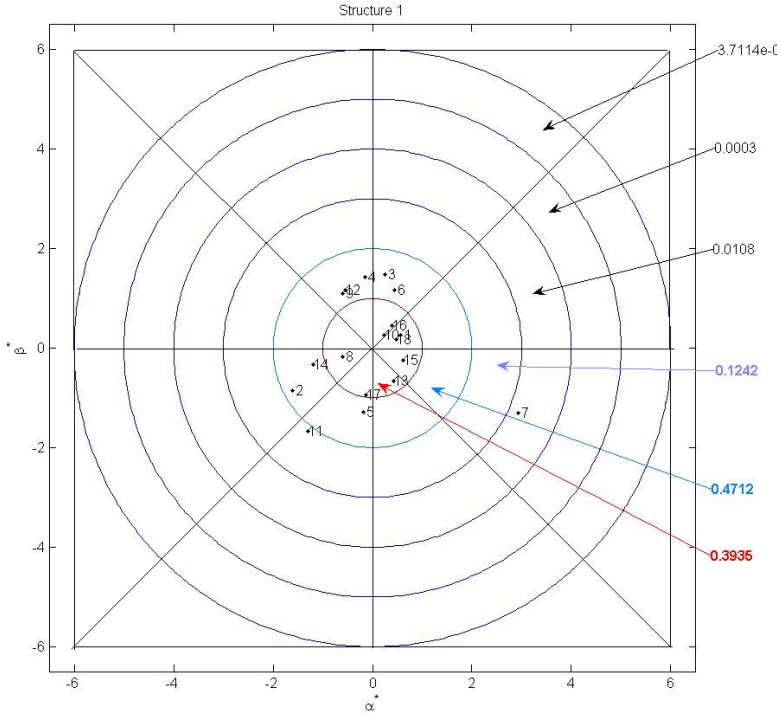

Andrews Curves
Theta
15
10
5
0
-5
-10
-4
-3
-2
-1
0
1
2
3
4

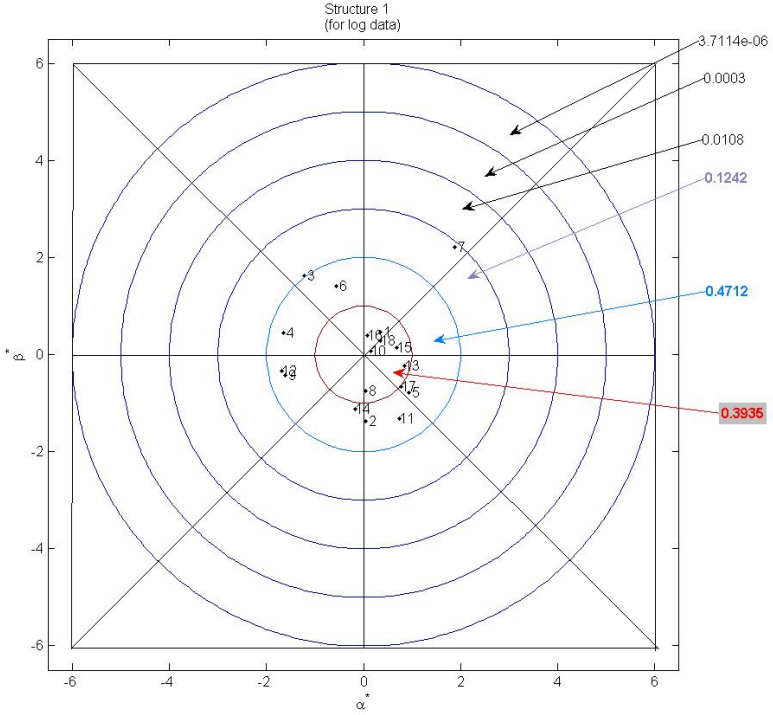

Andrews Curves
(for log data)
Andrews Function
Theta


Andrews Curves
(log data)
Andrews Function
Theta


Projection of the cases (raw data) on the factor-plane (1 × 2)
Cases with sum of cosine square >= 0,00
Labelling variable: Seq
Factor 2: 39,28%
Factor 1: 60,72%
Active


Projection of the cases (log-data) on the factor-plane (1 × 2)
Cases with sum of cosine square >= 0,00
Labelling variable: Seq
Factor 2: 34,08%
Factor 1: 65,92%
Active


Structure 1
3.7114e-0
0.0003
0.0108
0.1242
0.4712
0.3935
β*
α*


Structure 1
(for log data)
3.7114e-06
0.0003
0.0108
0.1242
0.4712
0.3935
β*
α*

The conclusion is obvious: all EDA technics confirm that subject 7 for raw and log-transformed data has outlying values.

In order to asses bioequivalence the U.S. Food and Drug Administration –FDA- recommend the following 2x2 crossover design:

$$Y_{ijk} = \mu + S_{ik} + F_{(j,k)} + C_{(j-1,k)} + P_j + e_{ijk},$$

where i = 1, 2, …, n_k; j = 1, 2; k = 1, 2 and the following normality assumptions:

1. $\{S_{ik}\}$ are i.i.d. normal with mean 0 and variance σ_s^2.
2. $\{e_{ijk}\}$ are i.i.d. normal with mean 0 and variance σ_e^2.
3. $\{S_{ik}\}$ and $\{e_{ijk}\}$ are mutually independent. (1.2)

The validation of the above assumptions has an influence on the assessement of bioequivalence. Consider the intrasubject residual for subject *i* within sequence *k* during period *j*, denoted by $\hat{e}_{ijk}$ and defined as the difference between the observed reponse Y_{ijk} and its predicted value $\hat{Y}_{ijk}$. Similarly, the intersubject residuals can be used to evaluate the normality assumption imposed on the intersubject variability of S_{ik}. The intersubject residuals, denoted by $\hat{S}_{ik}$, are given as:

$$\hat{S}_{ik} = Y_{i.k} - \overline{Y}_{..k}, \; i = 1, 2, \ldots, n_k, \; k = 1, 2.$$

Table 2 provides studentized intrasubject and intersubject residuals of the raw and log-transformed AUC data.

Table 2 Intrasubject and Intersubject Residuals of raw and log-transformed AUC for Clayton and Leslie's Study at the first Period

Sbj	Seq	$\hat{e}$	$\hat{S}$	$\hat{e}$ for log data	$\hat{S}$ for log data
1	CD	-0,757	-0,471	-0,143	-0,304
2	CD	2,733	5,249	0,547	1,44
3	CD	-0,012	-5,421	-0,279	-1,745
4	CD	0,648	-4,961	0,204	-1,202
5	CD	0,258	6,359	0,182	1,682
6	CD	-0,382	-4,161	-0,321	-1,092
7	CD	-4,982	4,919	-1,019	0,187
8	CD	1,133	1,909	0,325	0,965
9	CD	1,358	-3,421	0,504	-0,538
10	DC	0,499	-2,064	0,126	-0,316
11	DC	-1,741	6,736	-0,257	1,17
12	DC	-0,946	-5,514	-0,334	-1,42
13	DC	0,984	1,686	0,139	0,433
14	DC	-1,796	1,066	-0,35	0,318
15	DC	1,229	-0,164	0,238	0,041
16	DC	0,739	-2,984	0,227	-0,609
17	DC	0,094	3,106	-0,023	0,708
18	DC	0,939	-1,864	0,234	-0,326

Applying the consecutive test of up to *k*= 3, 2, 1 outliers in a normal sample of size *n* = 18 with the significance level $\alpha = 5\%$ for the data in Table 2 we obtain the following results

$\hat{e}$					
k	Outlier	T	$T_{3,\alpha}$	$T_{2,\alpha}$	$T_{1,\alpha}$
1	7	5.50	4.77	4.57	4.15
2	2	2.84	3.84	3.67	
3	14	2.34	3.50		

$\hat{S}$					
k	Outlier	T	$T_{3,\alpha}$	$T_{2,\alpha}$	$T_{1,\alpha}$
1	11	1.79	4.77	4.57	4.15
2	5	1.87	3.84	3.67	
3	12	1.89	3.50		

$\hat{e}$ for log-transformed data					
k	Outlier	T	$T_{3,\alpha}$	$T_{2,\alpha}$	$T_{1,\alpha}$
1	7	3.84	4.77	4.57	4.15
2	2	1.82	3.84	3.67	
3	14	1.72	3.50		

$\hat{S}$ for log-transformed data					
k	Outlier	T	$T_{3,\alpha}$	$T_{2,\alpha}$	$T_{1,\alpha}$
1	3	2.10	4.77	4.57	4.15
2	2	2.07	3.84	3.67	
3	4	2.27	3.50		

It can be seen that subject 7 is the single outlier and only for the raw data.
The assumptions (1.2) are examined using the Shapiro-Wilk test for normality and either Pearson's correlation coefficient or Spearman's rank correlation coefficient. Table 3 summarize the results obtained with and without subject 7.

Table 3 Summary of test results for assumptions (1.2)

	Shapiro–Wilk[a]		Pearson[b]		Spearman[c]	
Data set	$\tilde{e}_{ilk}$	$\tilde{S}_{ilk}$	Test	P	Test	P
Raw AUC	0.028	0.267	−0.204	0.417	−0.013	0.958
Log (AUC)	0.105	0.903	0.211	0.400	0.214	0.395
Raw AUC[d]	0.657	0.276	0.054	0.836	0.098	0.708
Log (AUC)[d]	0.117	0.876	0.339	0.183	0.275	0.286

[a] p-Value of Shapiro–Wilk's test for normality.
[b] Pearson correlation coefficient.
[c] Spearman's rank correlation coefficient.
[d] The results are obtained with deletion of subject 7.

As a result, if subject 7 is excluded, assumptions (1.2) hold for the raw data too.
Consequently, inclusion and exclusion of the possible outlying subject has a tremendous influence on assessment of the bioequivalence. Because one cannot determine whether the apparently nonconforming data result from laboratory error, data transcription, or other causes unrelated to bioequivalence FDA are against data removal.
FDA guidance on average bioequivalence recommend that logarithmic transformation be applied to the parmacokinetic responses AUC and C_{max} but not encourage firms to test for normality of data distribution after log-transformation.

REFERENCES

Andrews D. (1972) Plots of high-dimensional data, *Biometrics*, **28**, pp 125-136
Barnett V. and Lewis T. (1994) *Outliers in Statistical Data*, John Wiley, NY
Chow S.C. and Liu J.P. (2000) *Design and Analysis of Bioavailability and Bioequivalence Studies*, 2nd Edition, Marcel Dekker Inc., NY
Enachescu D. (2003) *Tehnici de Data Mining*, Ed. Universitatii din Bucuresti., Bucuresti
Enachescu C. and Enachescu D. (2000) Some simple rules for interpreting outputs of principal components and correspondence analysis, *Anal.Univ.Buc*., **XLIX**, seria Informatica, pp.3-8
Freidman J. (1987) Exploratory Projection pursuit, *JASA*, **82**, pp 249-266
Posse Ch. (1995a),Projection pursuit exploratory data analysis, *Comp. Stat. and Data Anal*., **29**, pp 669-687
Posse Ch. (1995b) Tools for two-dimensional exploratory data analysis, *J. of Comp. and Graphical Stat.* **4**, pp.83-100]