# Comparing methods for dimensionality reduction when data are density functions

**P. Delicado**[1]

[1] Universitat Politècnica de Catalunya, Barcelona, Spain; *pedro.delicado@upc.edu*

**Abstract**

Functional Data Analysis (FDA) deals with samples where a whole function is observed for each individual. A particular case of FDA is when the observed functions are density functions, that are also an example of infinite dimensional compositional data. In this work we compare several methods for dimensionality reduction for this particular type of data: functional principal components analysis (PCA) with or without a previous data transformation and multidimensional scaling (MDS) for different inter-densities distances, one of them taking into account the compositional nature of density functions. The different methods are applied to both artificial and real data (households income distributions).

**Key words:** Compositional data, Functional Data Analysis, income distributions, multidimensional scaling, principal components analysis.

# 1 Introduction

Observing and saving complete functions as results of random experiments is nowadays possible by the development of real-time measurement instruments and data storage resources. For instance, continuous-time clinical monitoring is a common practice today. Ramsay and Silverman (2005) express it saying that random functions are the *statistical atoms* in these cases. A particular case of functional data appears when the observed functions are density functions, that are also an example of infinite dimensional compositional data (Egozcue, Díaz-Barrero, and Pawlowsky-Glahn 2006).

Functional Data Analysis (FDA) deals with the statistical description and modeling of samples of random functions. Functional versions for a wide range of statistical tools (ranging from exploratory and descriptive data analysis to linear models to multivariate techniques) have been recently developed. Others techniques are specific of FDA, because they exploit the functional nature of this kind of data: *principal differential analysis* is a kind of principal component analysis made on the derivatives of the observed functions; *registration* is a pre-process step where a change of variable is done in each observed function in order to made them as similar as possible. See Ramsay and Silverman (2005) (the second edition of Ramsay and Silverman 1997) for a general perspective on FDA and Ferraty and Vieu (2006) for a non-parametric approach. Ramsay and Silverman (2002) present applications of FDA to a wide range of problems and disciplines. Special issues recently dedicated to this topic by several journals (Davidian, Lin, and Wang 2004, González-Manteiga and Vieu 2007, Valderrama 2007) bear witness to the interest for this topic in the Statistics community.

It is well worthwhile noting that random functions can also be obtained from standard random samples, by the application of non-parametric curve estimation methods. For instance, Kneip and Utikal (2001) use non-parametric density estimation methods to obtain annual income densities allowing them to study the temporal evolution of income density functions in United Kingdom from 1968 to 1988. The most frequent situation, however, is that of having observations densely sampled over time, space or other continuous parameter spaces. In these situations interpolation techniques (if the underlying sampled functions are smooth and there is no sampling noise) or smoothing methods (in other cases) allow us to transform the discrete observations to continuous functional objects.

Assume we have observed $n$ functions $f_1, \ldots, f_n$. In general, they belong to an infinite-dimensional functional space. The dimensionality reduction problem consists in looking for a low dimensional configuration $X$ (a $n \times q$ matrix, $q < n$, with rows $x_i$, $i = 1, \ldots, n$) and an application $\rho$ from $\mathbb{R}^q$ to the functional space such that $\rho(x_i)$ is close (in some sense) to the observed $f_i$, for $i = 1, \ldots, n$. Usually, dimensionality reduction aims at visualizing data, which requires a plane representation, that implies $q = 2$.

In this work we compare several methods for dimensionality reduction for this particular type of data: functional principal components analysis (FPCA) with or without a previous data transformation and multidimensional scaling (MDS) for different inter-densities distances, one of them taking into account the compositional nature of density functions. The different methods are applied to both artificial and real data (households income distributions in European countries; see Delicado 2007).

# 2 Dimensionality reduction for density functions

Ferraty and Vieu (2006) define a *functional variable* as a random variable $\boldsymbol{f}$ taking values in an infinite functional space, usually

$$L_2(I) = \{f : I \to \mathbf{R}, \text{ such that } \int_I f(t)^2 dt < \infty\},$$

where $I = [a, b]$, with $a$ and $b$ real numbers or $\pm\infty$, and $a < b$. An observation $f$ of $\boldsymbol{f}$ is called a *functional data*. A *functional data set* $f_1, \ldots, f_n$ is the observation of $n$ independent functional variables $\boldsymbol{f}_1, \ldots, \boldsymbol{f}_n$ identically distributed as $\boldsymbol{f}$.

We are particularly interested in the case where the observed functions are density functions in $[a, b]$: they are positive and integrate up to 1 on $[a, b]$. So we assume that the functional space in our case is

$$\mathcal{F}(I) = \{f : I \to \mathbf{R}, \text{ such that } f(t) \geq 0 \text{ for all } t \in I, \text{ and } \int_I f(t)dt = 1\}.$$

## 2.1 Functional Principal Component Analysis

In the context of FDA on $L_2(I)$, a version of the Principal Component Analysis (PCA) has been developed: the *Functional Principal Component Analysis (FPCA)*. The objective of FPCA can be stated as follows. Given a functional random sample with mean function $\bar{f}(t) = (1/n)\sum_{i=1}^{n} f_i(t)$, for all $t \in I$, we look for functions $g_1, \ldots, g_q$ (*principal functions* or *principal components*) in $L_2(I)$ and real numbers $\psi_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, q$, such that

$$\sum_{i=1}^{n} \int_I \left( (f_i(t) - \bar{f}(t)) - \sum_{j=1}^{q} \psi_{ij} g_j(t) \right)^2 dt$$

was minimum. Moreover, the functions $g_1, \ldots, g_q$ are asked to be orthonormal: $\int_I g_i(t)g_j(t)dt$ is equal to 0 if $i \neq j$ and equal to 1 if $i = j$. In other words, we are looking for a representation of functional data in a $q$-dimensional space (that spanned by the functions $g_1(\cdot), \ldots, g_q(\cdot)$):

$$f_i(t) \approx \bar{f}(t) + \sum_{j=1}^{q} \psi_{ij} g_j(t), \ t \in I, \ i = 1 \ldots n. \tag{1}$$

It can be proven that the principal functions are *eigen-functions* of the sampling covariance operator:

$$\int_I \Gamma_n(t, s) g_j(s) ds = \lambda_j g_j(t), \text{for all } t \in I, \tag{2}$$

where

$$\Gamma_n(t, s) = \frac{1}{n} \sum_{i=1}^{n} (f_i(t) - \bar{f}(t))(f_i(s) - \bar{f}(s)),$$

and that

$$\psi_{ij} = \int_a^b (f_i(t) - \bar{f}(t)) g_j(t) dt, \ i = 1, \ldots, n, \ j = 1, \ldots, q.$$

Coefficient $\psi_{ij}$ is the score of the observation $i$ on the $j$-th principal component. The numbers $\lambda_1, \ldots, \lambda_q$, known as *eigen-values*, they are sorted in decreasing order and they are proportional to the proportion of total variability explained by the corresponding principal functions.

The dimensionality reduction problem is approached in this context by defining the matrix $X$ with elements $(i, j)$ equal to $\psi_{ij}$, $i = 1, \ldots, n$, $j = 1, \ldots, q$. The application $\rho : \mathbb{R}^q \longrightarrow L_2(I)$ is defined by the right hand side of Equation (1).

There are different approaches to solve Equation (2) in practice. Ramsay and Silverman (2005) propose to express observed functions as linear combinations of B-splines functions forming an approximate base of $L_2(I)$. This way Equation (2) can be re-expressed as a matrix equation to be solved by standard methods.

A different solution is suggested by Kneip and Utikal (2001). Once the original functions have been properly smoothed (if required), the centered functions are evaluated in a fine grid of evenly

spaced points of $I$: $t_1 = a, \ldots, t_M = b$. Let $F$ be the $n \times M$ the resulting data matrix. It can be proven that for large values of $M$ the solutions of (1) can be derived from eigenvalues and eigenvectors of $FF^T$ or $F^TF$, the last one having the advantage of having dimension $n \times n$, what is very convenient given that usually $n << M$. In this paper we follow this approach.

A way to interpret the meaning of the principal component functions is that they represent the main variation modes of the observed functions around the global mean function. The mean function $\bar{f}$ represents what is common to all the data, the centered functions $(f_i - \bar{f})$ account for individual differences and the principal component functions summarize what is common in the way individual are diverse.

The following observations help to interpret principal components. Principal component scores are the scalar product between observed functions and eigen-functions. So argument values $t \in I$ with large values in a particular eigen-function have a great importance in the corresponding principal component. A convenient graphical way to interpret a principal component is to add and subtract from the mean the eigen-function multiplied by an appropriate constant. This gives us an idea of how the observed functions differ from the mean for observations that have significant positive or negative values in this principal component (looking at parts of $I$ where the shifted mean function is above or below the original mean function and where the maxima and minima values are). Principal components show several orthogonal patterns in a decreasing order of importance.

Let us return to the case of observing density functions, so our functional data belongs to $\mathcal{F}(I)$, that in general does not coincide with $L_2(I)$. Therefore FPCA previously introduced could present some problems now, as the examination of the right hand side of Equation (1) reveals: it is not sure that $\tilde{f}_i(t) = \bar{f}(t) + \sum_{j=1}^{q} \psi_{ij} g_j(t)$ is always a density function, even having observed density functions. It is easy to prove that $\bar{f}(t)$ is a density function and that $\int_I g_j(t)dt = 0$, but $\tilde{f}_i(t)$ can be negative for some $t \in I$ because both $\psi_{ij}$ and $g_j(t)$ can take negative values.

A standard solution to this problem is to look for a transformation $\Psi : \mathcal{F}(I) \mapsto L_2(I')$ and then to apply FPCA to the transformed functions. For instance, $\Psi(f)(\cdot) = \log(f(\cdot))$ is a sensible choice when the transformed densities are in $L_2(I)$.

Consider now the case of $f_{ri}$ being close to normal density functions (in this case $I \equiv (-\infty, \infty)$). The following property is verified when $f_{\mu,\sigma}(x)$ is the density function of a $N(\mu, \sigma^2)$ random variable:

$$\frac{\partial}{\partial x} \log(f_{\mu,\sigma}(x)) = -\frac{x - \mu}{\sigma}.$$

So the functional

$$\Psi_N(f)(x) \equiv \frac{\partial}{\partial x} \log(f(x))$$

would transform a density function $f$ (assumed to be close to normality) to a function that would be close to a straight line with negative slope. Therefore applying the transformation $\Psi_N$ could be appropriate when observed densities $f_{ri}$ are close to normality. This property is exploited in Ramsay and Silverman (2002) in the context of Functional Principal Component Analysis when data are density functions: they first transform the observed densities with this functional and then they perform FPCA on the transformed functions. The main drawback of this practice is that the transformed functions $g(x) = -(x - \mu)/\sigma$ are not in $L_2((-\infty, \infty))$. Therefore the interval $(-\infty, \infty)$ must be reduced to a compact interval $[a, b]$. The choice of $[a, b]$ is arbitrary and it could influence the final result.

For the case of densities $f_{ri}(y)$, $y \in (0, \infty)$, close to log-normal density functions, a suitable functional is

$$\Psi_{lN}(f)(x) = \Psi_N(f(\exp(x)) \exp(x)),\ x \in (-\infty, \infty),$$

given that $f(\exp(x)) \exp(x)$ is the density function of $X = \log(Y)$, $Y$ having density $f$.

## 2.2 Multidimensional Scaling

Multidimensional Scaling (MDS) is a generalization of PCA when the information about data is given by a inter-individuals distance matrix, instead of by a standard data matrix. Assume that there are $n$ individuals and that a distance (or dissimilarity) function between individuals is available. Let $d_{ij} \geq 0$ be the dissimilarity between individuals $i$ and $j$. It is assumed that $d_{ij} = d_{ji}$ and that $d_{ii} = 0$ for all $i, j = 1 \ldots, n$. Let $\Delta$ be the $n \times n$ matrix with element $(i, j)$ equal to $d_{ij}$. Assume that for $q \leq n$ there exists a $n \times q$ data matrix $X$ such that the Euclidean distance between the $i$-th and $j$-th rows of $X$ is $d_{ij}$. We say that $X$ is an *Euclidean configuration* of $\Delta$. Such a configuration does not always exist. When it does, $\Delta$ is said to be *Euclidean*. In this case the $X$ can be chosen having orthogonal columns, that are called *principal coordinates*.

Define the $n \times n$ matrix $D$ with element $(i, j)$ equal to $d_{ij}^2$. It can be proved (Borg and Groenen 2005) that $\Delta$ is Euclidean if and only if

$$Q = -\frac{1}{2} PDP$$

is positive definite, where $P = I - (1/n)\mathbf{1}\mathbf{1}^{\mathrm{T}}$ is a centering matrix ($\mathbf{1}$ is the $n \times 1$ vector of ones). In this case, let $Q = V \Lambda V^{\mathrm{T}}$ be the spectral decomposition of $Q$ ($V$ is a $n \times n$ orthonormal matrix, and $\Lambda = diag(\lambda_1, \ldots, \lambda_n)$ with $\lambda_1 \geq \cdots \geq \lambda_n$). Let $\tilde{X}_q = V_q \Lambda_q^{1/2}$, where $V_q$ is formed by the first $q$ columns of $V$ and $\Lambda_q = diag(\lambda_1, \ldots, \lambda_q)$. Then $Q \approx \tilde{X}_q \tilde{X}_q^{\mathrm{T}}$ and $\tilde{X}_q$ is a $q$-dimensional approximate Euclidean configuration of $\Delta$.

In terms of dimensionality reduction, in MDS the low dimensional configuration we are seeking is $\tilde{X}_q$. In MDS what is desired is that the Euclidean distance between the rows $i$ and $j$ of $\tilde{X}_q$ is as similar as possible to $d_{ij}$, the distance between data $i$ and $j$.

The relation between PCA and MDS is as follows. Let $X$ be a $n \times p$ data matrix. Let $d_{ij}$ be the Euclidean distance between the rows $i$ and $j$ of $X$. Then $\tilde{X}_q$ ($q \leq p$) coincides with the matrix of the first $q$ principal components of $X$ (see Peña 2002, Section 6.5).

In this paper we apply MDS to the specific case of density functions. Many distances can be computed between density functions. For instance, given that density functions are always in $L_1$, we can base MDS on $L_1$ distances between observed densities: $\|f_i - f_j\|_1 = \int_a^b |f_i(x) - f_j(x)| dx$. Other distances can also be used: for instance, the Hellinger distance (that is the $L_2$ distance between squared root of densities), the $L_2$ distance between densities (assuming they are well defined; for instance assuming that densities are bounded on $[a, b]$) the $L_2$ distance between logs of densities (assuming they are well defined) or the symmetrized version of the Kullback-Leibler divergence:

$$d_{KL}(f_i, f_j) = \int_a^b \log\left(\frac{f_i(x)}{f_j(x)}\right) f_i(x) dx + \int_a^b \log\left(\frac{f_j(x)}{f_i(x)}\right) f_j(x) dx.$$

More in general, let $\Psi : \mathcal{F}(I) \mapsto \mathcal{F}_\Psi$ be a transformation from $\mathcal{F}(I)$ to another functional space $\mathcal{F}_\Psi$, and let $d^*(\cdot, \cdot)$ be a distance between elements of $\mathcal{F}_\Psi$. Let $f$ and $g$ be two density functions, then

$$d_\Psi(f, g) = d^*(\Psi(f), \Psi(g))$$

is a distance between $f$ and $g$. For instance, assuming that $\mathcal{F}_\Psi = L_2(I')$ and using the $L_2$ distance in $L_2(I')$ as $d^*$, the results of MDS from distance $d_\Psi$ coincide with those obtained by doing FPCA on the transformed functions $\Psi(f_i)$, $i = 1, \ldots, n$.

## 2.3 Density functions as compositional data

The set $\mathcal{F}(I)$ of density functions on $I$ is a convex set of $L_1(I)$ that is not a linear subspace when using ordinary sum and multiplication by real constant, as pointed out by Egozcue, Díaz-Barrero, and Pawlowsky-Glahn (2006). They note that density functions are in fact infinite dimensional

compositional data. Therefore they propose to extend Aitchison's geometry (well developed for data living in the finite dimensional simplex; see (Pawlowsky-Glahn and Egozcue 2001)) to the infinite dimensional simplex $\mathcal{F}(I)$. In particular, when $I = [a, b]$ is a finite interval they give a Hilbert space structure to the subset of $\mathcal{F}(I)$ formed by densities whose logarithm is square-integrable. In this space, the distance between two densities $f$ and $g$ is defines as

$$d_A(f,g) = \left[\frac{1}{2\eta}\int_a^b\int_a^b\left(\log\frac{f(x)}{f(y)} - \log\frac{g(x)}{g(y)}\right)^2 dxdy\right]^{1/2},$$

where $\eta = b - a$ and the subscript $A$ refers to Aitchison's geometry. Once $d_A$ has been defined, MDS can be applied to the matrix $\Delta_A$ of inter-densities distances $d_A(f_i, f_j)$. Therefore all we have said in Subsection 2.2 applies here.

Observe that

$$d_A(f,g) = \frac{1}{\sqrt{2\eta}} d_{L_2(I\times I)}(f^*, g^*)$$

where $f^* : I \times I \mapsto \mathbb{R}$ is defined as $f^*(x, y) = \log(f(x)/f(y))$, and $g^*$ is defined analogously. We conclude that $\Delta_A$ is a Euclidean matrix.

# 3 Application to artificial data

We have generated three sets of density functions and then we analyze them with the different dimensionality reduction methods presented in the previous section. The sets of densities are the following:

Set 1: For $i = 1, \ldots, n_1 = 21$, $f_i$ is the density of a $N(\mu_i, 1)$, with $\mu_i = -3 + 6(i-1)/20$.

Set 2: For $i = 1, \ldots, n_2 = 21$, $f_i$ is the density of a $N(0, \sigma_i^2)$, with $\log(\sigma_i) = -.5 + (i-1)/20$.

Set 3: For $i = 1, \ldots, 9$ and for $j = 1, \ldots, 9$, $f_{ij}$ is the density of a $N(\mu_i, \sigma_j^2)$, with $\mu_i = -2+4(i-1)/8$ and $\log(\sigma_i) = -.5 + (j-1)/8$. Therefore there are $n_3 = 81$ functions in the set.

From this definition it follows that Sets 1 and 2 are intrinsically one-dimensional in the sense that they are included in a one-dimensional manifold of $\mathcal{F}(I)$: the image of an application from a subset of $\mathbb{R}$ to $\mathcal{F}(I)$. In the same sense, Set 3 is two-dimensional. Therefore good dimension reductions techniques must discover these intrinsic dimensions.

## 3.1 FPCA for density functions

In Subsection 2.1 we have noted that FPCA is not a well suited technique for density functions. We are showing now the practical drawbacks of FPCA when applied to density functions.

Figure 1 summarizes the FPCA for densities in Set 1. Several erroneous conclusions could be derived from the analysis of this graphical output. First, the percentage of explained variance for the first two principal components (56.50% and 29.98%, respectively) indicates that the intrinsic dimension of the set is at least two. But if we look at the projection of functions on the components plane (top right panel in Figure 1) we observe that a second dimension is needed only to accommodate the nonlinear structure of the set, that can not entirely be reflected in the first principal component. The reason of this fact is that the one-dimensional manifold of $\mathcal{F}(I)$ where Set 1 is included has a curvature that FPCA (a linear method) is not able to detect. Observe that the components plane clearly shows a nonlinear relationship between first and second principal components: if they were independent their interpretation could be done separately, but FPCA only guarantees that they are uncorrelated.

The graphic labeled as "mean +/- PC 1" is also misleading by several reasons. The main one is that it is based on representing the mean function (width black line) but arithmetic mean is not an appropriate location measure for densities in Set 1 because the mean function does not belong to the one-dimensional manifold where Set 1 is included: it does not correspond to a normal density; it does not have standard deviation equal to 1; in the components plane the mean function has coordinates $(0, 0)$, being a point that is far from the points corresponding to the densities in Set 1. Moreover, this graphic reinforces the idea that the data set could be approximated by adding and subtracting a multiple of the first principal component from the mean function, and this interpretation is strongly dependent on the assumption of a linear structure for the data. Given that Set 1 is not linearly structured, this graphic only lead us to false conclusions. The same applies for the graphic labeled as "mean +/- PC 2". The two bottom panels of Figure 1 show the density functions in Set 1 and the approximated functions derived from FPCA (using the right hand side of Equation 1), from top to bottom. It is clear that the approximation is quite bad.

We analyze now Figure 2, the graphical output of FPCA for densities in Set 2. In this case FPCA is able to detect that the right dimension of the set is one (98.93% of the variability is explained by the first principal component). Nevertheless, other drawbacks of FPCA for density functions (already pointed out when talking about for Set 1) are also present here: the nonlinear relationship between first and second principal components, the mean function is not in the one-dimensional manifold of $\mathcal{F}(I)$ where Set 2 is included (but in this case it is closest than in the case of Set 1), the approximated density functions do not look like the true functions in Set 2.

When looking at FPCA for densities in Set 3 (see Figure 3) we observe that the first two components account for the 92.48% of the total variability, what is erroneously indicating that the intrinsic dimension of the data is larger that two. The first principal function coincides with that of the FPCA for Set 1, and the second one is the same (up to a change of sign) as the first one obtained for Set 2. This fact shows that FPCA at least has been able to discover the two main sources of variation in Set 3. Nevertheless, the projection of functions on the components plane indicates that the two main modes of variation do not correspond with changes in mean and changes in standard deviation. Instead of that, this plane reflects a complicated nonlinear relation between principal components and changes on $\mu$ and $\sigma$. Again, the approximated density functions derived from FPCA are far from those in Set 3.

### 3.2 FPCA for transformed density functions using $\Psi_N$

The bad behaviour of FPCA for density functions can be amended if densities are transformed by an appropriate functional, as we have noted in Subsection 2.1. For normal densities we have indicated that $\Psi_N$ is a good functional, because it transforms normal densities to straight lines, that will always be in a two dimensional vector space, where linear operators (as FPCA) work properly.

Figure 4 is the graphic output corresponding to FPCA applied to Set 1 transformed by $\Psi_N$. It is clear that the dimension of Set 1 is one. The first principal function is almost constant and the second one is practically equal to zero. The present noise is due to numeric errors in the computation of eigen-functions, that are also the responsible of the surprising position of function 1 in the second principal component (see top right panel).

Instead of the graphic "mean +/- PC i" (that works on the space of transformed densities, that is, in the set of straight lines) we have drawn the densities that would be obtained inverting the operator $\Psi_N$. So we have the representation of three densities for the principal component $i$ ($i = 1, 2$): $\Psi_N^{-1}$(mean - PC i), $\Psi_N^{-1}$(mean), $\Psi_N^{-1}$(mean + PC i). In this case these three functions calculated for the first principal component give an exact idea about the way densities in Set 1 vary around its center (that could be located in the standard normal density). The corresponding graphic to the second principal component only shows very small noisy deviations from this center. Observe that the approximated functions are now almost identical to the true densities.

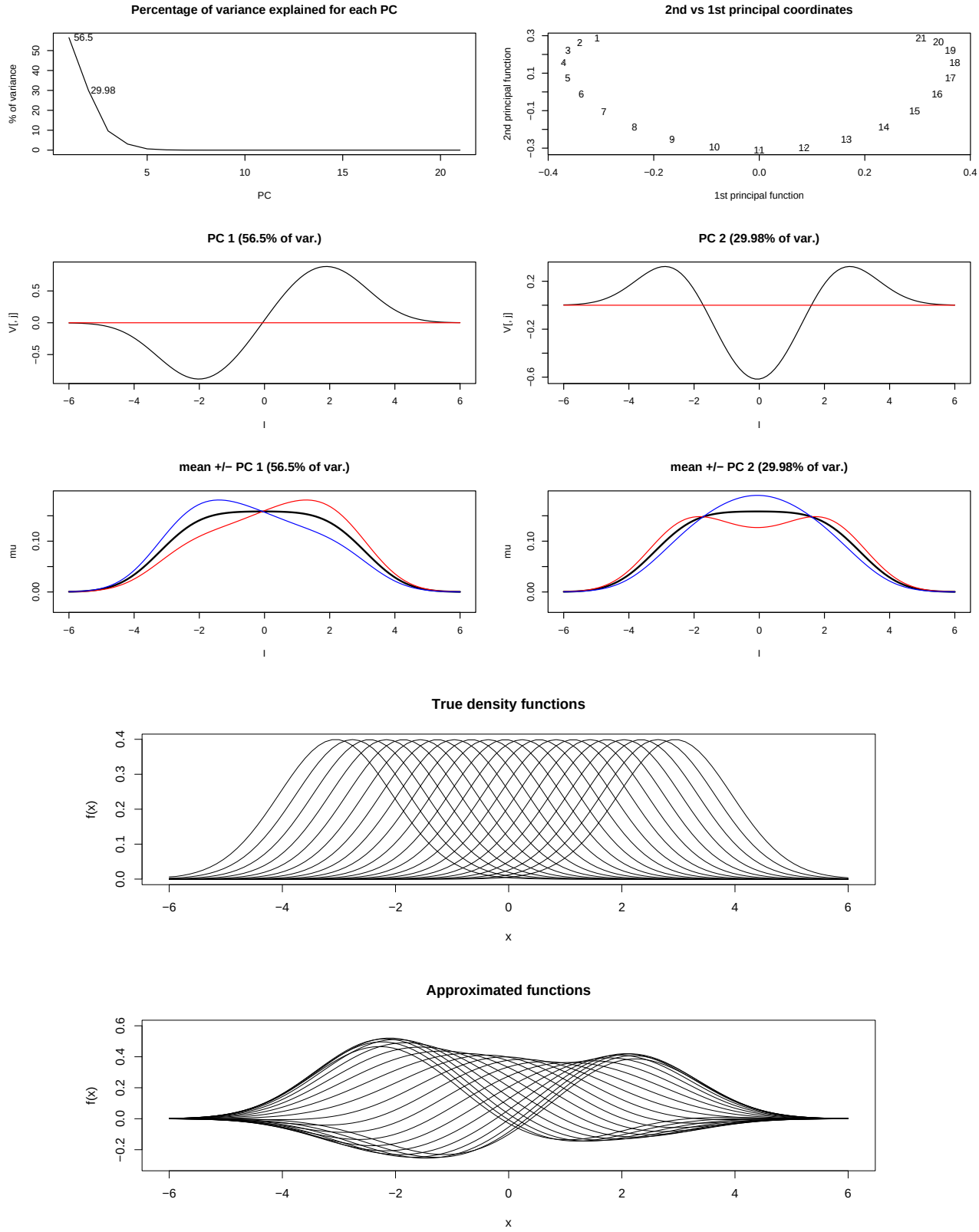


**Figure** 1: FPCA for densities in Set 1.

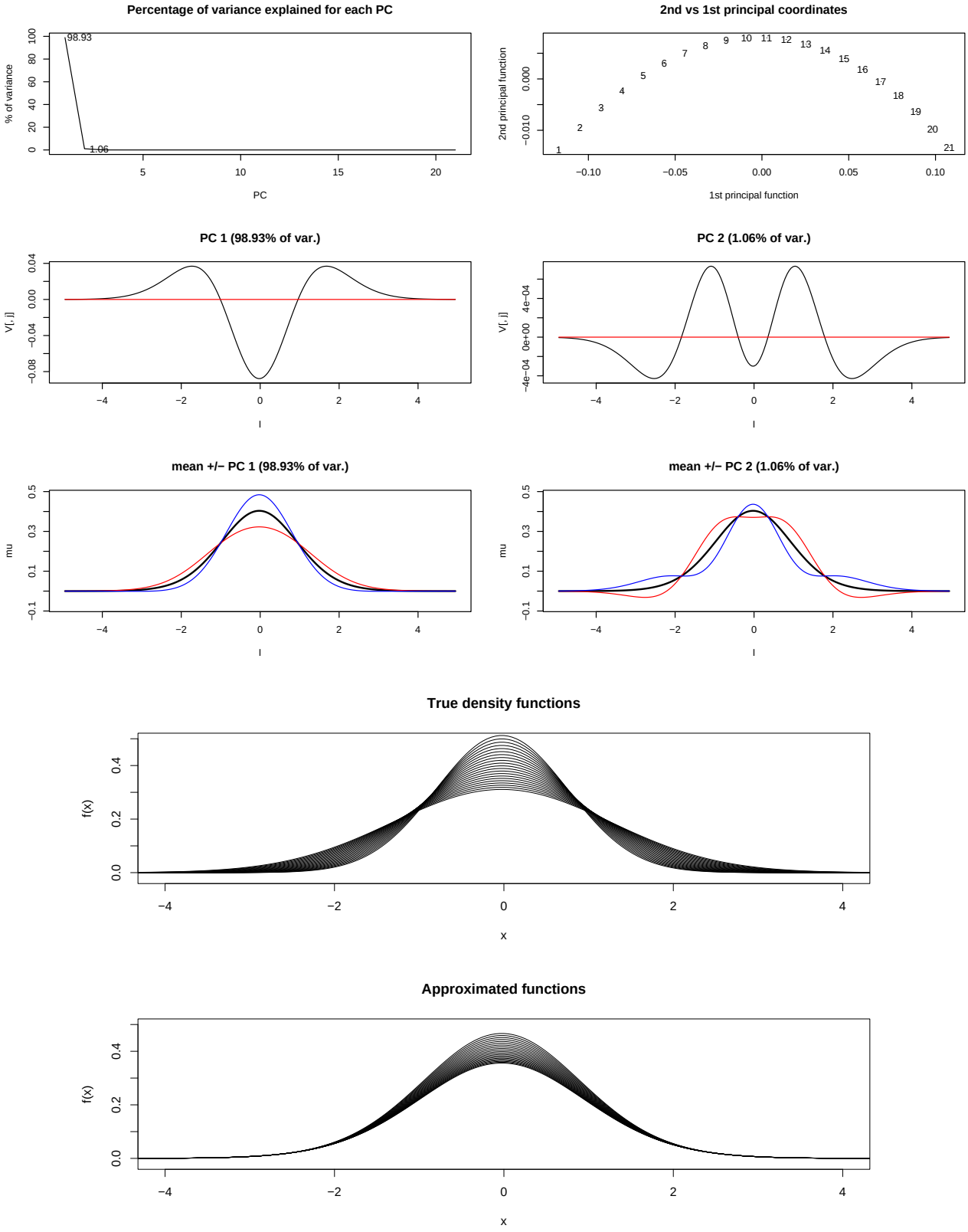


**Figure** 2: FPCA for densities in Set 2.

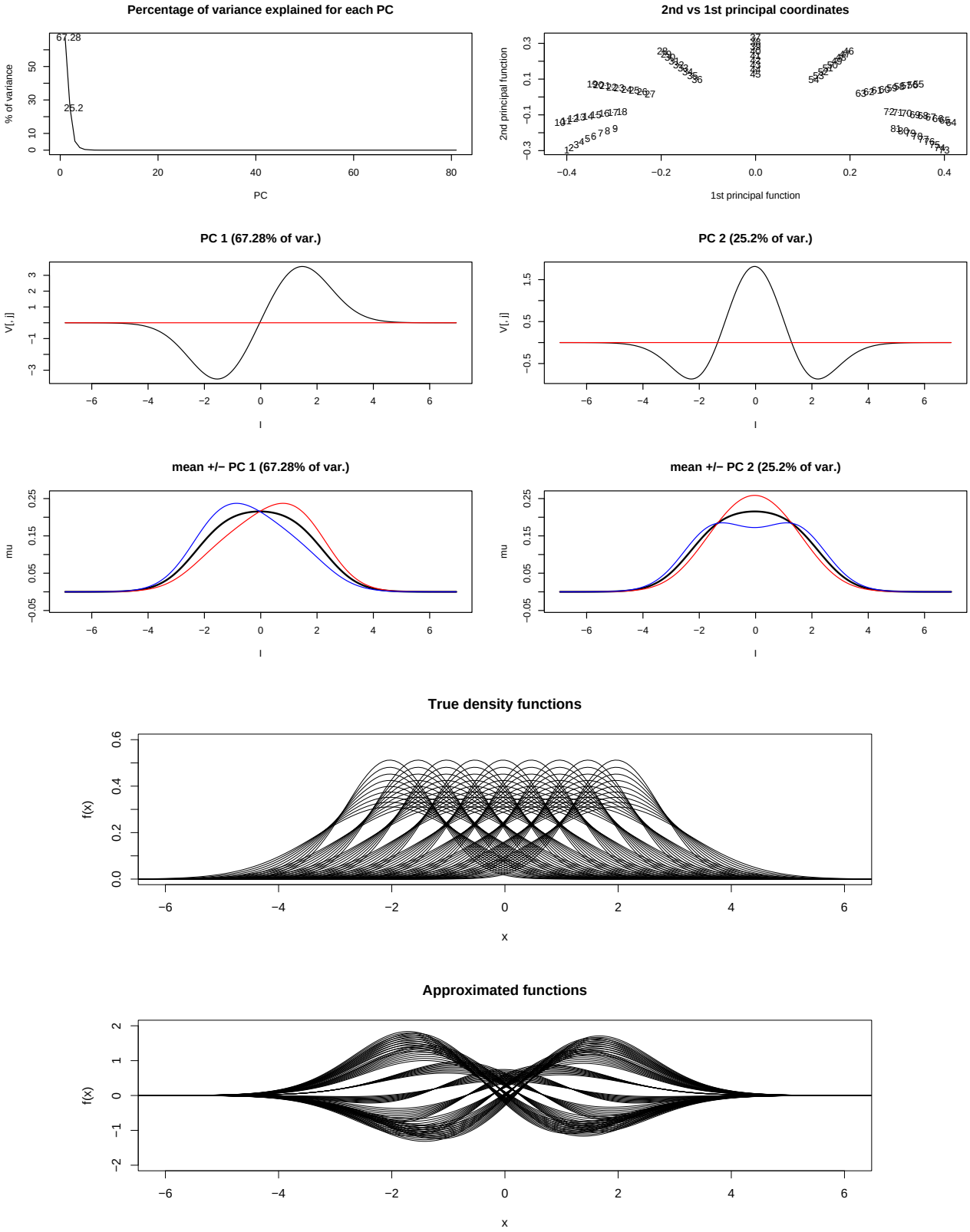


**Figure** 3: FPCA for densities in Set 3.

Similar comments on the well behaviour of FPCA combined with transformation $\Psi_N$ also apply to Sets 2 (Figure 5) and 3 (Figure 6). Only a remark for Set 3 is in order. Observe that the projection of functions on the components plane shows that the first principal components are not independent, even if they present a dependent structure much more simple than that obtained applying FPCA directly on density functions in Set 3.

It is appropriate to remember that the good results of FPCA illustrated in Figures 4, 5 and 6 have been possible because in this artificial data case we know the right transformation $\Psi_N$ that linearize the functional data sets. In general such a transformation would be unknown. So we need dimension reduction methods working without the knowledge of any specific linearizing transformation.

### 3.3 MDS for several distances between density functions

We analyze in this section the performance of MDS based on six different distances when applied to Sets 1, 2 and 3. The distances used are: $L_1$, $L_2$, Hellinger, $L_2$ between logarithms, Symmetrized Kullback-Leibler divergence and $d_A$ (the distance that take into account that density functions are compositional data).

The results of MDS for Sets 1, 2 and 3 are illustrated in Figures 7, 8 and 9, respectively. Each row of graphics correspond to a distance. The left panel of the row is the plane of the two first principal coordinates (similar to the plane of the first two principal components). The percentage of explained variability are calculated as the quotient between the corresponding eigenvalues of matrix $Q$ and the sum of the absolute values of all the eigenvalues of $Q$ (not all of them must be positive if the distance is not Euclidean; $L_1$ distance and Symmetrized Kullback-Leibler divergence are the only not Euclidean distances among those we are used). The other two panels are analogous to the graphics labeled as "mean +/- PC i" ($i = 1, 2$) in Figures 1 to 6, but with some particularities: the function represented with width black line is calculated as the geometric mean of those functions having both first and second principal coordinates values between first and third quartiles (if no function verifies both conditions simultaneously, then functions verifying one of them are taken to compute the geometric mean), conveniently normalized for being a density function. Then the (rescaled) geometric mean of functions having first principal coordinate lower than the first quartile is drawn in blue, and finally the same is done for those with values larger that the third quartile and the corresponding function is drawn in red. The same is done for the second principal coordinate. Observe that MDS does not provide any function such as the principal functions obtained by FPCA. Therefore it is not possible to do graphics directly comparable with those labeled as "mean +/- PC i". This is the reason why we have developed the kind of graphics explained before.

For Set 1 of densities, Figure 7 shows that distances $L_1$, $L_2$ and Hellinger has a similar behaviour to FPCA done on density functions (two dimensions are required, and nonlinear relations between the first and second principal coordinates are present). Distance $d_A$ reproduce the output obtained doing FPCA on densities transformed by $\Psi_N$. This is true in this case because $d_A(f, g)$ is equivalent to $L_2$ distances between $f^*$ and $g^*$, where

$$f^*(x, y) = \log(f(x)/f(y)) = \frac{\mu}{\sigma^2}(x - y) + \frac{y^2 - x^2}{2\sigma^2} = \frac{1}{\sigma^2}\left(\mu(x - y) + (y^2 - x^2)/2\right),$$

in the case of $f$ being the density function of a $N(\mu, \sigma^2)$. When moving $\mu$, the set of functions $f^*(x, y)$ moves over a one-dimensional vector space. Therefore MDS on Set 1 using $L_2$ distances between $f^*$ and $g^*$, that is equivalent to doing FPCA over functions $f^*$, identifies perfectly this one-dimensional linear manifold. Using distances $L_2$ between logarithms or Symmetrized Kullback-Leibler divergence, is almost equivalent to using $d_A$, with the advantage that numeric errors has less influence on the second principal coordinate (compare right panels in the rows 4, 5 and 6 of Figure 7).

The behaviour of MDS applied to Set 2 and 3 (Figures 8 and 8) is very similar to that described

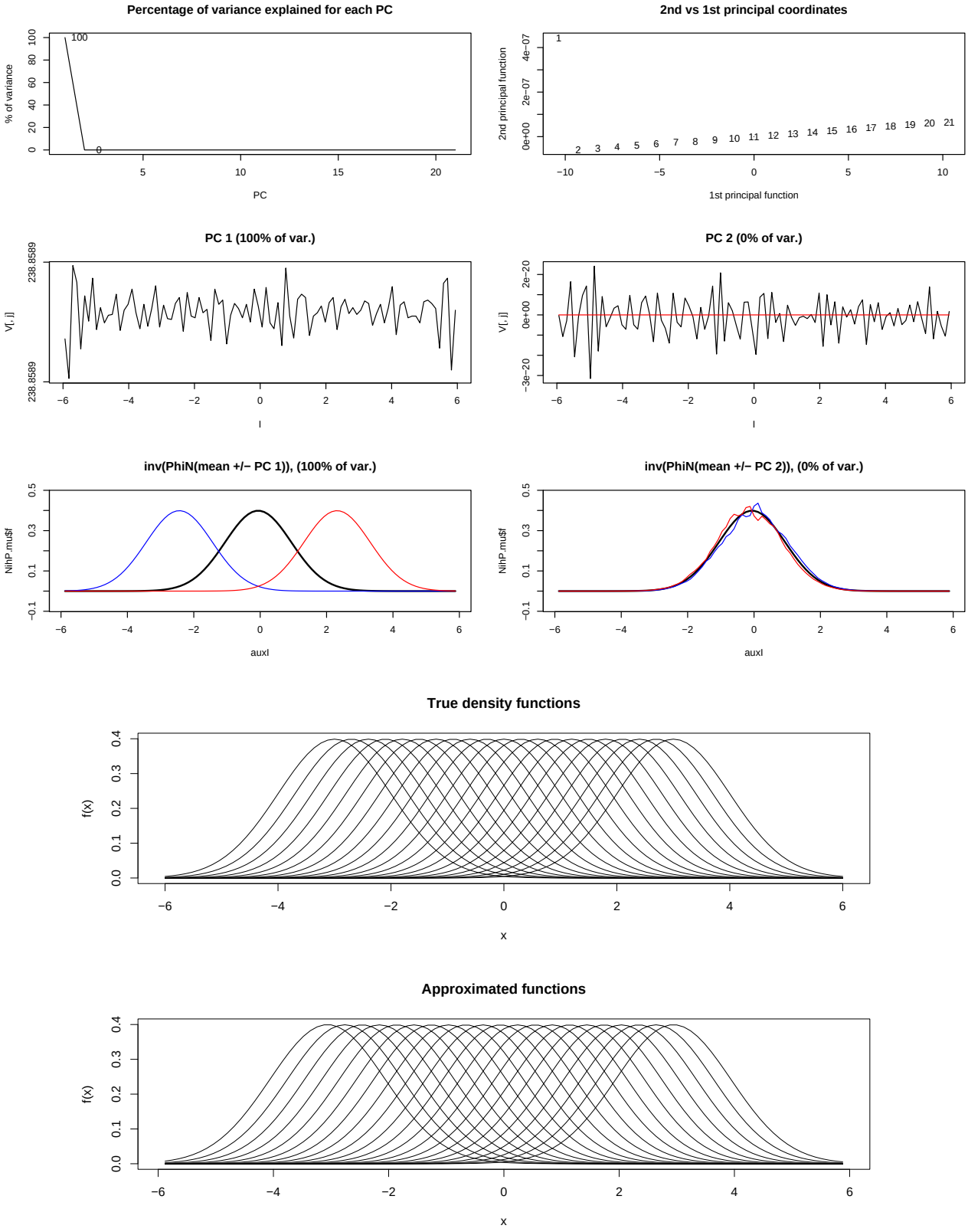


**Figure** 4: FPCA for transformed densities in Set 1.

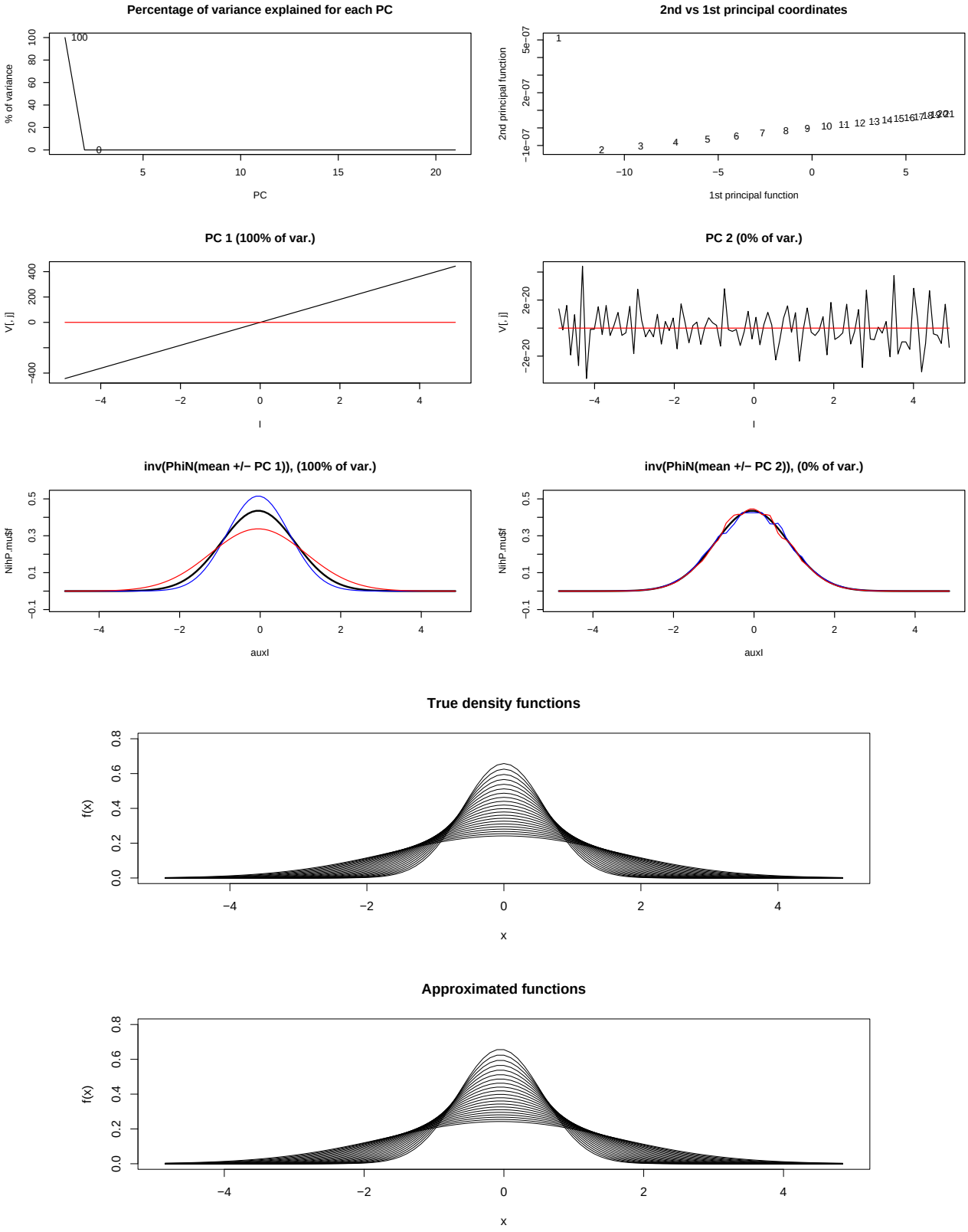


**Figure** 5: FPCA for transformed densities in Set 2.

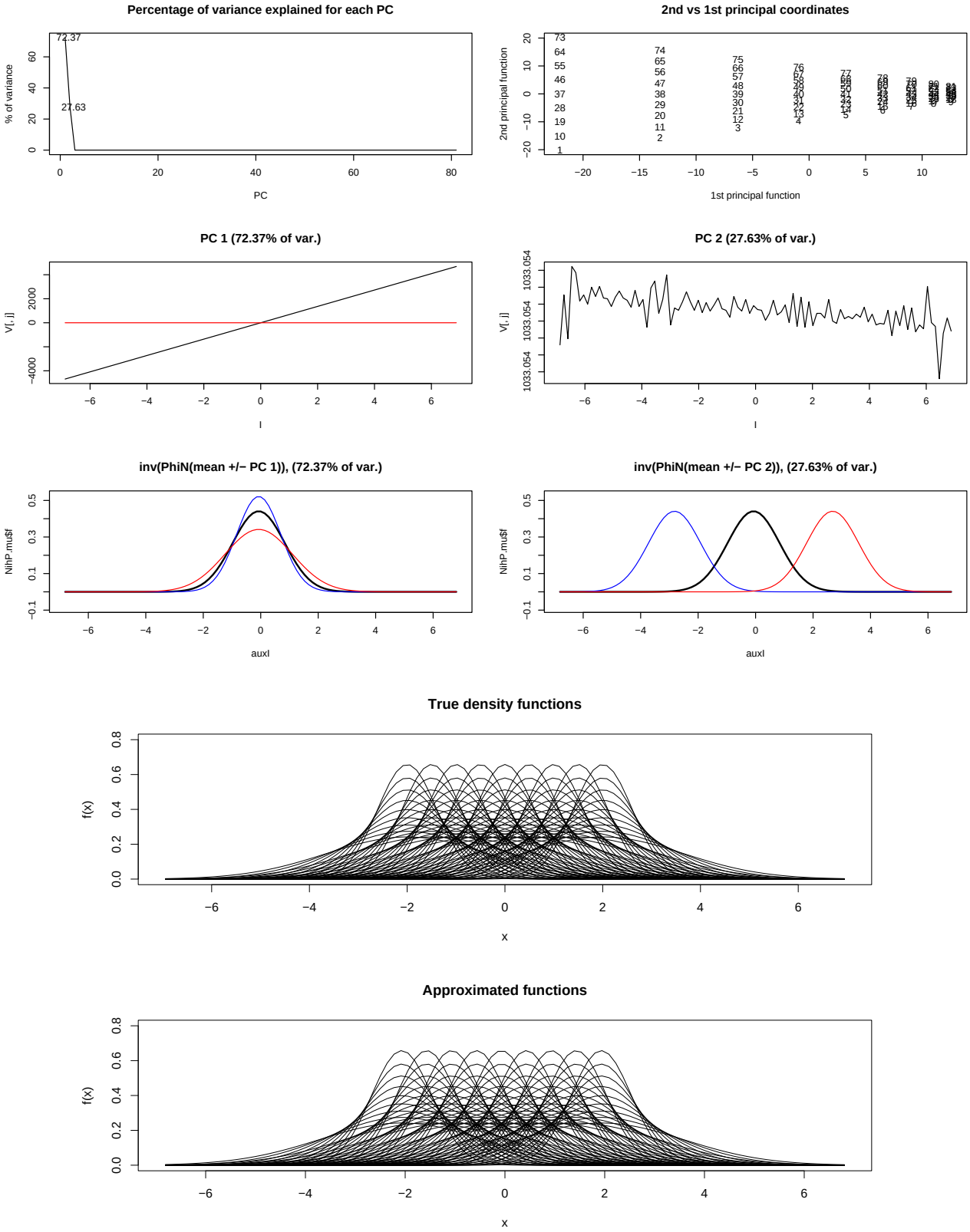


**Figure** 6: FPCA for transformed densities in Set 3.

for Set 1. We conclude that distances $L_2$ between logarithms, Symmetrized Kullback-Leibler divergence and $d_A$ give the best results, comparable to those achieved when the specific linearizing transformation is known.

# 4 A real data example: Households income distributions

In this section we analyze income distributions for European countries. Let $f_i(x)$ be the relative equivalent disposable income (after taxes and benefits) density function of country $i$, one of the 15 countries forming the European Union before May 2004.

The true densities $f_i(x)$ are not available, so we are working with non-parametric estimates of them. The used incomes are *disposable* (or *net*) because they are the result of applying taxes and social benefits to the household gross income. They are *equivalent incomes* in the sense that the household incomes are divided by the equivalent number of adults living in there, according to the modified OECD scale: one adult (person aged 14 or plus), plus one half of the additional number of adults, plus 0.3 times the number of children. Finally they are *relative* because in each country the observed equivalent incomes are divided by the country median. The information about the income distribution in European countries comes from the 8th wave of the European Community Household Panel (ECHP-w8) corresponding to year 2001. Any household in the sample has a specific weight and this fact has to be taken into account in the estimation process.

In order to do the non-parametric density estimation, we take the log of the data because their marked right asymmetry. Given that not all the relative income data are positive, a positive constant $c$ has to be added to each observation before taking logs (we have chosen $c = 1$). Then usual kernel estimation is done in the transformed scale, and a change-of-variable formula is used to recover a density estimation in the original scale. See Delicado (2007) for a similar estimation process for regional income.

We select the bandwidth using the *normal reference rule* for weighted data. It is well known that this rule is appropriate only when data are near normality (that is the case for $\log(x_i + c)$) and that it tends to over-smooth (to produce too high values for the bandwidth). In order to correct the over-smoothing, a common practice is to multiply the proposed values by a positive constant lower than 1. In our case, we always take 2/3 times the values provided by the normal reference rule (Luxembourg, where the normal reference rule is respected, is an exception because otherwise the estimated density would be very bumpy). The constant 2/3 was chosen by visual inspection. The same applies for the constant $c$ choice.

We use the library `sm` (Bowman and Azzalini 2001) in the package `R` (R Development Core Team 2005), that implements the normal reference bandwidth choice rule and kernel estimation for weighted data. All densities are evaluated in 1001 points evenly spaced from -1 to 5. The estimated densities are shown in Figure 10 (panel labeled as Estimated density functions).

Figure 10 is the graphical output of FPCA applied directly to the estimated density functions (observe that in this case we do not know what transformation $\Psi$ would transform the densities into a linear set of functions). It seems that the set is two-dimensional. The first principal function can be interpreted as a polarization dimension: Countries with positive value of the first principal component (Greece, Portugal, Luxembourg, Ireland, Spain) has densities with less weight around the median ($x = 1$) and more in lower and higher values, so their polarization is bigger. The opposite happens for countries with negative values (Denmark, Sweden, Germany, Finland). The second component has a difficult interpretation, and it seems only take into account the particularities of Luxembourg. These conclusions have to be taken carefully, because the approximated functions are considerable different to the estimated densities.

The MDS analysis has been done using four different distances ($L_1$, $L_2$, Hellinger and Symmetrized Kullback-Leibler divergence). The other two ($L_2$ between logarithms and $d_A$) present numerical problems when computed on this data set because $\log(f_i(x))$ are not in $L_2([-1, 5])$. Figure 11

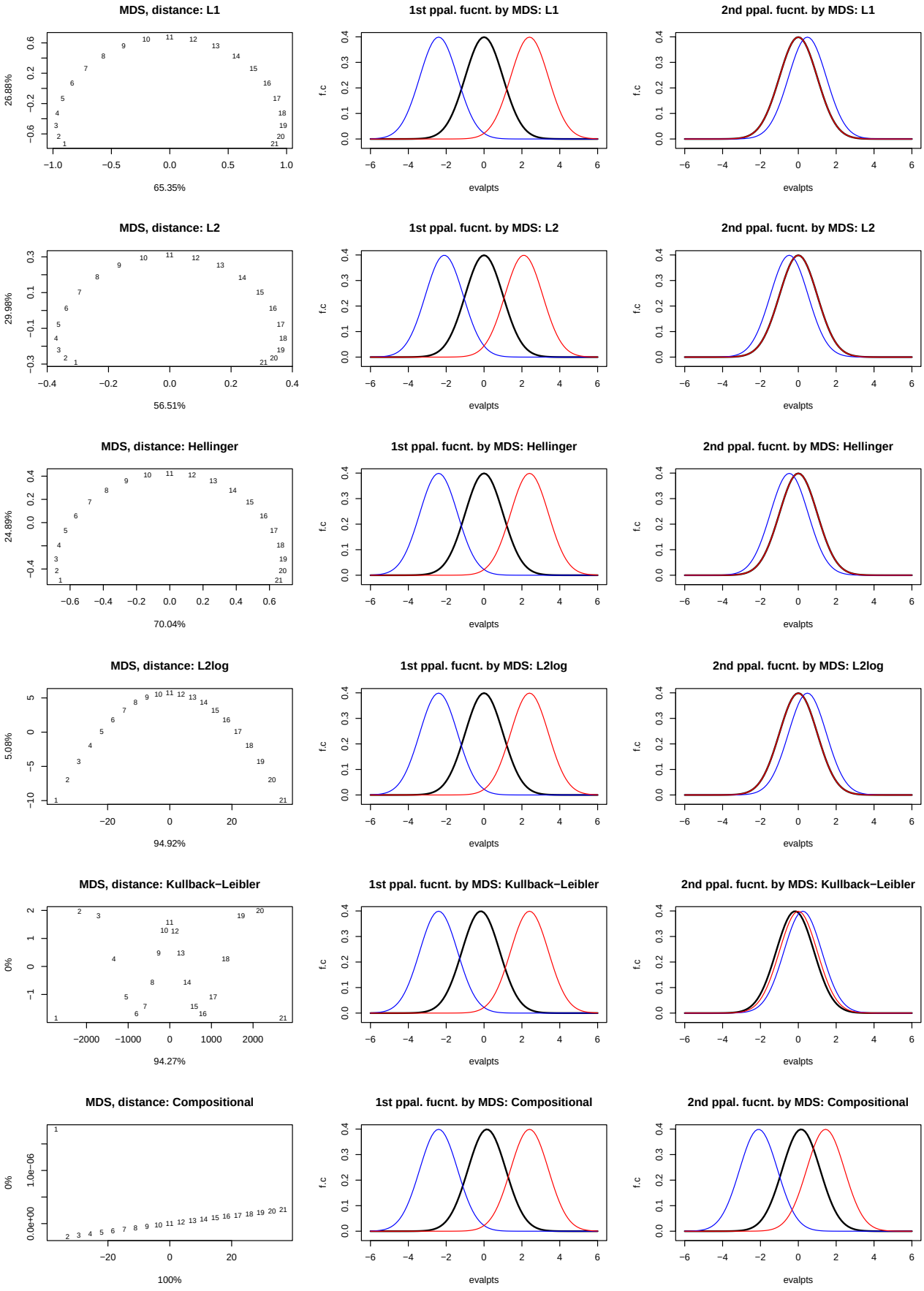


**Figure** 7: MDS for densities in Set 1.

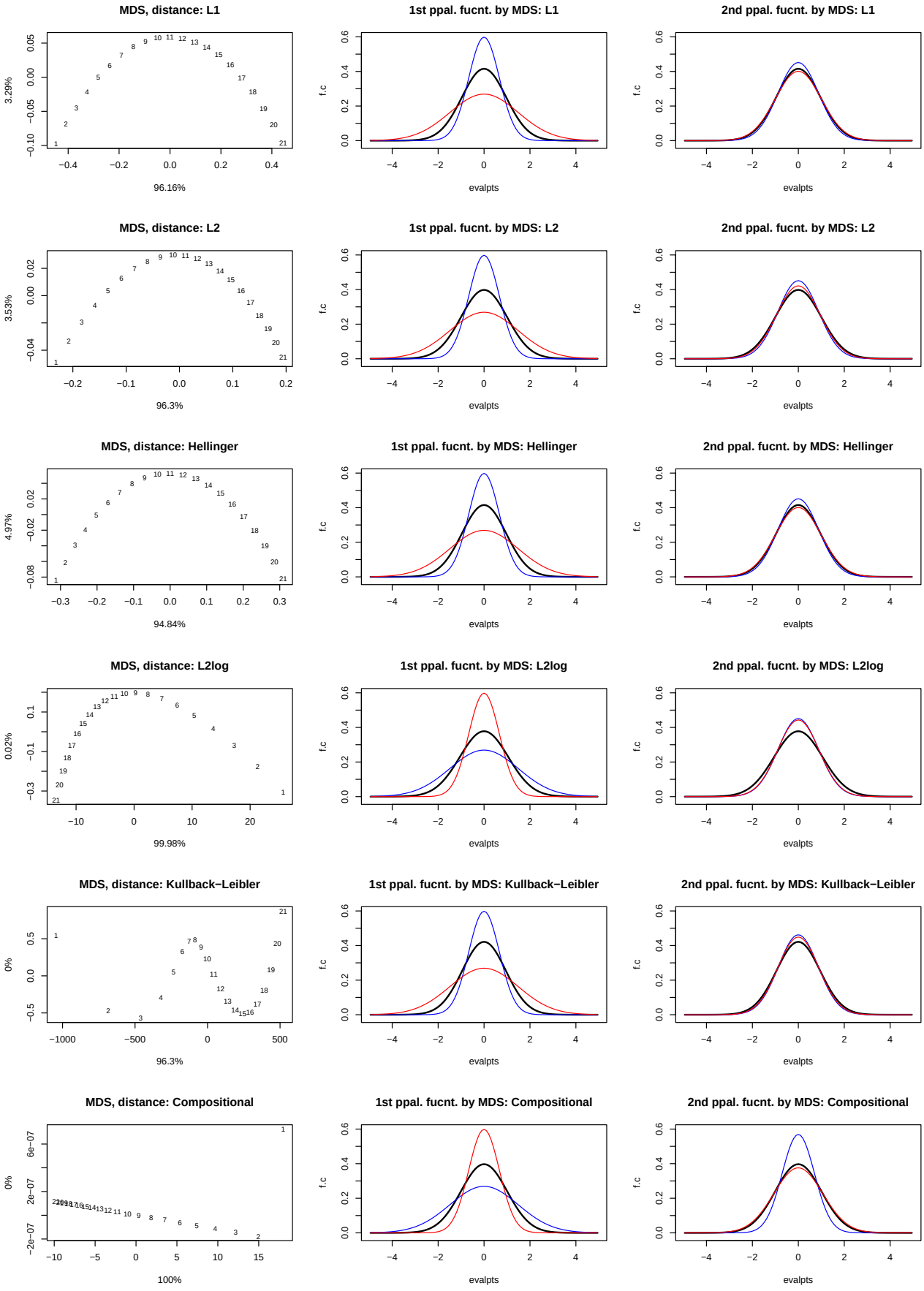


**Figure** 8: MDS for densities in Set 2.

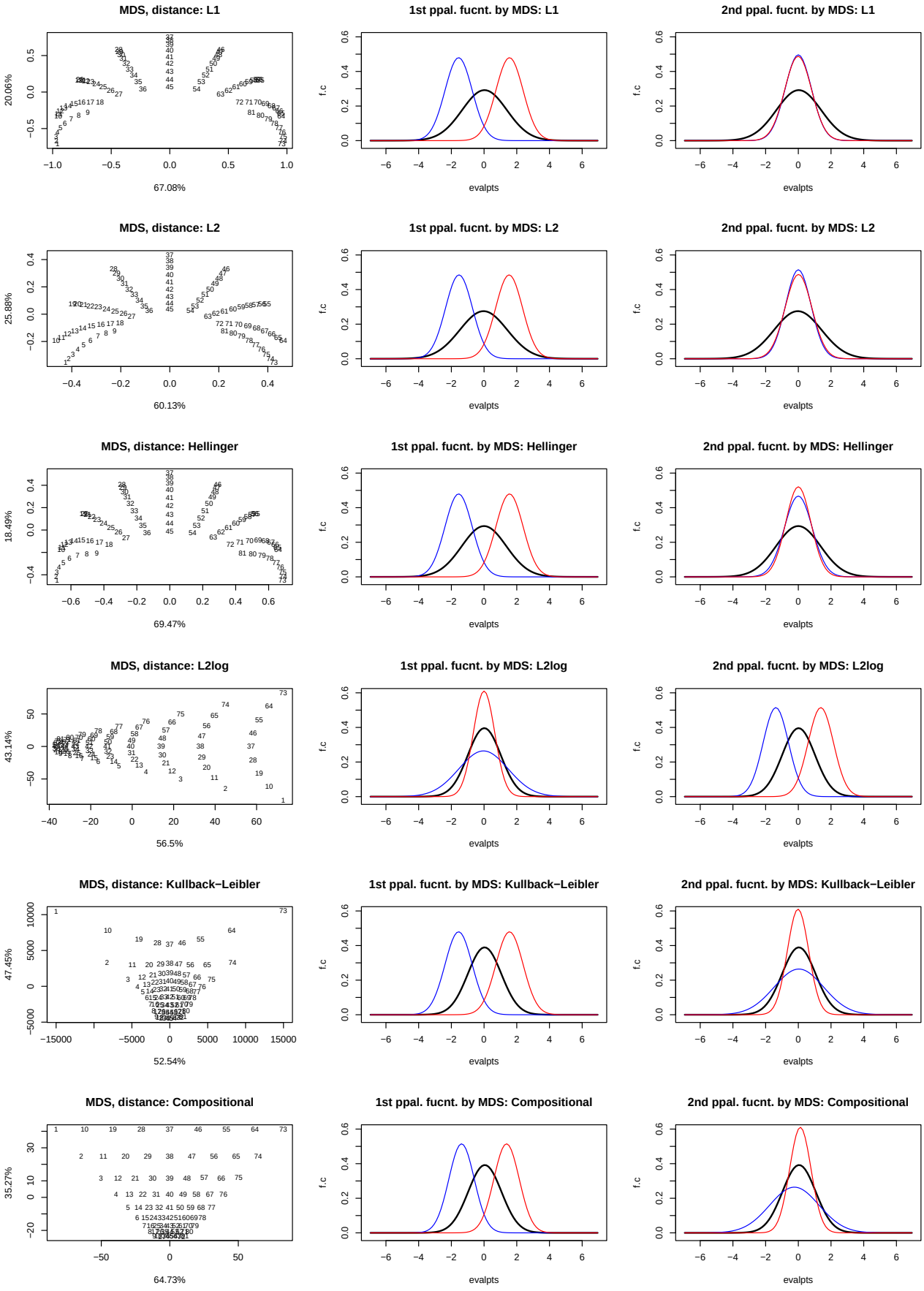


**Figure** 9: MDS for densities in Set 3.

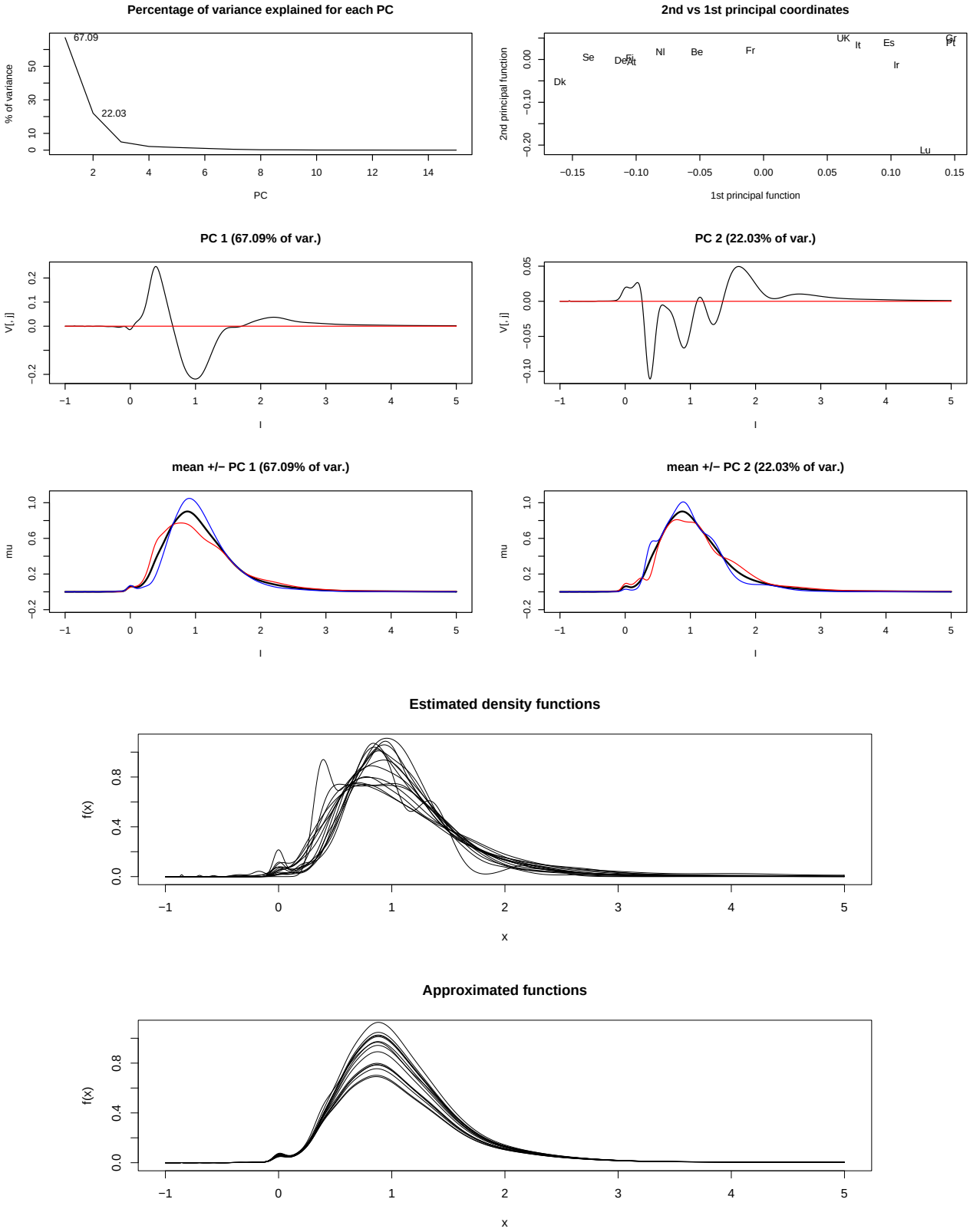


**Figure** 10: FPCA for household income densities for European countries.

shows the results. As it happens with artificial data, distances $L_1$, $L_2$ and Hellinger give similar results to those of FPCA. The results derived from the Symmetrized Kullback-Leibler divergence are very different, indicating that the set could be considered one-dimensional. The interpretation of the first principal direction is not clear.

## Acknowledgements

Work supported in part by the Spanish Ministerio de Educación y Ciencia and FEDER grant MTM2006-09920. I'm grateful with Magda Mercader that provided me with the relative equivalent disposable income data.

# REFERENCES

Borg, I. and P. Groenen (2005). *Modern Multidimensional Scaling: Theory and Applications (2nd ed)*. New York: Springer-Verlag.

Bowman, A. and A. Azzalini (2001). *Applied Smoothing Techniques for Data Analysis*. Oxford: Oxford University Press.

Davidian, M., X. Lin, and J. Wang (2004, July). Introduction: Emerging issues in longitudinal and functional data analysis. *Statistica Sinica 14*, 613–614.

Delicado, P. (2007, September). Functional $k$-sample problem when data are density functions. *Computational Statistics 22*(3), 391–410.

Egozcue, J. J., J. L. Díaz-Barrero, and V. Pawlowsky-Glahn (2006, July). Hilbert space of probability density functions based on aitchison geometry. *Acta Mathematica Sinica 22*(4), 1175–1182.

Ferraty, F. and P. Vieu (2006). *Non parametric functional data analysis. Theory and practice*. Springer.

González-Manteiga, W. and P. Vieu (2007). Editorial: Statistics for functional data. *Comput. Stat. Data Anal. 51*, 4788–4792.

Kneip, A. and K. Utikal (2001). Inference for density families using functional principal component analysis. *Journal of the American Statistical Association 96*, 519–542.

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

Peña, D. (2002). *Análisis de datos multivariantes*. McGraw-Hill. (In Spanish).

R Development Core Team (2005). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. ISBN 3-900051-07-0, http://www.R-project.org.

Ramsay, J. and B. W. Silverman (1997). *Functional Data Analysis*. New York: Springer.

Ramsay, J. and B. W. Silverman (2002). *Applied Functional Data Analysis: Methods and Case Studies*. New York: Springer-Verlag.

Ramsay, J. and B. W. Silverman (2005). *Functional Data Analysis* (Second ed.). New York: Springer.

Valderrama, M. J. (2007, September). Editorial: An overview to modelling functional data. *Computational Statistics 22*, 331–334. Special issue on Modelling Functional Data in Practice.

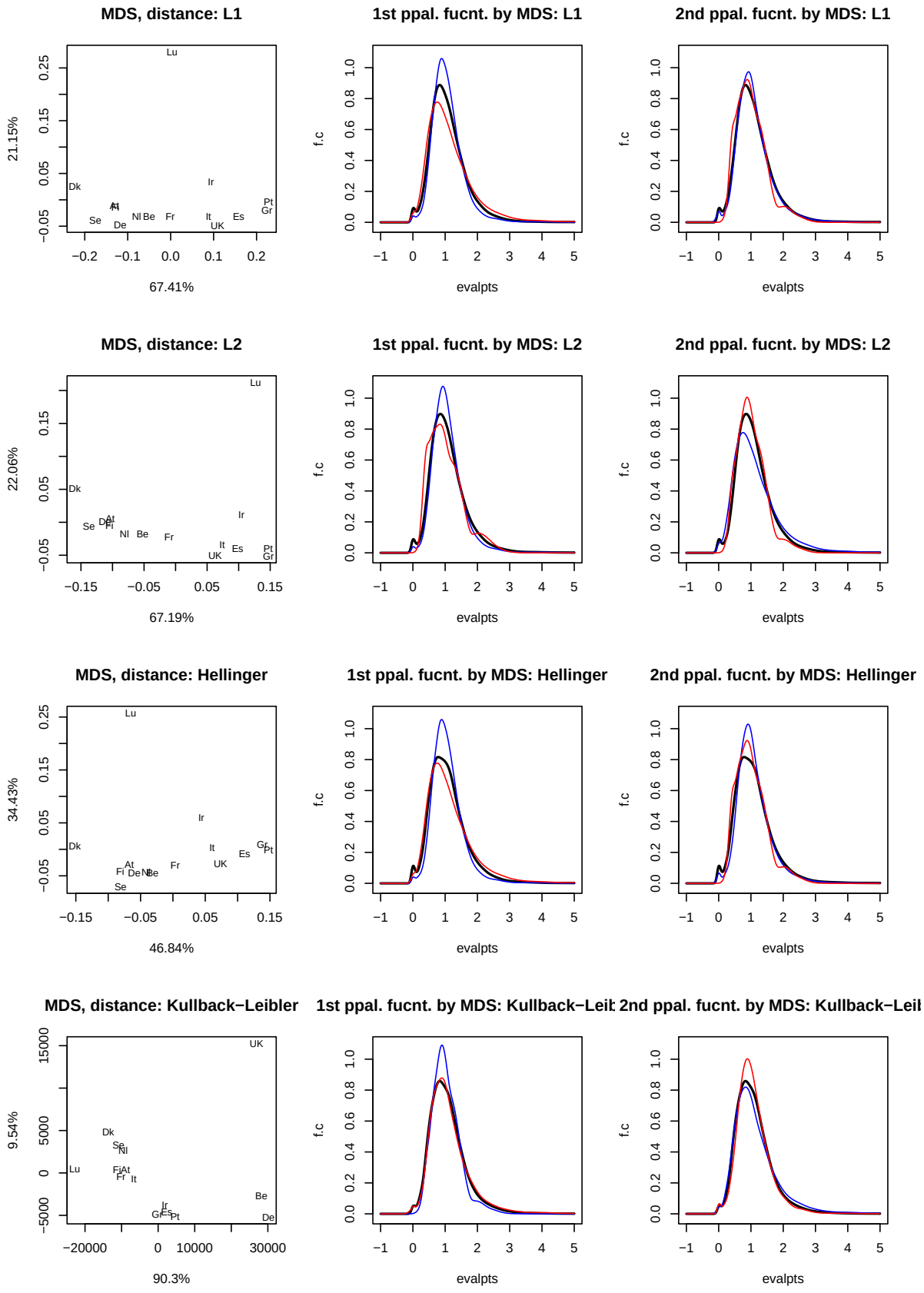


**Figure** 11: MDS for household income densities for European countries.