Functional data analysis: Introduction and applications to financial electricity contracts

Jostein Lillestøl Fridthjof Ollmar

Department of Finance and Management Science
The Norwegian School of Economics and Business Administration
Bergen, Norway

E-mail: Jostein.Lillestol@nhh.no, Fridthjof.Ollmar@nhh.no

Jostein Lillestøl
Fridthjof Ollmar

Functional data analysis: Introduction and applications to financial electricity contracts

1. Introduction

Financial decisions and pricing of assets are often based on process models in continuous time. However, the data analysis of financial time series is typically based on either discrete models or discrete sampling from a chosen process model. In many situations this choice is made by convenience, e.g. to have a simple interpretable model with nice analytical properties. Each model may be compared with competing models with respect to the fit to data, simplicity and ease of implementation. However, we often have vague ideas of a suitable model and the aim of the data analysis is to explore the data in order to reveal some aspects of the dynamics of the underlying process. The objects of study are essentially functions of time, and this should ideally also be the context of the data analysis. Such a context is developed during the 1990's by Ramsey and Silverman, among others.

In the common multivariate context the object under study is a data matrix

$$\{x_{ij}; i = 1,2,...,I, j = 1,2,...,J\}$$

while, in the functional context, the object under study is a set of functions defined over an interval, say [0, T].

$$\{x_i(t); i = 1,2,...,I; 0 \leq t \leq T\}$$

In finance we may consider prices of I assets observed at J instances or over continuous time. In multivariate analysis the two subscripts typically refers to I repeats of J variables, where the order of the variables does not matter. If order matters, as with time, more structure is added and common multivariate methods are not necessarily adequate. In problems involving time the label "repeat" will depend on the context. If the focus is on the correlation between prices of various assets, then repeats are observed through time to get better estimates of this correlation, and the possible correlation over time is then a nuisance, which has to be accounted for by choosing an appropriate (multivariable) time series model. On the other hand, if focus is on the dynamics of the financial market, then we may regard the observation of several price series through time as repeats to get a better insight to the dynamics, say by using some kind of functional principal component method. This is in fact one of the new developments in functional data analysis, which we will return to in the sequel.

Software for functional data analysis is developed by Ramsay and Silverman and freely available on the net from http://www.psych.mcgill.ca/faculty/ramsay/fda.html. They provide functions for R, S-PLUS versions 3.4, 5, and 6 (Unix) and S-PLUS 2000 and 6 (Windows 98/2000/NT), and also for MATLAB, versions 5 and 6. They also provide written programs for various examples in their book, which can easily be modified to accommodate own needs.

2. Theory

We will subsequently imagine the functions under study as processes in continuous time over an interval, [0, T]. In practice they are observed at discrete instants t_j ; j = 1, 2, . . . , J. We will here assume that all processes are observed at the same instants, but in principle they can be different, i.e. at instants t_{ij} depending on i. Our observations are therefore

$$\{x_i(t_j); i = 1,2,...,I; j = 1,2,...,J\}$$

which can be organized in an $I \times J$ matrix.

The first step in a functional data analysis is to convert the data to functional form, so that each function can be evaluated for all values of t in the interval in question. The functions thus obtained are usually referred to as *functional data objects*. This conversion is done by representing the data by interpolation or smoothing in a number of different ways, most notably by choosing a set of basis functions appropriate for the problem at hand. The functions are then expressed by

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$$

where $\{\phi_k(t)\}$ is the set of basis functions. The conversion of data into a functional data object amounts to compute and store the coefficients $\{c_{ik}\}$ of the basis representation in an I x K matrix. The most common basis functions are

- Fourier basis
- B-spline basis
- Monomial basis
- Exponential basis
- Wavelet-basis
- Constant basis

The first two are the most frequently used, the Fourier basis being appropriate for periodic data. A basis is specified by four elements

- Type of basis
- Range of argument values $[T_0, T_1]$
- Number of basis functions (K)
- Basis parameters (depend on basis)

The range is the interval over which we want the functional object to be evaluated.

The Fourier basis for periodic data consists of the functions 1, sin ωt, cos ωt, sin 2ωt, cos 2ωt, ... where $\omega = 2\pi / S$ and S is the period. For instance, S=12 for data observed on a monthly time scale with yearly periodic variation, like temperature. The B-spline basis consists of piecewise polynomials and requires specification of the order O and the location of break points (knots), often chosen at the sampled points of the data. Here O = 2 corresponds to line segments and O = 3 quadratic segments etc. To some extent the basis representation is a compromise between interpolation and smoothing of the original data. However the variability in the raw data should be transferred to the functional data object. If smooths are needed, say for modelling or prediction purposes, this comes in the later steps of the functional data analysis.

More specifically if $\{x_{ij}\}$ represents the observation of the i ´th function at instant t_j we may obtain the basis representation of $x_i(t)$ by a least squares criterion, that is minimization for each i

$$\sum_{j=1}^{J}\left(x_{ij} - \sum_{k=1}^{K} c_{ik}\phi_k\left(t_j\right)\right)^2$$

with respect to the c_{ik}'s (Here $K \le J$). The smoothness of fit is determined by K, a small/large K gives smooth/less smooth fit respectively.

Given a functional data object

$$x_i(t) = \sum_{k=1}^{K} c_{ik}\phi_k(t)$$

we can modify this to $\tilde{x}_i(t)$ by smoothing. This can be done by a least squares criterion with an added roughness penalty, that is, by minimization for each i

$$\int \left(x_i(t) - \tilde{x}_i(t) \right)^2 dt + \lambda\, g\left(\tilde{x}_i(t) \right)$$

where g (x) is a measure of roughness and λ is the weight to be put on smoothness, more smooth for increasing λ. A common measure of roughness is based on the second derivative function of the functional object and given by

$$g(x) = \int \left(x''(t) \right)^2 dt$$

often referred to as the *total curvature* of x(t). The smoothed objects can be in the span of the same basis (of high dimension wanted) or another basis (if wanted). The penalty function can typically be expressed as a quadratic form involving the basis coefficients and the minimization is within the realm of linear algebra.

Note that by representing the function as a functional data object, we can easily compute derivatives. However, even if the object itself is reasonably smooth, the derivative may be erratic, unless we do additional smoothing, taking second order derivatives into account.

In some situations it is advantageous to use two sets of basis functions that complement each other, a small set $\{\phi_k\}$ of K_1 elements able to capture the large-scale features in the data and a larger set $\{\psi_k\}$ of K_2 elements that can capture local and specific features. The idea is then to smooth by penalizing only the second part in

$$x_i(t) = \sum_k c_{ik}\phi_k(t) + \sum_k d_{ik}\psi_{ik}(t)$$

Again such penalizing can be expressed in matrix terms. For example if we have monthly data with a yearly season, it makes sense to take $K_1 = 3$ with basis functions $1, \sin(\omega t), \cos(\omega t)$ with $\omega = 2\pi/12$. A possibility is to use further Fourier series terms as the second set of basis functions. $\mathrm{K}_2 = 9$ of them are sufficient to interpolate the data perfectly $(\mathrm{K}_1 + \mathrm{K}_2 = 12)$.

An important tool in the functional data analysis toolbox is PCA, i.e. *principal component analysis*. The main idea is as for multivariable data, but contrary to this, principal components weight or harmonics now are functions of time. They carry the main features of the functional data object, so that each harmonic can, in some sense, be interpreted separately. Consider an integral transform of all $x_i(t)'s$

$$y_i = \int \xi(t) x_i(t) dt$$

where the common weight function $\xi(t)$ is determined by minimizing

$$\sum_{i=1}^{I} y_i^2$$

within a space of functions subject to the constraint $\int \xi^2(t) dt = 1$. The minimizing function is denoted $\xi_1(t)$ and called the *first principal component function* (or *first harmonic function*). The corresponding minimizing y_i's are called the *principal component scores*.

Consider the minimizing problem again to obtain another weight function $\xi_2(t)$ subject to the added orthogonal restriction $\int \xi_1(t)\xi_2(t) dt = 0$. This is the *second principal component function*. This process can be continued to get a sequence of principal component functions (or harmonics) $\{\xi_k(t); k = 1,2,...,K\}$ orthonormal to one another. Here K must be less than I. Each harmonic accounts for a certain portion of the variability in the original functional data objects, decreasing for increasing k.

Another way to introduce functional principal components is to find least squares approximating functions in terms of an orthonormal basis $\{\xi_k(t); k = 1,2,...,K\}$, i.e.

$$\hat{x}_i(t) = \sum_{k=1}^{K} y_{ik} \xi_k(t)$$

so that

$$\sum_{i=1}^{I} \int \left(x_i(t) - \hat{x}_i(t) \right)^2 dt \qquad \text{is minimized.}$$

The optimal choice of this basis is identical to the solution of the problem above. Since these functions are determined by the data they are often named *empirical orthonormal functions.*

The principal curve determination is not unique and solutions that are just as good, but perhaps more easily interpreted, may be obtained by rotation. Mathematically the principal component problem above leads to a so-called *functional eigenvalue problem* for the covariance function

$$v(s,t) = \frac{1}{I} \sum_{i=1}^{I} x_i(s) x_i(t)$$

The numerical implementation of this is based on an approximation that brings the problem into the realm of standard linear algebra. The simplest approach is by discretization of the functional objects to a fine grid spanning the interval in question, leading to a standard matrix eigenvalue problem. The solution can then be brought back to continuous time by a suitable interpolation method, which one does not matter if the grid is sufficiently fine, usually much finer than the data itself.

Using the basis function representation of the functional objects

$$x_i(t) = \sum_{k=1}^{K} c_{ik} \phi_k(t)$$

we can express the covariance function v(s, t) in terms of the vector of basis functions $\{ \phi_k(t); k = 1,2,..., K \}$ and the matrix of coefficients C = (c_{ik}). The corresponding representation of the eigenfunction

$$\xi(t) = \sum_{k=1}^{K} b_k \phi_k(t)$$

can be written in matrix notation as well. The functional eigenvalue problem can then be expressed by a matrix equation with functional elements, which simplifies to a pure matrix equation within the realm of standard matrix algebra.

The number of eigenfunctions obtainable is limited by K. The above does not necessarily assume that the basis is orthonormal, but in that case the computation is just ordinary eigenvalue/vector computations on the matrix C'C. In some cases the harmonic functions may come out fairly erratic, and it might be tempting to smooth them in order to get a more clear cut picture of what is going on. However, the desire to get smooth harmonics can alternatively be fulfilled by using the penalty approach, i.e. adding extra regularity condition to the minimization problem, for instance on the derivatives. This gives rise to a modified eigenvalue problem, which together with modified orthonormal conditions, can be solved by standard matrix analysis, provide we express the problem in terms of the chosen basis. In particular this is so for periodic situations where we typically use the Fourier basis. The penalty approach requires specification if a smoothing parameter that controls the desired smoothness. Alternatively this can be determined by *cross validation*.

An alternative to the penalized harmonic approach is to start with a functional data object that is smoothed prior to the principal component analysis. Discussions on the benefits and drawbacks of each approach are given in the literature.

3. Applications

We will now give some examples of the application of functional principal component analysis to financial contracts in the Nordic electricity market. Our main objective is to illustrate the main features of the functional principal component method. We hope that the various plots will expose known dynamics of the market, and perhaps also raise some questions for discussion. The examples involve futures and forward contract. A forward contract is an agreement today to deliver a commodity at a future maturity date, in the case of electricity 1MW during a settlement period. A futures contract is similar to forward except that changes in value are balanced on an account over time.

Example 1: Term structure for electricity base load contracts in the year 2001.

The input data is a 10 by 12 matrix of monthly prices based on the following 10 contract categories: one future contract (GB13-01) and nine forward contracts (FWV1-02, FWSO-02, FWV2-02, FWV1-03, FWSO-03, FWV2-03, FWYR-02, FWYR-03, FWYR-04). Here the coding refers to the year and season for delivery (V1=Winter-spring, SO=Summer, V2=Winter-fall, YR=Year). The data is average prices recorded on the first trading day of the month, averages taken over all contracts with the given delivery period. The prices of the year contracts is expected to come out as an average of the three seasonal ones, and will perhaps just duplicate the information.

The observed (smoothed) price functions are given in Figure 1.1.

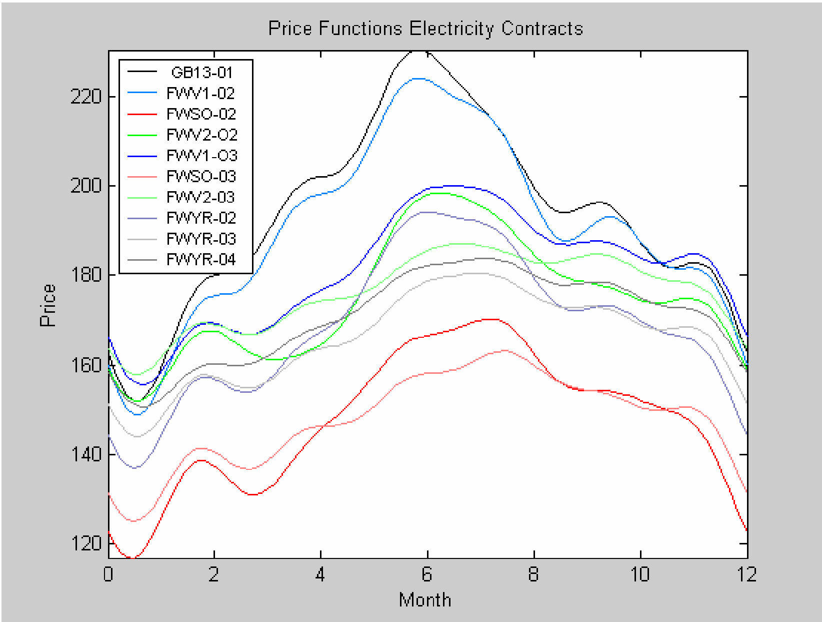

Figure 1.1 Observed price functions for the electricity base load contracts year 2001.

The level of individual curves reflects the seasonal patterns of the spot price, the summer delivery contracts having a lower value than the winter ones. The apparent common appearance over the year should not be taken as a periodic yearly phenomenon. The reason for the high prices in the middle may be due to the low reservoir fillings combined with less than normal snowfall in the spring of 2001. This will presumably affect the contracts with shorter time to maturity more than the longer ones.

In Figure 1.2 we illustrate the variance-covariance structure of prices, the peak at the middle of the diagonal representing the largest variance of the midsummer prices.

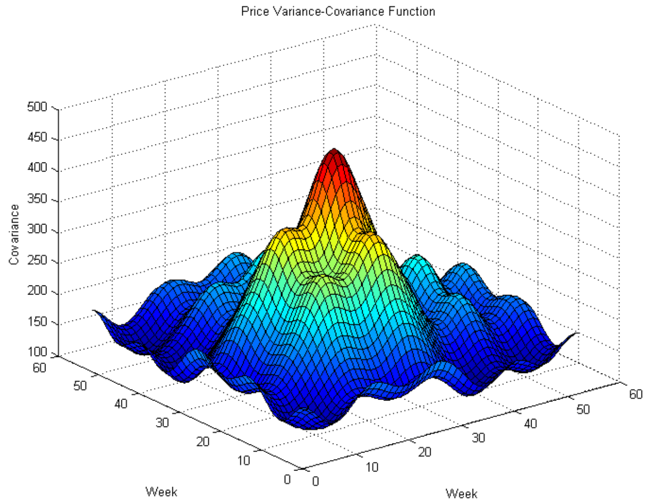

Figure 1.2. Variance-covariance structure of prices.

The principal component functions are illustrated in Figure 1.3a-c as perturbations of the overall mean function, by adding and subtracting a multiple of each principal component function. We see that the first principal component function, which accounts for 92 percent of the variation, is essentially the common price variation over the year believed to be due to the hydrological balance the year in question. The second principal component function, which accounts for 7 percent of the variation, is seen to pick up the difference between the shorter and longer contracts with respect to the height of the peak summer price. The third principal component function, which accounts for only 1 percent of the variation, could be spurious and should hardly be interpreted. If anything, it could be connected to the revelation of the hydrological balance, i.e. snow melting, which may affect the contracts differently.

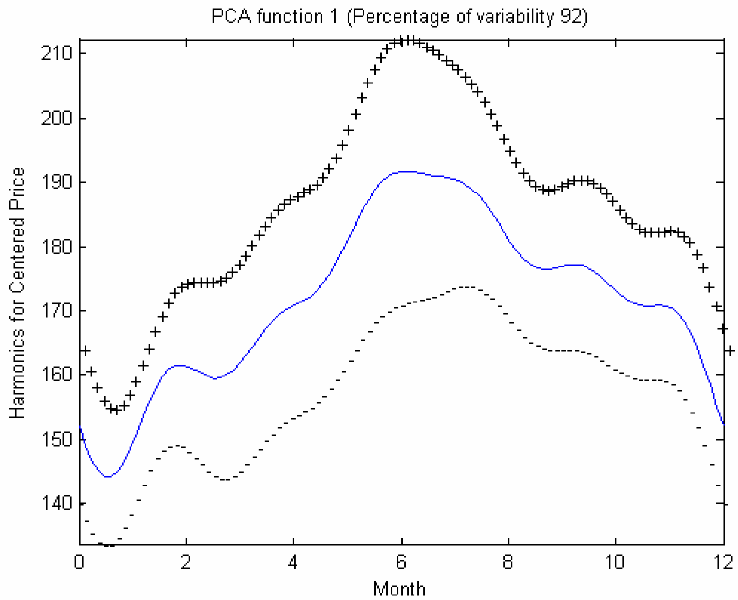

Figure 1.3a First Principal Component Function.

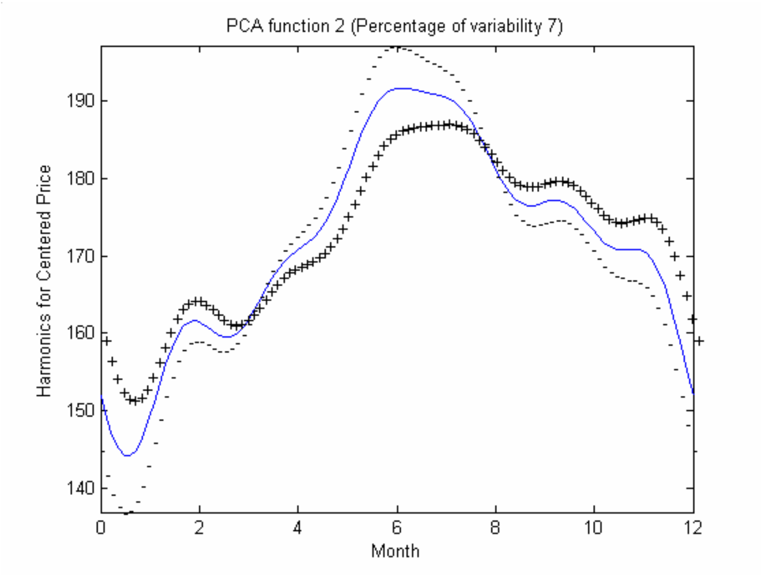

Figure 1.3b. Second Principal Component Function.

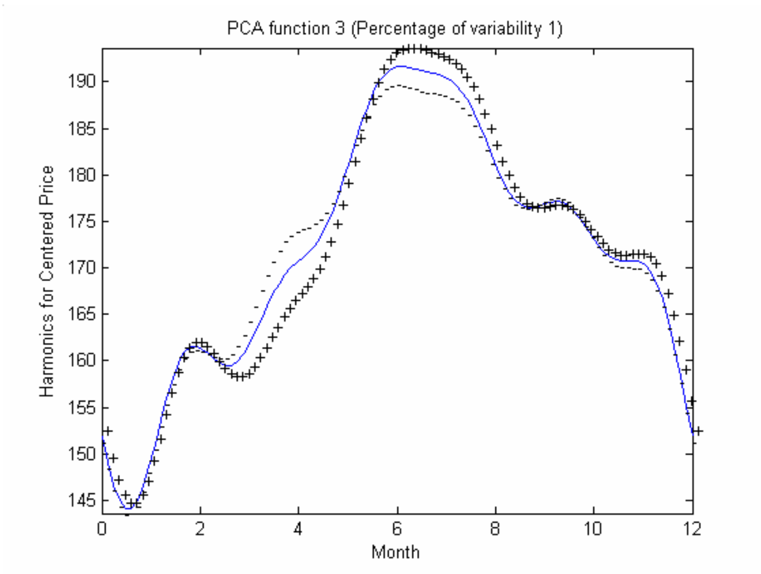

Figure 1.3c. Third Principal Component Function.

In Figure 1.4 we have plotted the scores of the two first principal components. We recognize the approximate ordering of curves from bottom to top of Figure 1 as points with scores from left to right.

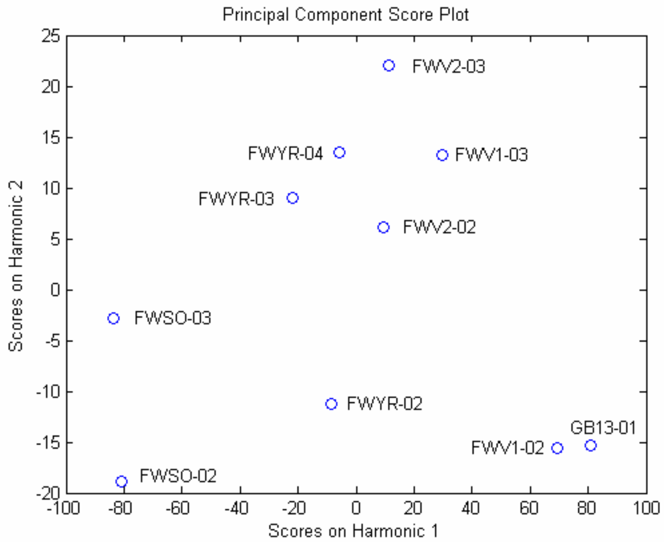

Figure 1.4. Principal Component Score Plot

The analysis above was repeated omitting the yearly contracts. The results were almost indistinguishable from the above (leaving out the yearly points in Figure 1.4).

For comparison we have looked at the corresponding data from year 2000 which shows a weak upward trend with some common fluctuation and a "bump" for some of the contracts in the Fall. As for the year 2001 the first principal component function is the common variation, which this time accounts for as much as 98 percent of the variation. The second component accounts for just 1 percent and accounts the contracts with extra bump in the Fall, which is the only future contract GB13-01 and the forward contract FWV1-01. This may be connected to the revelation of the supplementary rainfall that may affect the contracts first to be delivered.

Example 2 : Electricity futures monthly in year 2000 in term of weeks to maturity

In this example we study the prices of electricity monthly future curves, i.e. an agreement today to deliver 1MWh at a future maturity date (strictly speaking during a settlement period). The prices of various contracts with respect to weeks to maturity are observed each month of year 2000, on the first trading day of the month. Since the marketplace cannot provide data to cover all times to maturity, practitioners construct a smoothed term structure using some prior harmonic assumptions. The input data available and used here are of this kind. The input to our fd-construction (using a B-spline basis) is a therefore a discretized fd-object.

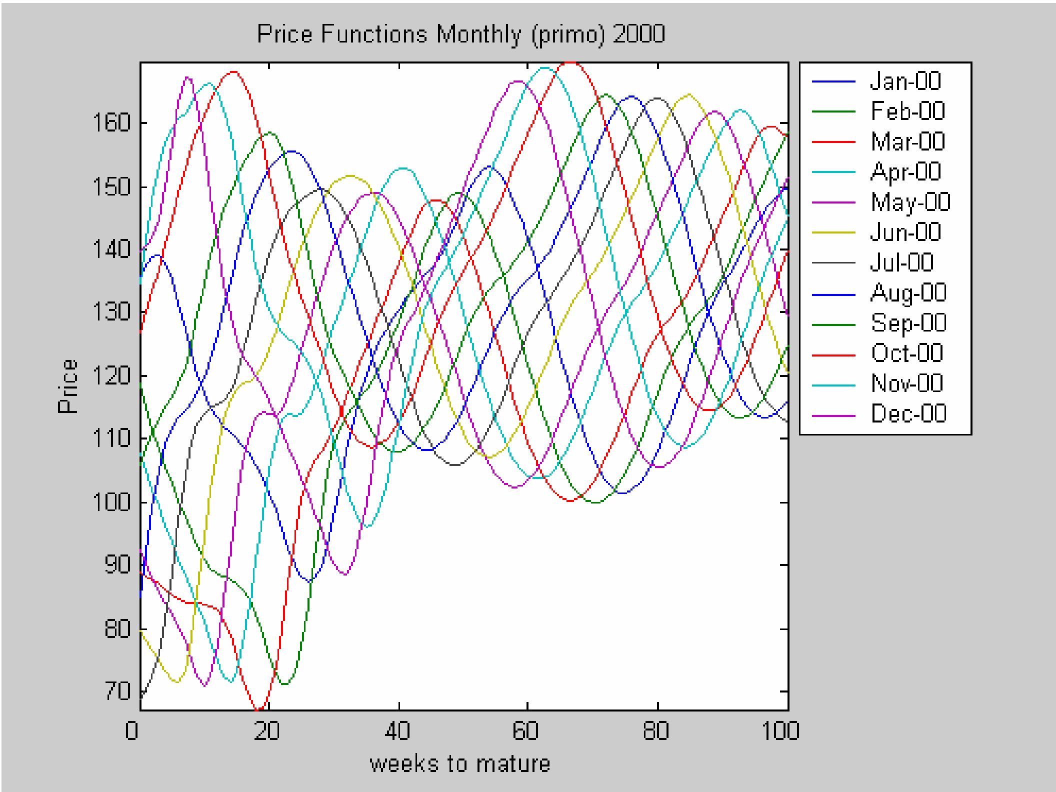

Figure 2.1 Electricity futures monthly in year 2000 in term of weeks to maturity

Figure 2.1 should be interpreted as follows: Each curve represents a functional data object for a specific month of 2000 as seen from the first trading day of the month, and gives the future prices for contracts with different number of weeks to maturity (ranging from 1 to 101). For each curve we see the cyclic variation over the year with highest prices for maturity in January to March. The year 2000 had above normal precipitation in the spring leading to lower prices for the contracts with short time to maturity (left half of figure), while the contracts with maturity the same months one year ahead are not affected (right part of figure).

The first four principal component functions are given in Figure 2.2a-d. The first principal component function in Figure 2.2a accounts for 57% of the variation, and is seen to be similar to a harmonic variation over a one-year period. The second principal component function in Figure 2.2b accounts for 38% of the variation and is also seen to be similar to a harmonic variation over a one year period, a quarter out of phase with the first, as sine versus cosine. Taken together this accounts for 95% of the functional variation and is able to reflect any phase translation of a curve. The third and fourth principal component in Figure 2.2c and d accounts for just 2% and 1% of the variation respectively, which raises the total to 98%. They may be seen as a higher frequency adjustment, which is more pronounced for maturities one year ahead. For short time to maturity the harmonic effect is blurred, and the adjustment tries instead to make up for the lower spring prices in the contracts with short time to maturity. Analysis of the year 1999 shows the same pattern, and the first four principal components accounts for 53%, 40%, 5% and 1% respectively, totalling 99%.

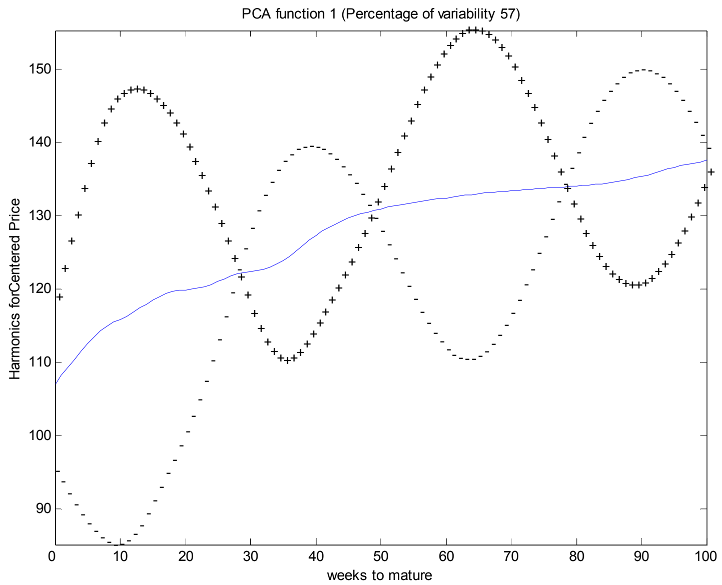

Figure 2.2a First principal component function

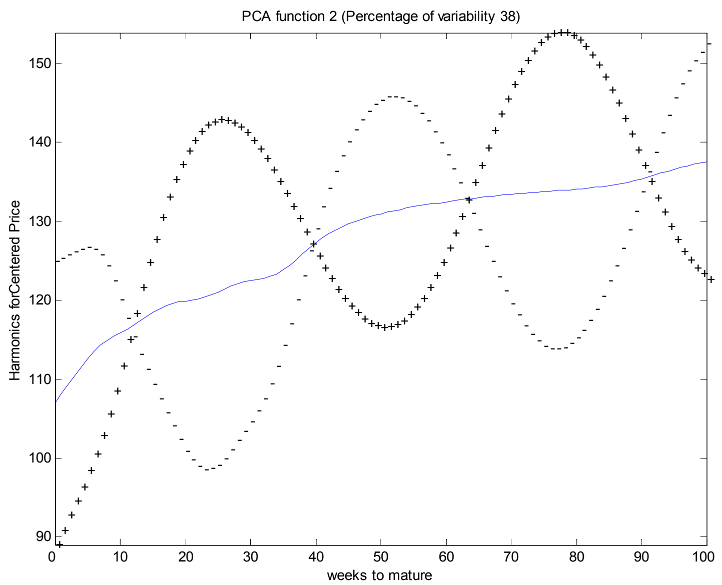

Figure 2.2b Second principal component function

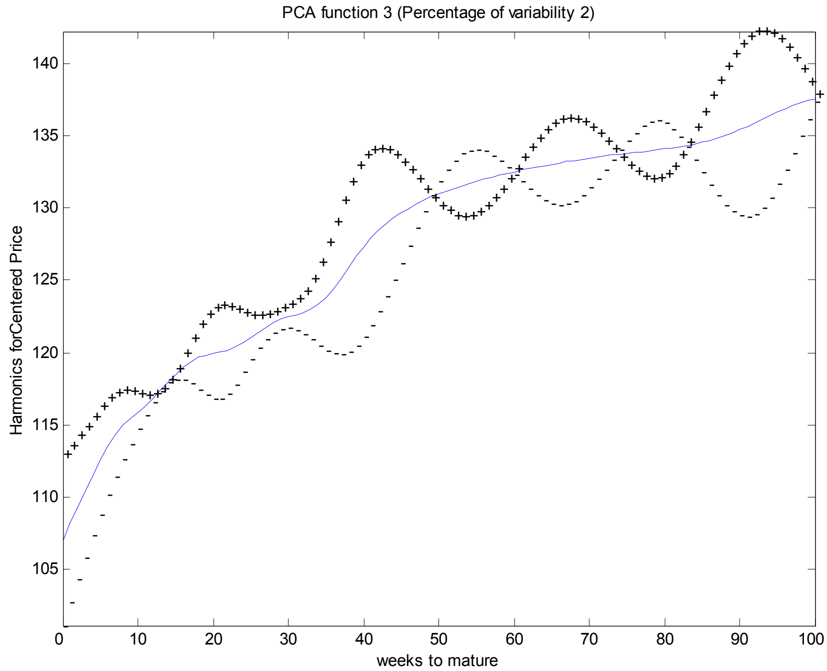

Figure 2.2c   Third principal component function

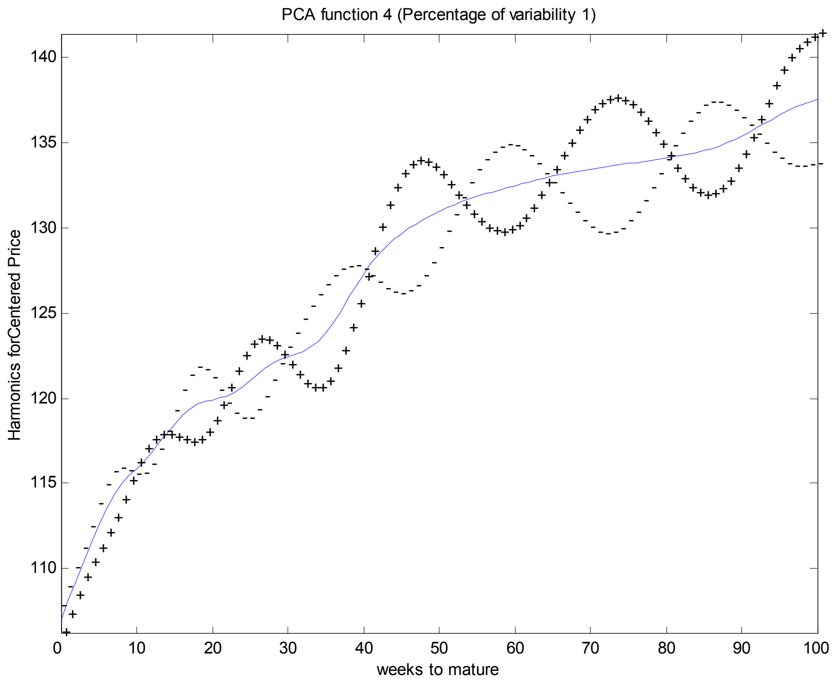

Figure 2.2d   Fourth principal component function

In Figure 2.3a-c we see bivariate plots of the principal component scores for the first three principal components. In Figure 2.3a the second is plotted against the first, and we see that the harmonic features appear clearly, as the months appear in natural order along a near perfect circle. However, there is a danger that we have just rediscovered the harmonic assumptions used in the pre-processing of the data. In Figure 2.3b and 2.3c we see a loop pattern, i.e. like the infinity symbol.

It would be of interest to use data pre-processed by weaker assumptions to see how much the harmonic pre-processing affects the principal components and their interpretations.

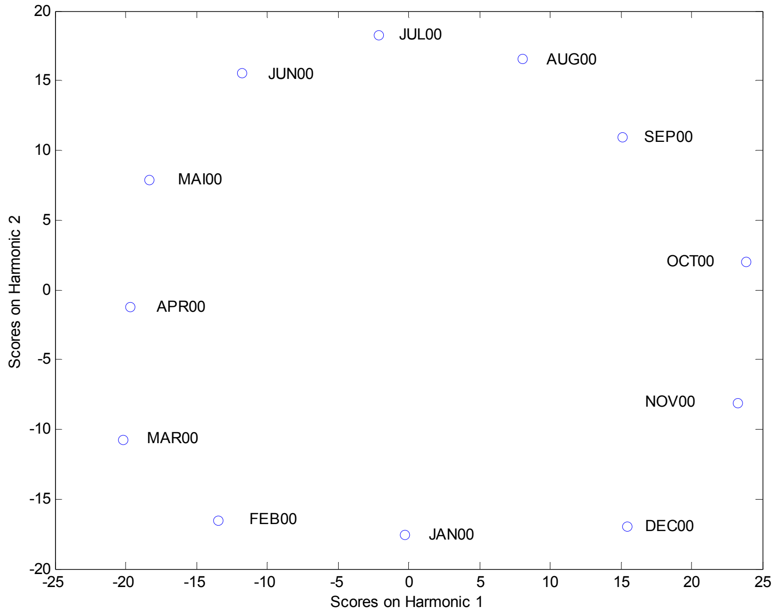

Figure 2.3a Principal component scores of second PC vs. first PC

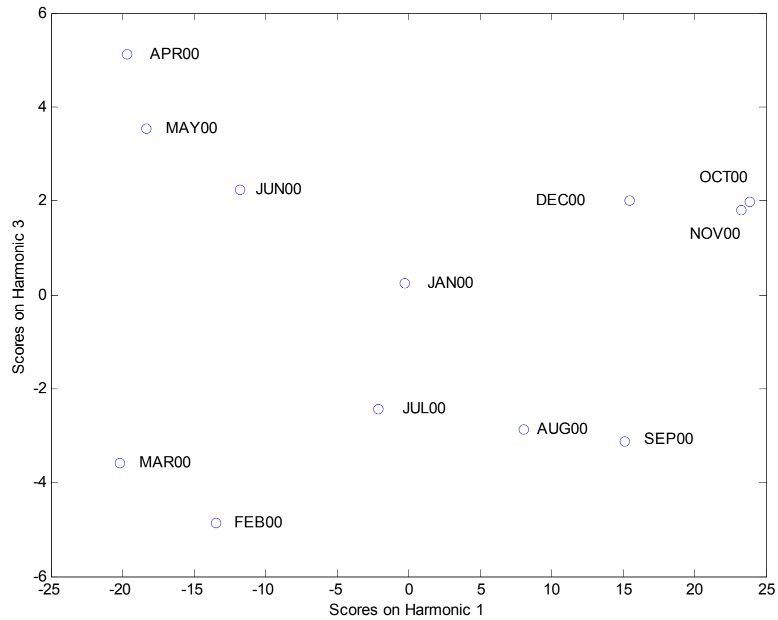

Figure 2.3b Principal component scores of third PC vs. first PC

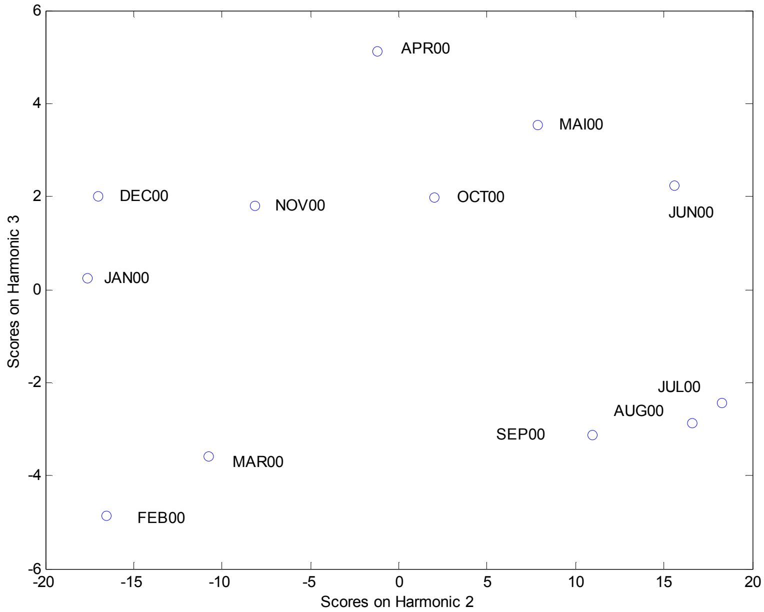

Figure 2.3c Principal component scores of third PC vs. second PC

Example 3 : Electricity futures observed in terms of maturity date

In this example each of 13 curves represent the prices on dates 4 weeks apart in the year 1998, for 50 maturity dates (start of settlement week) 4 weeks apart starting 10.01.00 and ending 13.10.03. The input is therefore a (13 x 50)-matrix, the upper left corner being

	11.01.1999	**08.02.1999**	**08.03.1999**	**05.04.1999**	**03.05.1999**	**31.05.1999**
05.01.1998	194,02	196,71	191,41	180,86	169,41	156,83
02.02.1998	182,99	185,34	180,55	171,50	162,54	152,21
02.03.1998	176,59	179,90	175,43	165,94	155,78	144,31
30.03.1998	171,69	174,02	169,49	160,99	152,88	143,45

Figure 3.1 shows the 13 curves. The dominant feature is the seasonal variation with respect to the settlement date. Note that the variation is increasing for settlement dates two years ahead. and beyond. A possible explanation of this is: While the left hand part of the figure is based on more than one contract a year, the right hand part is constructed on basis of yearly contracts, i.e. contracts with settlement period of one year with typically time to maturity more than two years, thus providing lower resolution. We also see that prices tend to decrease during the sample period, e.g. the curves are approximately ordered from top to bottom with respect to the recording dates according to the legend of the figure. It is also seen that the wider swing for some dates (among them date 13) occur one year later than most of them.

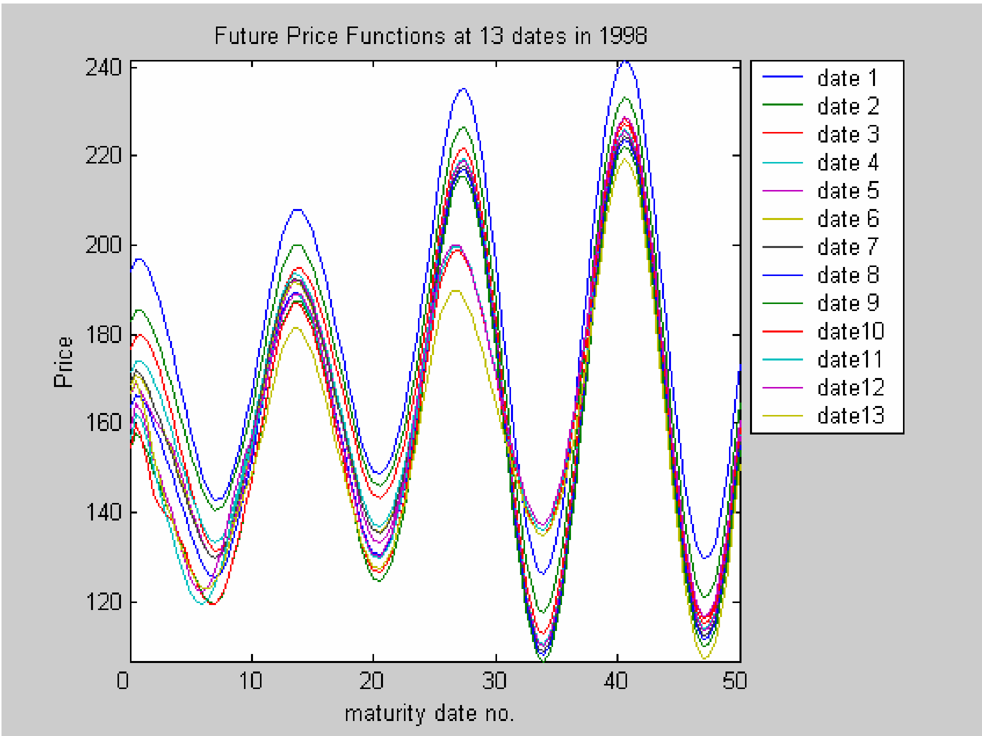

Figure 3.1 Futures Price Functions as seen from various dates of 1998

The first three principal component functions account for 71%, 24% and 3% of the variation respectively, totalling 98%. Figure 3.2a-b shows the first two principal component functions as deviates from the overall mean function. We see that the first component function mainly accounts for the differences in peak and bottom prices, as well as mid-winter to summer prices, mostly for settlement periods two years ahead and less, i.e. the left hand side of the bundle of price curves. The second principal component function accounts for differences in the peak and bottom for settlement dates more than two years ahead. Note that the extra swing at maturity date around 28 is represented in both. The third component, not shown, is just a minor adjustment, and carry no essential information.

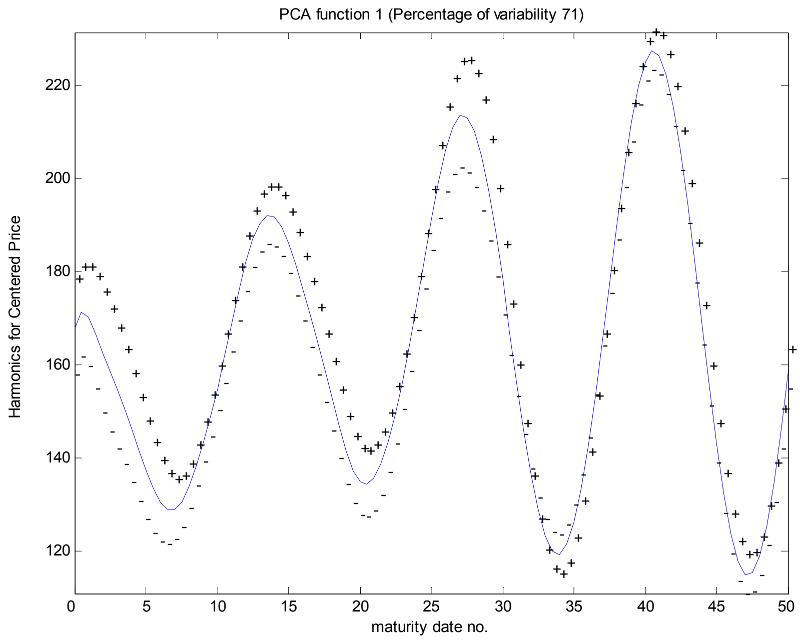

Figure 3.2a First Principal Component Function

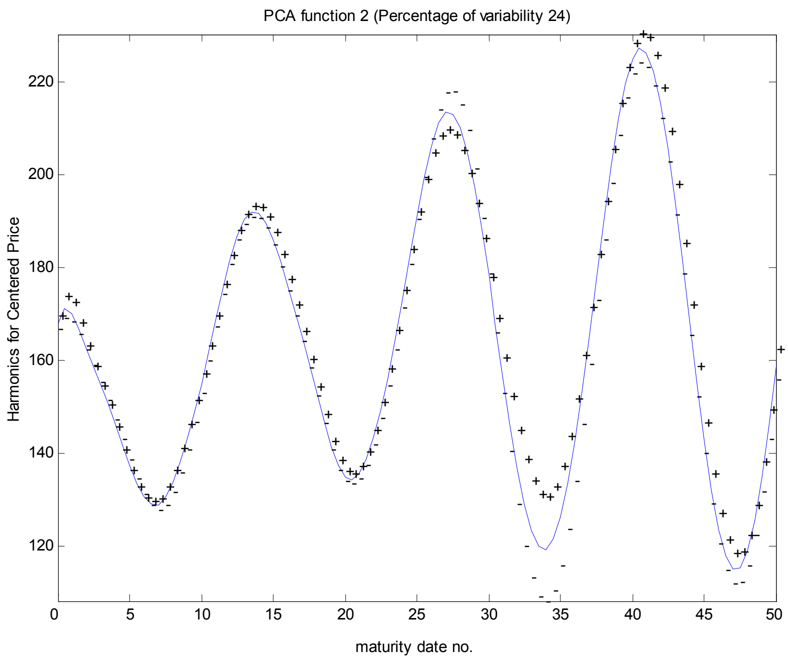

Figure 3.2b Second Principal Component Function

Figure 3.3 show the principal component scores for the second component against the first. We see that there is a clear split between the first 9 sampling dates and the last 4. The first 9 dates are also well ordered with respect to the both principal components, and the last 4 try to return to date1, except a disorder caused by date 13.

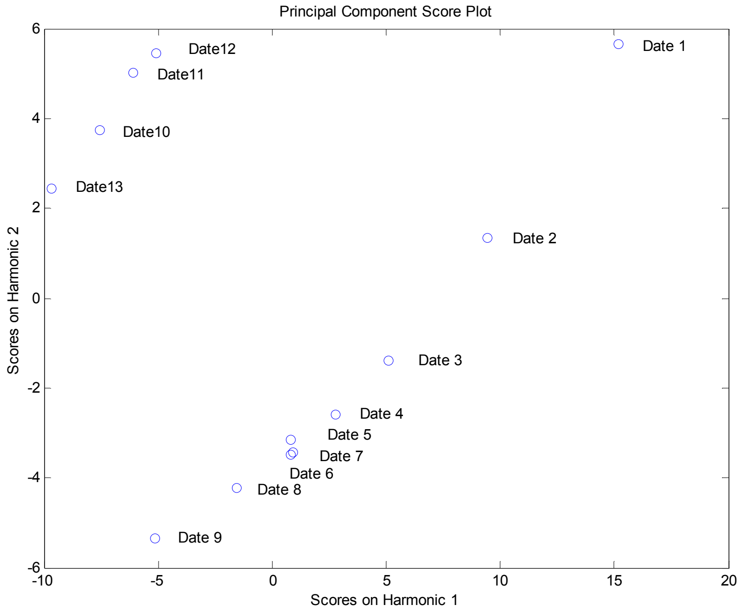

Figure 3.3 Principal Component Scores Plot

Reference:

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