Ledolter J, J Ophthalmic Clin Res 2019, 6: 057
DOI: 10.24966/OCR-8887/100057

HSOA Journal of Ophthalmology & Clinical Research

Review Article

Assessing Trend Changes in Functional and Structural Characteristics: Combining Principal Components Methods and Functional Data Analysis

Johannes Ledolter[1,2*]

[1]*Department of Business Analytics and Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA*

[2]*Center for the Prevention and Treatment of Visual Loss, Iowa City Veterans Affairs Health Care System, USA*

Abstract

We classify subjects into one of several groups based on their time progression. Medical time series are available for each subject, but time series are short and observations are collected at different time periods. Instead of characterizing each subject by its average and slope, we use principal components analysis to determine the appropriate summary features and classify the subjects on their most important principal components scores. Since observations on subjects are collected at different time periods, we first use functional data analysis to transform the irregularly-spaced data into a complete data array with rows representing subjects and columns representing time.

We illustrate the technique - functional data analysis to create a complete data matrix and principal components analysis to determine the most important features for classification - on the average thickness of the retinal nerve fiber layer that come from two groups of patients.

Keywords: Classification; Functional data analysis; Neuro-ophthalmology; Principal components; Time series; Trend detection

***Corresponding author:** Johannes Ledolter, Department of Business Analytics and Department of Statistics and Actuarial Science, University of Iowa, Iowa City, IA 52242, USA, E: Mail: johannes-ledolter@uiowa.edu

Citation: Ledolter J (2019) Assessing Trend Changes in Functional and Structural Characteristics: Combining Principal Components Methods and Functional Data Analysis. J Ophthalmic Clin Res 6: 057.

Received: August 30, 2019; **Accepted:** September 12, 2019; **Published:** September 20, 2019

Mathematics Subject Classification: 62M10, 62H25, 62P10

Introduction

In reference [1] we discuss how to 1) Best analyze multiple short time series of functional and structural characteristics collected on different groups of subjects and (2) Classify subjects on the basis of their time progression. These methods help us assess the progression of measurements taken over time, such as the visual acuity of normal and glaucoma subjects (measured by either a visual field test or by Optical Coherence Tomography (OCT) measurements on the optic nerve) or the weight gains of children on different diets. Common difficulties with the analysis of such data are the relatively short length of each time series and the fact that measurements are taken at time periods that vary across subjects.

In reference [1] we consider random effects (repeated measurement) time-trend models of the form $Y_t = \alpha + \beta Time_t + noise_t$ and discuss how such models can be estimated on time series observations from multiple subjects of several groups. Model results can be used to test whether or not the average slopes (or the variances of slopes across subjects) of different groups are the same.

We also discuss how subject-specific least squares estimates of intercept and slope in the trend model can be used to classify subjects into one of several groups. Such classification can be carried out with Fisher's linear or quadratic discriminant functions [2,3], or with a non-parametric support vector machine algorithm [4,5]. However there are shortcomings with this approach as it assumes that the investigator has already decided on the intercept and slope as the relevant summary statistics for a subject's time progression. An alternative strategy that is discussed in this paper determines the relevant summary features through principal components analysis and classifies the subjects on their scores that result from the most important principal components. Since observations on the subjects are usually not taken at the same time, we first need to transform the irregularly-spaced data into a complete matrix of observations with rows representing subjects and columns representing time. We use functional data analysis techniques to construct this data matrix. We discuss the methods in Section 2, and illustrate the analysis in Section 3 on OCT data collected on normal and glaucoma subjects.

A Nonparametric Approach for Classifying Time Series Data

Reducing the dimensions of a time series through principal components and classifying subjects on their implied principal components scores

The goal is to classify subjects on their time progression. Imagine the case when we have observations on m subjects taken at the very same T time points. If T is large (say observations at monthly intervals over a period of 10 years, and hence $T = 120$), the classification takes

place in a very high-dimensional (120-dimensional) feature space. A way to simplify the problem is to reduce the *T* observations to a smaller number of subject-specific summary features and classify the subjects on their summary features. Summary features can be pre-selected, such as the intercept and the slope of the trend regression considered in reference [1]. Another approach is to treat the measurements at the *T* time periods as features, determine the principal components, hence reducing the dimensionality of the classification by classifying the subjects on their scores that are implied by the most important principal components. The (length-standardized) weights of the first principal component represent the coefficients in the linear combination of the *T* measurements that has the largest variance. The weights of the second principal component determine the linear combination that has the largest variance among all linear combinations whose weights are orthogonal to the weights of the first principal component; and so on. The weights are also referred to as the principal component loadings on the features. The mathematics is rather simple and involves the calculation of the largest eigenvalues of the *T* x *T* sample covariance matrix (which is estimated from the time series data on the *m* subjects). The ordered eigenvalues of the covariance matrix, $d_1, d_2, \ldots$, represent the variances that are explained by the first, second, … most-important principal components. And the corresponding eigenvectors are the weights (loadings) of the principal components. Note that the rank of the $T \times T$ sample covariance matrix can never be larger than $\min(m-1, T)$. For $m \leq T$, the sample covariance matrix is singular and at least $T - m + 1$ of its eigenvalues are zero.

We are dealing with the time progression of multiple subjects when measurements on all subjects are taken at the same time points. Graphically, the information can be visualized as follows. We start by drawing the mean curve, which we obtain by averaging the responses across the *m* subjects and at each of the *T* time points and by connecting the successive averages. A multiple of the first principal component weights (with multiple $2\sqrt{d_1}$, where d_1 is the variance explained by the first principal component) is added to and subtracted from the mean curve, resulting in two (a lower and an upper) PCA weight curves for the first principal component. Assume, for illustration, that there is a considerable subject effect to the time progression and suppose that the time progression curve shifts up or down depending on the subject. In this case the *T* x *T* correlation matrix has large positive correlations at all off-diagonal elements (as being above average at a particular time point means also being above average at other time points). The weights of the largest principal component in this case are all equal, implying that the first principal component score for a subject is the average of its *T* observations. For this particular illustration the resulting graph of the (lower and upper) first PCA weight curves reflects a band of equal width around the mean curve, with bands far from the mean curve if the principal component explains much of the total variance. In other words, a graph with such "banded" PCA weight curves suggests that subject levels (averages) carry useful information for classification. [Mathematically, for a *T* x *T* matrix of all ones, the largest eigenvector is *T* while all other eigenvectors are 0, and the eigenvector corresponding to the largest eigenvector is proportional to the vector of all ones].

As a second illustration, assume that the time progression curves of subjects are linear, but with different subject-specific slopes. In this particular case the weights of the (first) principal component are linear in time with mean zero. Adding and subtracting $2\sqrt{d_1}$ multiples of the principal component weights from the mean curve results in two crossing lines, and the closeness of the lower and upper PCA weight curves from the average curve reflects how much variability is explained by the principal component. Because of the linearity of the weights, a subject's principal component score assesses the magnitude of its slope (but is unaffected by the level)."Crossing" PCA weight curves are an indication that there is useful information in the slopes. The usefulness of the corresponding PCA scores for classification purposes will be small if the lower and upper PCA weight curves are close to the average curve. [Mathematically, consider the time trend $Y_t = c + \beta(t - \bar{t})$, for $t = 1,2,\ldots,n$, and let the slope β vary across subjects with $Var(\beta) = \sigma_\beta^2$. In this case the *T* x *T* covariance matrix of the observations has rank 1 and the eigenvector corresponding to the one positive (largest) eigenvector is proportional to $(t - \bar{t})$, a linear function of time with mean zero. The resulting PCA score measures a subject's slope, but is not affected by the subject's level].

PCA weight curves provide information on the important features of multiple time series. For example, two major principal components with "banded" and "crossing" PCA weight curves indicate that the levels and the slopes of the time progression functions carry useful information for classification. The variances explained by these two principal components (and the closeness of the PCA weight functions from the average response curve) express how well a subject's level and slope represent the time progression. Often only scores from the first or the first two principal components are needed for classification, but there could be situations where more than two principal components are needed. Furthermore, the scores resulting from the most important principal components need not be the subject's mean and the slope.

In paper [1] we characterize the time-progression of each subject with a parametric (linear) model. We compare and classify subjects on the estimated summary statistics from that model which – for the linear model – are the intercept and the slope of the fitted trend regression line. The approach in this note is non-parametric and more general as it uses the scores from the most important principal components; the resulting scores don't necessarily have to be the intercept (average) and the slope, and there can be more than two scores.

The functional data analysis approach for data collected at different time periods

But there is a problem when trying to apply this approach to data that do not follow an array-type structure. And this situation arises with the data that we analyze in the next section. There the measurements on 104 glaucoma and 55 normal subjects are observed at many distinct time points (in fact, there are 402 distinct time points), and the few observations on each subject (there are between 4 and 13 observations per subject) are taken at time periods that are not common to all or most subjects. Consequently, most of the entries in the data matrix with its 159 rows and 402 columns are missing. The initial step in the analysis is to estimate the responses for a smaller set of times that are common to all subjects, and for this we use the Functional Data Analysis (FDA) approach. FDA provides a nonparametric method for estimating a very general time response function and can be used to obtain estimated responses for a set of time points that are the same for all subjects.

Functional data analysis is described in references [6-8]. FDA approximates the relationship between a response and time (time in the case of progression but, more general, any continuous covariate) as a linear combination of either Fourier or spline basis functions. Fourier basis functions are sine and cosine functions of increasing frequency. Splines are polynomial (degree g) segments which join at points that are referred to as knots. Knots are placed at distinct time points within the observed time interval, and the polynomial segments (usually linear, quadratic, or cubic) join at the knots. The segments are constrained to be smooth at the joins. Any spline can be generated from basis or B-splines and easy recursive algorithms exist for their construction. One needs to select knots at the two end points and at internal points of the relevant time interval (with more internal knots leading to more spline functions, just like additional frequencies lead to more Fourier functions) and the degree of the polynomial segments (linear segments lead to a "tent" like appearance, whereas quadratic and cubic splines allow for more general patterns). A B-spline is a continuous function at the knots, and when internal knots are distinct, its derivatives are also continuous up to the derivative of degree g-1.

Even quite complicated response functions of time can be approximated by linear combinations of either Fourier or spline basis functions. A rich basis with many basis functions allows for a very general representation. Ordinary least-squares estimates are available if the number of basis functions is less (or equal) than the number of observations. However, with a rich class of basis functions and its many estimates, the fitted response function will be rather noisy. A smoother response function results when introducing a penalty for the roughness of the fitted function. The square of the second derivative of a function at a fixed argument expresses the deviation from linearity as the second derivative disappears for a linear function. Hence, a useful measure of roughness of a function is the integrated squared second derivative of the function, with integration extending over the function's range. Penalized regression determines the estimates by minimizing the sum of the error sum of squares and a multiple of the roughness of the function; the multiple is referred to as the smoothing parameter or the penalty λ . By increasing the penalty, the fitted response function reduces to the linear least squares line. Cross-validation techniques determine the appropriate penalty λ that needs to be imposed, letting the data themselves dictate the smoothness of the response function. Introduction of a penalty into the estimation has yet another advantage because it allows us to consider more basis functions than the number of available observations; with least squares estimation alone, this would not be possible. The general cross-validation measure developed [9], standardizes the error sum of squares by the (square of the) effective degrees of freedom of the fit (which depend on the smoothing parameter and the hat regression matrix). Software is readily available to obtain the smoothing parameter that minimizes the cross-validation measure; we use the *fda package* of the R Statistical Software [10].

A penalized regression on B-splines of degree $g = 2$ with knots at all distinct time periods across all subjects is used to estimate the response curve for each subject. A penalty for the roughness of the fitted function is introduced, and the optimal smoothing coefficient is obtained by minimizing the average of the general cross-validation scores across all subjects. This allows the data to determine the appropriate functional form of the time progression and does not limit the analyst to a specific functional form such as the linear time-trend model considered previously in reference [1]. The response for each of the m subjects is then evaluated at the $T^* = 50$ equally-spaced time periods covering the interval from the smallest to the largest time point. The function *pca.fd* from the *fda package* is used to carry out the principal components analysis on the regular array-type data structure for the m subjects and $T^* = 50$ periods, display the PCA weight curves discussed in the previous section, and calculate the implied principal components scores. The mean response and the PCA weight functions can also be smoothed for better interpretability. Subjects are then classified on their implied principal components scores, using either linear or quadratic discriminant functions.

Example

We use the data on 104 glaucoma patients and 55 normal subjects analyzed in the previous paper. We analyze OCT measurements on the average Retinal Nerve Fiber Layer (RFNL) and the average thickness of the ganglion cell layer derived from the Macula Scan (GCL). We consider a penalized regression on B-splines of degree $g = 2$ with knots at all $T = 402$ distinct available time points. We select the penalty that optimizes the average general cross-validation score across all 159 time series; this penalty estimate is given by $\lambda = 0.5$ .The response for each of the m subjects is evaluated at the $T^* = 50$ equally-spaced time periods covering the interval from the smallest to the largest time point. This creates a regular array-type data structure for $m = 159$ subjects and $T^* = 50$ periods, and principal components can be determined from the resulting covariance matrix.

We find that we need no more than two principal components to describe the variability in the time response function. The first principal component alone explains 98.1 (GCL) and 99.0 (RFNL) percent of the total variation. The wide and roughly parallel bands in the PCA curves for the first principal component (left panel) indicate a very large subject effect (Figure 1). The crossing lines for the PCA curves for the second principal component (right panel), with their small deviations from the average trend line, indicate very minor variability in the subjects' slopes. This fact is confirmed by the small proportions of variability (1.7 percent for GCL and 1 percent for RNFL) that are explained by the second principal components. Scores of the first two principal components are calculated for all subjects, and linear and quadratic discriminant functions are used for classification. Scatter plots of first and second PCA scores are shown in Figure 2. Normal subjects are shown as red dots and glaucoma subjects are shown as blue dots. Subjects that are misclassified by Fisher's quadratic discriminant function are displayed with an open circle in a color that reflects the incorrect classification. A red dot with a blue circle indicates a normal subject incorrectly classified as coming from the glaucoma group. A blue dot with a red circle indicates a glaucoma patient incorrectly classified as coming from the normal group.

Misclassification errors for quadratic discriminant functions are listed in Table 1. The misclassification rates (8.8 percent for GCL and 13.2 for RFNL) are very similar to the misclassification rates we reported in [1], when using least squares estimates of intercepts and slopes. Misclassification errors for quadratic discriminant functions when using the scores of the first principal component alone are only slightly larger; 17/159 = 0.107 for GCL and 25/159 = 0.157 for RFNL. This was expected as the first principal component (which measures the average level of a subject) alone explains about 99 percent of the total variation. The second principal component (which measures the slope) contributes little to the classification. Here we report misclassification rates that are based on the complete data set; results from resampling test-training splits are similar.

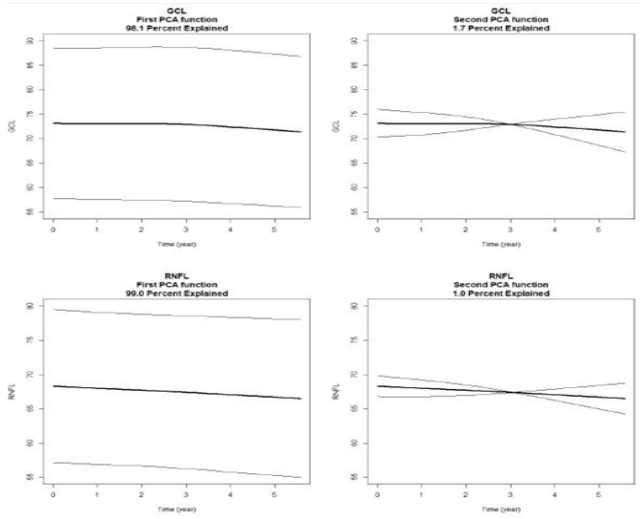

Figure 1: Mean response function (bold) with added first (and second) principal components weights, for GCL (top) and RFNL (bottom). Levels are much more important than slopes for classification.

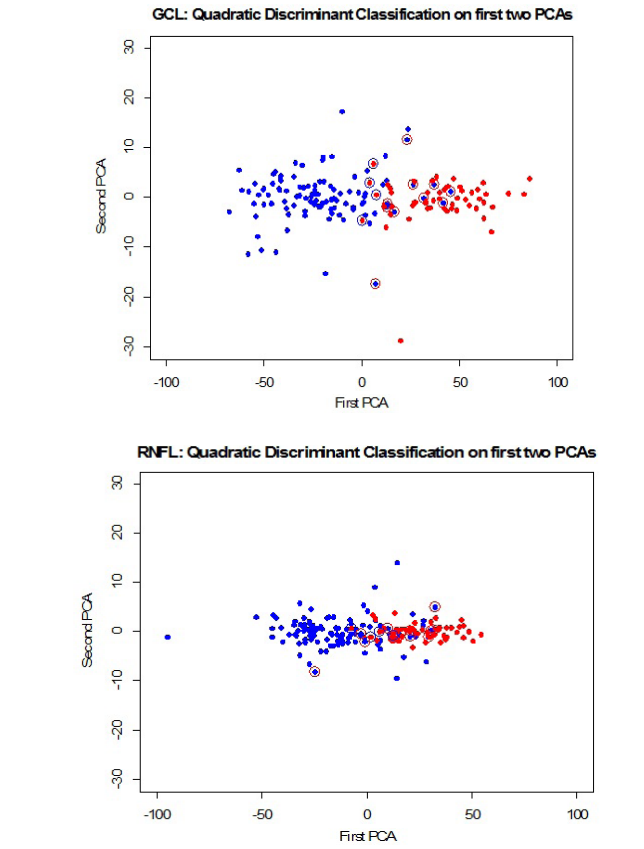

Figure 2: Scatter plots of first and second PCA scores. Normal subjects are shown as red dots and glaucoma subjects are shown as blue dots. Subjects that are misclassified by the quadratic discriminant function are displayed with an open circle in a color that reflects the incorrect classification. A red dot with a blue circle indicates a normal subject incorrectly classified as coming from the glaucoma group. A blue dot with a red circle indicates a glaucoma patient incorrectly classified as coming from the normal group.

GCL		
	Classified as	
True Group	NORMAL	GLAUCOMA
NORMAL	51	4
GLAUCOMA	10	94
Overall misclassification rate: 14/159 = 0.088		
RNFL		
	Classified as	
True Group	NORMAL	GLAUCOMA
NORMAL	49	6
GLAUCOMA	15	89
Overall misclassification rate: 21/159 = 0.132		

Table 1: Cross-classification matrix and misclassification rates of quadratic discriminant functions applied to the first two principal components scores, for GCL and RFNL.

Conclusion

This note and the accompanying example illustrate how functional data analysis and the method of principal components can be combined for the purpose of classifying irregularly-spaced patient records. This approach is useful for supervised classification and unsupervised clustering of irregularly-spaced multiple time series and multiple time series with missing observations.

Funding

This research was supported through grant C9251-C from the US Department of Veterans Affairs Office of Rehabilitation Research & Development.

References

1. Ledolter J, Kardon RH (2018) Assessing trends in functional and structural characteristics: A survey of statistical methods with an example from ophthalmology. Transl Vis Sci Technol 7: 34.

2. Johnson RA, Wichern DW (2012) Applied Multivariate Statistical Analysis (6th Edn). Pearson, Essex, UK.

3. Ledolter J (2013) Data Mining and Business Analytics with R. John Wiley & Sons, New York, USA.

4. Boser BE, Guyon IM, Vapnik VN (1992) A training algorithm for optimal margin classifiers. COLT '92 Proceedings of the fifth annual workshop on Computational Learning Theory Pg no: 144-152.

5. Cristianini N, Shawe-Taylor J (2000) An Introduction to Support Vector Machines and other Kernel-based Learning Methods. Cambridge University Press, Cambridge, UK.

6. Ramsay JO, Hooker G, Graves S (2009) Functional Data Analysis with R and MATLAB. Springer-Verlag New York, New York, USA.

7. Ramsay JO, Silverman BW (2002) Applied Functional Data Analysis: Methods and Case Studies. Springer-Verlag New York, New York, USA.

8. Ramsay JO, Silverman BW (2005) Functional Data Analysis. Springer, New York, USA.

9. Craven P, Wahba G (1978) Smoothing noisy data with spline functions Estimating the correct degree of smoothing by the method of generalized cross-validation. Numerische Mathematik 31: 377-403.

10. Ramsay JO, Wickham H, Graves S, Hooker G (2018) Package 'fda'.

Journal of Anesthesia & Clinical Care

Journal of Addiction & Addictive Disorders

Advances in Microbiology Research

Advances in Industrial Biotechnology

Journal of Agronomy & Agricultural Science

Journal of AIDS Clinical Research & STDs

Journal of Alcoholism, Drug Abuse & Substance Dependence

Journal of Allergy Disorders & Therapy

Journal of Alternative, Complementary & Integrative Medicine

Journal of Alzheimer's & Neurodegenerative Diseases

Journal of Angiology & Vascular Surgery

Journal of Animal Research & Veterinary Science

Archives of Zoological Studies

Archives of Urology

Journal of Atmospheric & Earth-Sciences

Journal of Aquaculture & Fisheries

Journal of Biotech Research & Biochemistry

Journal of Brain & Neuroscience Research

Journal of Cancer Biology & Treatment

Journal of Cardiology: Study & Research

Journal of Cell Biology & Cell Metabolism

Journal of Clinical Dermatology & Therapy

Journal of Clinical Immunology & Immunotherapy

Journal of Clinical Studies & Medical Case Reports

Journal of Community Medicine & Public Health Care

Current Trends: Medical & Biological Engineering

Journal of Cytology & Tissue Biology

Journal of Dentistry: Oral Health & Cosmesis

Journal of Diabetes & Metabolic Disorders

Journal of Dairy Research & Technology

Journal of Emergency Medicine Trauma & Surgical Care

Journal of Environmental Science: Current Research

Journal of Food Science & Nutrition

Journal of Forensic, Legal & Investigative Sciences

Journal of Gastroenterology & Hepatology Research

Journal of Gerontology & Geriatric Medicine

Journal of Genetics & Genomic Sciences

Journal of Hematology, Blood Transfusion & Disorders

Journal of Human Endocrinology

Journal of Hospice & Palliative Medical Care

Journal of Internal Medicine & Primary Healthcare

Journal of Infectious & Non Infectious Diseases

Journal of Light & Laser: Current Trends

Journal of Modern Chemical Sciences

Journal of Medicine: Study & Research

Journal of Nanotechnology: Nanomedicine & Nanobiotechnology

Journal of Neonatology & Clinical Pediatrics

Journal of Nephrology & Renal Therapy

Journal of Non Invasive Vascular Investigation

Journal of Nuclear Medicine, Radiology & Radiation Therapy

Journal of Obesity & Weight Loss

Journal of Orthopedic Research & Physiotherapy

Journal of Otolaryngology, Head & Neck Surgery

Journal of Protein Research & Bioinformatics

Journal of Pathology Clinical & Medical Research

Journal of Pharmacology, Pharmaceutics & Pharmacovigilance

Journal of Physical Medicine, Rehabilitation & Disabilities

Journal of Plant Science: Current Research

Journal of Psychiatry, Depression & Anxiety

Journal of Pulmonary Medicine & Respiratory Research

Journal of Practical & Professional Nursing

Journal of Reproductive Medicine, Gynaecology & Obstetrics

Journal of Stem Cells Research, Development & Therapy

Journal of Surgery: Current Trends & Innovations

Journal of Toxicology: Current Research

Journal of Translational Science and Research

Trends in Anatomy & Physiology

Journal of Vaccines Research & Vaccination

Journal of Virology & Antivirals

Archives of Surgery and Surgical Education

Sports Medicine and Injury Care Journal

International Journal of Case Reports and Therapeutic Studies

Journal of Ecology Research and Conservation Biology

Submit Your Manuscript: http://www.heraldopenaccess.us/Online-Submission.php

Herald Scholarly Open Access, 2561 Cornelia Rd, #205, Herndon, VA 20171, USA.
Tel: +1 202-499-9679; E-mail: info@heraldsopenaccess.us
http://www.heraldopenaccess.us/