Effiziente Dimensionsreduktion durch dynamische funktionale Hauptkomponenten

Efficient Dimension Reduction using Dynamic Functional Principal Components

Siegfried Hörmann

Durch die enormen technologischen Entwicklungen der letzten Jahrzehnte insbesondere im IT-Bereich werden laufend neue, genauere und größere Datenmengen generiert und gesammelt. In einer derartigen Datenflut wird es allerdings immer schwieriger, die richtigen Schlüsse zu ziehen und relevante Informationen herauszufiltern. In meiner Forschung versuche ich, Lösungsansätze zu Teilen dieser komplexen Problematik zu liefern.

Im Laufe der letzten Jahrzehnte wurde das Sammeln und Speichern von Daten ständig einfacher und preisgünstiger. In unserem alltäglichen Leben, aber auch in der Wissenschaft, gibt es viele Bereiche, wo Daten in hoher Auflösung gesammelt werden. Zum Beispiel können auf modernen Motorenprüfständen pro Sekunde mehrere Beobachtungen erhoben werden, und das für etliche Messgrößen gleichzeitig. Ähnliche Beispiele gibt es in vielen anderen Bereichen, man denke an Schadstoffmessreihen, hochfrequente Finanztransaktionen, funktionelle Magnetresonanztomografie etc. Um aus dieser Datenflut einen Nutzen zu ziehen, braucht es entsprechende statistische Methoden, die gewisse Problemstellungen auf das Wesentliche zu reduzieren vermögen und relevante Information extrahieren können. Die funktionale Datenanalyse (FDA) ist ein aufstrebendes Teilgebiet der Statistik, das sich mit dieser Problematik befasst. Methoden der FDA sind dann anwendbar, wenn jede einzelne Beobachtung eine auf einem Kontinuum definierte Funktion darstellt, zum Beispiel die Grauwerte in einem Schwarz-Weiß-Bild oder die Temperaturverteilung auf der Erdoberfläche, siehe Abbildung 1.

Wachstumskurven und PM10

Zur Illustration geben wir zwei einfache Beispiele: Wir betrachten zum einen Wachstumskurven von zehn Kindern im Alter von 0 bis 18 Jahren und zum anderen tägliche Feinstaubkurven in Graz an zehn aufeinanderfolgenden Wintertagen, siehe Abbildung 2. Jede dieser Kurven enthält eine >

With increasing complexity and the rapidly growing amount of data collected in almost all areas of our life, it becomes more and more difficult to draw meaningful conclusions and to filter relevant information. The field of statistics hasseen a big upsurge due to such new challenges. My research is devoted to some of these challenges.

Over the past decades storing and collecting data electronically has steadily become easier and cheaper. As a consequence, for many everyday life processes or scientific experiments nearly continuous data records exist. For example, on some engine test benches hundreds of variables can be collected and it is not uncommon to have for certain parameters of interest several measurement points per second. Similar examples can be given in environmental sciences (pollution levels), geophysics (strength of magnetic fields), medicine (fMRI images) or econometrics (tick-data), to just name a few. To benefit from increasing information, scientists need appropriate statistical tools which can help in finding the most important characteristics in such a big data context. Functional data analysis (FDA) is one of the emerging statistical disciplines which aims to extract relevant information from complex, intrinsically high-dimensional data objects. It is targeted for data samples where each underlying sampling point is a curve or some other process defined on a continuum, such as a grey level image or surface temperatures, etc. (Figure 1.)

Growth curves and PM10

To clarify ideas, let us look at two simple FDA examples. In the first, we consider growth curves of 10 children at the age of 0-18 years and in the second, we look at daily PM10 pollution level curves in Graz (Figure 2). In each curve we can check for many abstract features that may have practical or scientific relevance: e.g. the average level, the maximum, a potential trend or the position and number of peaks are important features in an environmental study on PM10 levels. >

© Institut für Statistik

Siegfried Hörmann ist seit Oktober 2017 Professor für Angewandte Statistik am Institut für Statistik der TU Graz. Zuvor war er zwei Jahre in den USA und dann acht Jahre in Belgien wissenschaftlich tätig.

Siegfried Hörmann has been Professor of Applied Statistics at the Institute of Statistics of TU Graz since October 2017. Before this, he held professorships in the USA and Belgium.

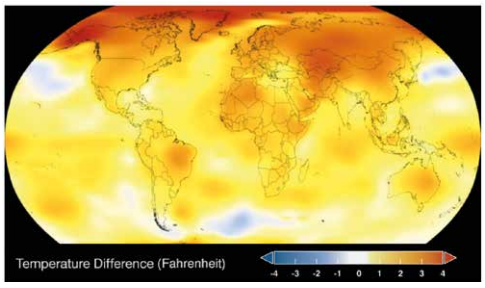


Abbildung 1:
Veränderung der globalen Oberflächentemperatur gegenüber 1951–1980.
Figure 1:
Change in global surface temperature relative to 1951–1980.

Reihe von abstrakten Eigenschaften, die je nach Fragestellung relevant sein können: Zum Beispiel sind das Tagesmittel, das Maximum, ein eventueller Tagestrend oder die Zeitpunkte und die Anzahl von Spitzenwerten interessanter Kenngrößen für PM10-Daten. Wenn wir nicht nur am Verlauf eines einzelnen Tages interessiert sind, sondern vielmehr an einer zugrundeliegenden Systematik, bieten sich statistische Methoden an. Anhand von wiederholten Messreihen (Wachstumskurven von Individuum 1, 2, 3 … und Feinstaubbelastung an den Tagen 1, 2, 3 …) versuchen wir, typische Muster zu erkennen. Im Idealfall können wir daraus Rückschlüsse auf die Schadstoffquellen ziehen oder Prognosen zum weiteren Wachstumsverlauf eines Kindes liefern. In beiden Beispielen gibt es eine natürliche Variation zwischen den Beobachtungen, die aus der Komplexität der zugrunde liegenden physikalischen und biologischen Prozesse rührt. Genau dann, wenn ein System zu komplex für ein exaktes naturwissenschaftliches Modell ist, helfen Methoden der Statistik und Wahrscheinlichkeitstheorie, um diese Variation zu modellieren.

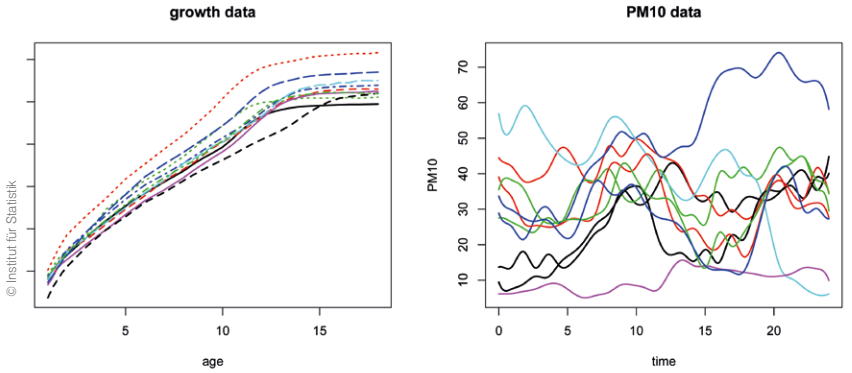


Abbildung 2:
Zehn Wachstumskurven (linke Tafel) und zehn tägliche PM10-Stufen (rechte Tafel).
Figure 2:
Ten growth curves (left panel) and ten diurnal PM10 levels (right panel).

Umgang mit hoher Dimension

Aus mathematischer Sichtweise sind unsere funktionalen Beobachtungen Realisierungen eines stochastischen Prozesses. Da, wie oben angedeutet, die Trajektorien dieser Funktionen viele Eigenschaften besitzen, sind sie aus mathematischer Sicht hochdimensionale (theoretisch gesehen gar unendlichdimensionale) Objekte. Es ist also aus vielerlei Hinsicht wünschenswert, deren Dimension zu reduzieren und für die weitere Analyse nur die wichtigsten Eigenschaften herauszufiltern. Eine fundamentale Rolle spielt in diesem Zusammenhang

Statistics comes into play since we have replicates of the same experiment (measuring growth curves of individuals 1, 2, 3, … and pollution levels on day one 1, 2, 3, …). It is usually not particularly interesting if there was a peak PM10 load on a certain day at a certain time, but we may be very interested if peaks arise in a systematic way throughout a period of time. This will allow us to draw better conclusions regarding the polluters or to give better forecasts regarding the growth of a child. In both examples, there is variation and uncertainty between replicates due to very complex physical and biological processes, such as the nutrition regime and genetic endowments in the growth curves example. When a system becomes too complex to model all of its aspects, probabilistic and statistical tools enter the stage.

Tackling high dimension

In mathematical terms the curves that we investigate are realizations of a stochastic process. The fact that these random curves contain many features means that they constitute intrinsically high (in theory infinite) dimensional mathematical objects. From a practical as well as from a theoretical point of view, one is interested to reduce the dimensionality of the problem and to retain for a further analysis only those features in our observations which best describe the curves. A key statistical tool to tackle the dimensionality of functional data is the so-called functional principal component analysis. *Functional principal components are orthogonal basis functions and as such we can use them to represent our functional observations as a superposition of these curves. This representation is called Karhunen-Loève (KL) expansion and its theoretical foundations date back to the early 20th century. Back then this approach was numerically unfeasible and hence it was not targeted for statistical applications. By expanding along a small number of basis-functions we obtain a low dimensional representation of the curve. The reader familiar with Fourier series may compare this to the Fourier expansion, where a curve is represented as a superposition of sinusoidal functions. The advantage of functional principal components is that, in some sense, they optimally adapt to the data. In Figure 3 we illustrate the approximation of a PM10 curve with 3 principal components and 5 and 25 Fourier basis functions, respectively.*

Incorporating serial correlation

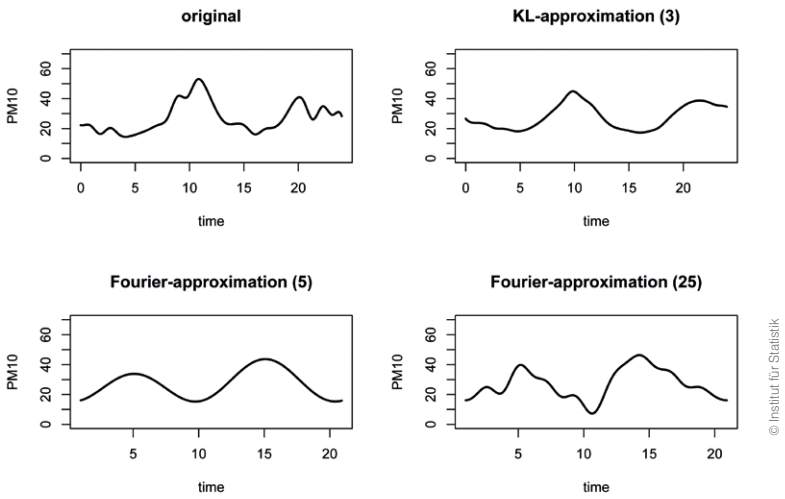
When looking at the PM10 and growth curve data, we observe several fundamental differences. For example, in contrast to PM10 data, the growth curves are monotone and smooth. Another important difference is that the growth data are statistically independent – there is no reason why the growth curve of one child should impact the growth curve of

die *funktionale Hauptkomponentenanalyse*. Hauptkomponenten sind orthogonale Funktionen. Durch Überlagerung dieser Funktionen lässt sich die ursprüngliche Funktion rekonstruieren. Man nennt dieses Verfahren Karhunen-Loève-(KL-)Entwicklung – ein Konzept, dessen theoretische Erforschung bereits Anfang des 20. Jahrhunderts stattfand. Zu diesem Zeitpunkt war eine statistische Anwendung nicht von Interesse, zumal es keine Möglichkeit einer numerischen Implementierung gab. Heute gibt es dazu Softwarepakete und wir können mittels einer KL-Entwicklung die Trajektorien unserer funktionalen Daten in beliebiger Dimension blitzschnell approximieren. Wer mit Fourierreihen vertraut ist, kann diese Methodik mit der Fourier-Entwicklung vergleichen. Hier werden Funktionen als Überlagerung von Sinus- und Cosinusschwingungen dargestellt. Der Vorteil der Hauptkomponenten liegt vor allem darin, dass sich diese in einem gewissen Sinne optimal an die Daten anpassen und damit eine ausgezeichnete Approximation bereits in kleiner Dimension gewähren. In Abbildung 3 illustrieren wir die Approximation eines PM10-Tagesverlaufs mit drei Hauptkomponenten bzw. mit 5 und 25 Fourier-Basen.

Einbindung von serieller Abhängigkeit

Beim Vergleich von PM10- und Wachstumskurven fallen schnell einige offensichtliche Unterschiede auf. Zum Beispiel sind Wachstumskurven, im Gegensatz zu den PM10-Kurven, monoton und sehr glatt. Ein anderer wesentlicher Unterschied ist, dass die Wachstumskurven statistisch unabhängig sind: Der Verlauf des Wachstums eines Kindes hat keinen Einfluss auf den eines anderen Kindes. Für die PM10-Daten gilt dies nicht. Wenig überraschend gibt es starke Korrelationen zwischen den aufeinanderfolgenden Tagen. Im Zusammenhang mit FDA treten solche zeitlichen Abhängigkeiten sehr häufig auf.

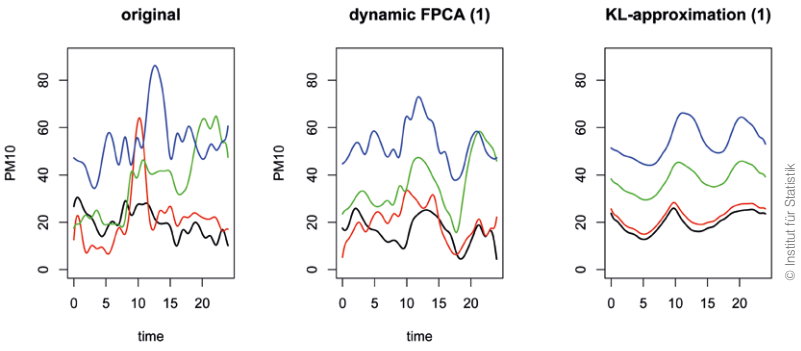
In einem meiner Forschungsprojekte zeige ich mit meinen Kolleg/innen, wie etwaige serielle Abhängigkeit genutzt werden kann, um die Dimension von funktionalen Daten noch effizienter zu reduzieren. Unsere Methode heißt *dynamische Hauptkomponentenanalyse* und basiert auf frequenzanalytischen Methoden. Aus diesem Ansatz ergeben sich nebst verbesserter Interpretation der Daten etliche Anwendungen zur vereinfachten statistischen Inferenz. Zur Illustration vergleichen wir die Approximation von vier PM10-Kurven mit einer einzigen Hauptkomponente sowie mit einer einzigen dynamischen Hauptkomponente, siehe Abbildung 4. ■



another child. This, however, is no longer true for the PM10 data. Not surprisingly, there is strong correlation between the PM10 loads on consecutive days. This problem is very common in FDA. It is related to the fact that many functional data are sampled sequentially in time (e.g. when data are obtained by segmenting a continuous process into natural units, such as daily data) which then often yields dependences.

In one of my recent research projects I showed with my collaborators that the dependence between functional data can be used in order to obtain much more efficient dimension reduction than with common functional PCA. Our method is called dynamic functional principal component analysis.

Abbildung 3:
PM10-Kurve (oben links) und Approximation mit drei Hauptkomponenten (oben rechts). Die unteren Abbildungen zeigen die Approximation um fünf (links) und 25 (rechts) Fourier-Basisfunktionen.
Figure 3:
PM10 curve (upper left) and approximation with 3 principal components (upper right). Lower figures show the approximation by 5 (left) and 25 (right) Fourier basis functions.



This approach, based on a so-called frequency domain analysis, not only allows for a better interpretation of the data, but is also useful in many problems of statistical inference. For the purposes of illustration, we show a 1-dimenional approximation of four consecutive PM10 curves by means of the usual KL-expansion and dynamic functional PCA (Figure 4) ■

Abbildung 4:
Vier PM10-Kurven (links) und Approximation durch eine dynamische Hauptkomponente (Mitte) sowie eine normale Hauptkomponente (rechts).
Figure 4:
Four PM10 curves (left) and approximation with one dynamic principal component (middle) as well as one ordinary principal component (right).