Fast decomposition on proper vectors in multidimensional space

D N Gerasimov and S B Morgunova
National Research University “Moscow Power Engineering Institute”,
Russia, 111250 Moscow, Krasnokazarmennaya, 14
deniger@mail.ru

Abstract. A recursive method to obtain proper vectors for a data set representation is proposed. This method faster than Karhunen-Loève expansion in SVD-realization for multidimensional space.

1. Introduction: Karhunen–Loève expansion
The problem of finding a simplified representation for a complex dataset usually requires decomposition in Karhunen-Loève expansion (KLE). For instance, this method can be used for analysis of time-space structures (such as coherent structures in turbulence), in problems of pattern recognition or even to find reconstruction equations [1–6].

KLE runs as follows. Suppose that we have dataset of $M$ points in $N$-dimensional space: $\{x^i\}, i=1..M$. We want to represent any vector of this set by expansion

$$\bar{x}^i = \sum_{j=1}^{N} a^j_1 \bar{v}^j.$$  (1)

Using new coordinates $\bar{q}^i = \bar{x}^i - \bar{x}^c$ (where $\bar{x}^c = \frac{1}{M} \sum_{i=1}^{M} \bar{x}^i$ is the 'center of mass'), we find orthogonal vectors $\bar{v}^j, j=1..N$, to satisfy conditions $\frac{1}{M} \sum_{i=1}^{M} (\bar{q}^i \bar{v}^j)^2 \rightarrow \max$ and $||\bar{v}^j||=1$. This problem leads to the determination of eigenvectors for the matrix $Q_{mn} = \frac{1}{M} \sum_{k=1}^{M} q^k_m q^k_n$, $m,n=1..N$:

$$Q \bar{v}^j = \lambda^j \bar{v}^j.$$  (2)

Thus, to find vectors $\bar{v}^j$ for expansion (1) we must solve the problem for eigenvectors (2).

2. Eigenvectors in KLE-method
To construct matrix $Q$, we must produce $\sim MN^2$ operations. As a result of these calculations we get a matrix $N \times N$; then we have the eigenvector problem for it. To avoid underestimation of KLE-
method, we consider calculation of eigenvectors and eigenvalues by reverse iteration method; thus, one iteration consists of the following steps (n is the iteration number):

1. \( \lambda_n = (Qv_n, v_n) \),
2. \( (Q - \lambda_n I)v_{n+1} = v_n \),
3. \( v_{n+1} = \frac{v_n}{||v_{n+1}||} \).

Following this way, we must produce \( S_1 \sim (2N-1)(N+1) \) operations on the first step, \( S_2 \sim 2IN^2 \) operations (\( I \) – amount of iterations to solve a linear system) on the second step, and \( S_3 \sim 3N \) operations on the third one. Denote the number of iterations on steps 1–3 as \( L \) and assume that we can use this method to find all of \( N \) eigenvectors, so to realize this method we have to produce \( \sim LN(S_1 + S_2 + S_3) \) calculations. Thus, total amount of operations is about \( \sim 2(I+1)LN^3 \) (for \( N \gg 1 \)) and comparable with the amount of operations for constructing matrix \( Q \) only for \( M \sim 2ILN \), that is in case when number of points is not larger than \( 2ILN \), the eigenvectors problem is the most expansive part of KLE-method.

However, for example, for the image recognition problem \( N \gg M \): there are \( M \sim 10^{2-3} \) images of size \( N \sim 10^{4-6} \) pixels, and SVD (Singular Value Decomposition) method is used to calculate the eigenvectors. SVD requires \( \sim 6NM^2 + O(M^3) \) operations [4], and the main benefit of SVD-method is the determination of the whole set of eigenvectors, i.e. all \( N \) vectors. But we do not need all these vectors, as a rule; only the first few of them provide useful information.

This – partial – basis can be found by another algorithm. Below we describe a fast method to find the main proper vectors (further we will use the term 'proper vectors' to distinguish them from the eigenvectors in equation (2)) of decomposition, which requires \( O(MN) \) operations.

3. Determining ‘the main direction’

The main idea of decomposition (1): we want to represent our data set by the orthogonal vectors oriented in the 'main directions' of dataset. In KLE we mean that 'main directions' are satisfied to conditions \( \frac{1}{M} \sum_{i=1}^{M} (q_i^j v_i^j)^2 \rightarrow \max \) i.e. deviation of \( \bar{v}^j \) from the set of \( q_i^j \) is minimal. In this work we suggest another way to define the term 'main direction'.

Let us consider a data set \( \bar{q}_i^j \) (as described above) and a hyperplane \( H \), which intersects this set at origin (that is at the center of mass of this set) and defined by its normal vector \( \bar{v} \). Then we find a center of mass for points \( \bar{q}_i^j \) in both half-spaces relatively to \( H \). In fine, we demand that both centers of mass must lie on the same stable 'main direction' \( \bar{v} \). Exactly, we find a vector \( \bar{v} \) from the map

\[ \bar{v}(n) = \frac{1}{M} \sum_{i=1}^{M} k_i^{(n-1)} \bar{q}_i^j, \text{ where } k_i^{(n-1)} = \text{sign}(\bar{v}^{(n-1)} \bar{q}_i^j). \]  

(3)

For further purposes, we must also normalize \( \bar{v} \).

For certain vectors \( \bar{q}_i^j \) deviation \( \delta \bar{v} = \frac{1}{M} \sum_{i=1}^{M} \delta k_i \bar{q}_i^j \) decreases (generally) as a number of non-zero items of this sum, so the less amount of items with \( \delta k_i \neq 0 \) on the given iteration, the less deviation \( \delta \bar{v} \) on the next one. Thus, a map (3) has stable solution – vector \( \bar{v} \) which lies along the most ‘stretched’ direction (and \( H \) is a ‘waist-plane’ of the set). For example, a ‘main direction’ for an ellipse is a long axis; see also figure 1.
Figure 1. The main vector for an ellipse. As one can see, only four iterations are needed to find the right direction starting from the absolutely wrong one.

Iterative map (3) can be used to find all the required vectors $\tilde{v}^j$, $j = 1..K$, but we must recalculate vectors $\tilde{q}^i, i = 1..M$ after finding vector $\tilde{v}^j$ by projecting $\tilde{q}^i$ on $H$:

$$q^i \rightarrow \tilde{q}^i - \tilde{v}^j (q^i \cdot \tilde{v}^j),$$

(4)

where vector $\tilde{v}^j$ is normalized.

Below we will refer this method as 'fast decomposition' (FD). To explain this name we calculate amount of operations needed to realize our method.

4. Comparison between FD and KLE

To calculate $L$ iterations of (3) (for one vector $\tilde{v}$), we must produce $\sim 4LMN$ operations ($\sim 2NM$ operations for $\text{sign}(\tilde{v} \tilde{q})$ and $\sim 2NM$ operations to find the sum on each iteration). Note that usually $L \sim 1$ is sufficient for reasonable accuracy, so to find the main vector of decomposition $O(MN)$ operations are needed.

To find the next vector, we have to realize projective operation (4), which costs $\sim 2MN$ calculations (no need to calculate scalar products on this stage because we already have them from the iterative stage). Total amount of required operations (for $K$ vectors) is $2(2L+1)KMN$. Thus, the determination of the first $K$ vectors can be faster than SVD decomposition for $M > (2L+1)K/3$ (for $N >> M$). For instance, for $M = 100$, $L = 5$, $K = 5$, and FD is $\sim 5$ times faster than SVD.

To calculate all of $N$ vectors $\tilde{v}$, we must produce $\sim 2(2L+1)N^2M$ operations. This is, theoretically, more slowly than SVD-method for the case $N > M$, but because of the sense of decomposition the value of $K$ must be even less than $M$. So, for reasonable number of proper vectors our FD method has, at least, comparable computational speed with SVD decomposition.
5. Application for hydrodynamics

For nonlinear equations of hydrodynamics, one of possible simplifying approaches is the representation of all functions as trigonometric series. For example, for the stream function:

$$\Psi(t,x,y) = \sum_{m} \sum_{n} a_{mn}(t) \sin \left( \frac{2\pi n}{L_x} x \right) \sin \left( \frac{2\pi m}{L_y} y \right).$$

(5)

Thus, we have a deal with a time evolution of vector $\vec{a}(t)$ defined in multidimensional space (components of this vector are coefficients $a_{mn}(t)$; there may be $\sim 10^4$ components or even more). The magnitudes $a_{mn}(t)$ are obtained, usually, from the numerical simulations, so in practical case we have the set of values $\vec{a}^i$: each vector corresponds to the point in time $t_i$.

To segregate the main relative values of magnitudes (i.e. to distinguish main spatial harmonics) one may use the fast decomposition described above: the result would be vectors $\vec{A}_1, \vec{A}_2, \ldots$ One may hope that the first few vectors $\vec{A}_k$ would contain main information about the spectrum, while other vectors are minor corrections. After that, evolution of vector $\vec{a}$ can be represented approximately as

$$\tilde{a}(t) = \sum_k c_k(t) \vec{A}_k,$$

(6)

where $c_k(t)$ are coefficients describing the time evolution of the system. The set of vectors $\vec{A}_k$ corresponds to main patterns of irregular (i.e. turbulent) flow.

Conclusion

We have proposed a recursive method to calculate proper vectors for a data set. This method consists of consequent stages to calculate each proper vector after another one, so it suits well to find the partial basis of decomposition. To determine the first few vectors, this method works faster than SVD realization of Karhunen–Loève transform.

Our method may be useful in many investigations, where the recognition of space-time structures is needed. Possible application is the analysis of turbulent patterns as well.

References

[1] Haken H 1996 Principles of Brain Functioning (Berlin Heidelberg: Springer-Verlag)
[2] Fukunaga K 1990 Introduction to Statistical Pattern Recognition (London: Academic Press)
[3] Rowley C W, Marsden J E 2000 Physica D 142 1
[4] Levy A, Lindenbaum M 2000 IEEE Transactions on Image Processing 9 1371
[5] Uenohara M 1997 IEEE Transactions on Pattern Analysis and Machine Intelligence 19 891
[6] Londoño E L, Lopez L C, de Souza Kazmierczak T 2005 Earth Science Research Journal 9 139