MATH590: Approximation in $\mathbb{R}^d$

Abstract

The methods of linear algebra are used to distinguish between different gaits.

Table of contents

1 Karhunen-Loève theorem

A *probability space* is a triplet $(\Omega, \mathcal{F}, P)$ with:

- Ω a sample space of all possible outcomes;
- $\mathcal{F}$ a set of events that is a set of subsets of Ω;
- $P: \mathcal{F} \to \mathbb{R}$ a probability function for each event.

Rather improperly named, a *random variable* $X: \Omega \to E$, is a function defined on a sample space with values in a measurable space (e.g., $\mathbb{R}^d$). For some measurable subset $S \subseteq E$, the probability of $X \in S$ is

$$\Pr(X \in S) = P(\{\omega \in \Omega | X(\omega) \in S\})$$

A *stochastic process* $X_t(\omega)$ is indexed collection of random variables. Often the index is time, and $X_t: \mathbb{R} \times \Omega \to E$. A centered stochastic process has mean value zero

$$\mathbb{E}[X_t(\omega)] = 0,$$

with $\mathbb{E}$ the expectation operator.

The Karhunen-Loève theorem affirms the existence of a canonical description of a stochastic process as a linear combination of random variables Z_k with time-dependent coefficients $e_k(t)$, or, conversely, as a linear combination of time-varying functions $e_k(t)$ with random coefficients Z_k

$$X_t(\omega) = X(t, \omega) = \sum_{k=1}^{\infty} Z_k \, e_k(t).$$

2 Singular-value decomposition

As will be explained in detail in the STC module, the singular value decomposition is a discrete form of the Karhunen-Loève theorem. Let the matrix $\mathbf{X}$ denote multiple samples of some real-valued, centered stochastic process

$$\mathbf{X}=\begin{pmatrix} X(t_1,\omega_1) & X(t_1,\omega_2) & \ldots & X(t_1,\omega_N) \\ X(t_2,\omega_1) & X(t_2,\omega_2) & \ldots & X(t_2,\omega_N) \\ \vdots & \vdots & \ddots & \vdots \\ X(t_m,\omega_1) & X(t_m,\omega_2) & \ldots & X(t_m,\omega_N) \end{pmatrix} \in \mathbb{R}^{m\times N}.$$

There exist orthogonal matrices $\mathbf{U}\in\mathbb{R}^{m\times m}$, $\mathbf{V}\in\mathbb{R}^{N\times N}$, and the quasi-diagonal positive matrix $\mathbf{\Sigma}=\operatorname{diag}(\sigma_1,\ldots,\sigma_r,0,\ldots,0)\in\mathbb{R}_+^{m\times N}$ such that

$$\mathbf{X}=\mathbf{U}\mathbf{\Sigma}\mathbf{V}^T,$$

known as the singular value decomposition (SVD). Introducing the column vectors of $\mathbf{U},\mathbf{V}$

$$\mathbf{U}=\begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \ldots & \mathbf{u}_m \end{pmatrix}, \mathbf{V}=\begin{pmatrix} \mathbf{v}_1 & \mathbf{v}_2 & \ldots & \mathbf{v}_N \end{pmatrix},$$

the SVD can be rewritten in two important forms:

Sum of rank-1 updates. $\mathbf{X}=\sum_{k=1}^r \sigma_k \mathbf{u}_k \mathbf{v}_k^T$

Bases for linear operator X.

$$\mathbf{X}\begin{pmatrix} \mathbf{v}_1 & \ldots & \mathbf{v}_N \end{pmatrix}=\begin{pmatrix} \mathbf{X}\mathbf{v}_1 & \ldots & \mathbf{X}\mathbf{v}_r & \ldots & \mathbf{X}\mathbf{v}_N \end{pmatrix}=\begin{pmatrix} \sigma_1\mathbf{u}_1 & \ldots & \sigma_r\mathbf{u}_r & \mathbf{0} & \ldots & \mathbf{0} \end{pmatrix}.$$

3 Covariance matrices

The covariance of two centered random variables $X(t_i,\omega)=X_i, X(t_j,\omega)=X_j$ is

$$\operatorname{cov}[X_i,X_j]=\mathbb{E}[X_i X_j],$$

typically approximated through a statistic from N observations

$$\mathbb{E}[X_i X_j]=\frac{1}{N}\sum_{k=1}^{N} X_i(\omega_k)X_j(\omega_k)$$

The covariance matrix $\mathbf{C}$ of m centered random variables $X(t_1,\omega),\ldots,X(t_m,\omega)$ has entries

$$C_{ij}=\operatorname{cov}[X_i,X_j],$$

and is expressed as the matrix product

$$\mathbf{C} = \frac{1}{N}\mathbf{X}\mathbf{X}^T \in \mathbb{R}^{m\times m}.$$

By construction the covariance matrix is symmetric positive definite (spd) and therefore admits an orthogonal eigendecomposition

$$\mathbf{C} = \mathbf{U}\mathbf{\Lambda}\mathbf{U}^T = \sum_{k=1}^{m} \lambda_k\, \mathbf{u}_k\mathbf{u}_k^T.$$

Assume (by re-labeling if necessary) that $\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_m$. It is often the case that p eigenvalues dominate over all others (the strongly correlated modes)

$$\lambda_1 \geqslant \lambda_2 \geqslant \ldots \geqslant \lambda_p \gg \lambda_{p+1} \geqslant \cdots \geqslant \lambda_m,$$

and the correlation matrix is approximated by the first p rank-1 updates

$$\mathbf{C} \cong \sum_{k=1}^{p} \lambda_k\, \mathbf{u}_k\mathbf{u}_k^T.$$

4 Data-driven bases

The dominant correlated modes form a natural basis set for analysis of data. Rather than solving the covariance matrix eigenproblem the SVD of $\mathbf{X}$ is used since

$$\mathbf{X} = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T, N\mathbf{C} = \mathbf{X}\mathbf{X}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{V}^T\mathbf{V}\mathbf{\Sigma}^T\mathbf{U}^T = \mathbf{U}\mathbf{\Sigma}\mathbf{\Sigma}^T\mathbf{U}^T.$$

5 Application to gait analysis

Consider the data obtained from many individual gait measurements (either from different individuals or at different times for the same individual). The goal is to identify differences w.r.t. a mean gait and use those differences to (1) identify either an individual or (2) a particular type of walking (climbing stairs versus level walking). The procedure is demonstrated here for the second problem.

```
octave> dir='/home/student/courses/MATH590/NUMdata';
        chdir(dir);
        LastName='Mitran';
        d=textread(strcat(LastName,'.data'));
        mu=max(size(d))/7; disp(mu)
```

```
14190
```

```
/home/student/courses/MATH590/NUMdata
```

```
octave> data=reshape(d,[7,mu])';
        data(1:6,1:7)
```

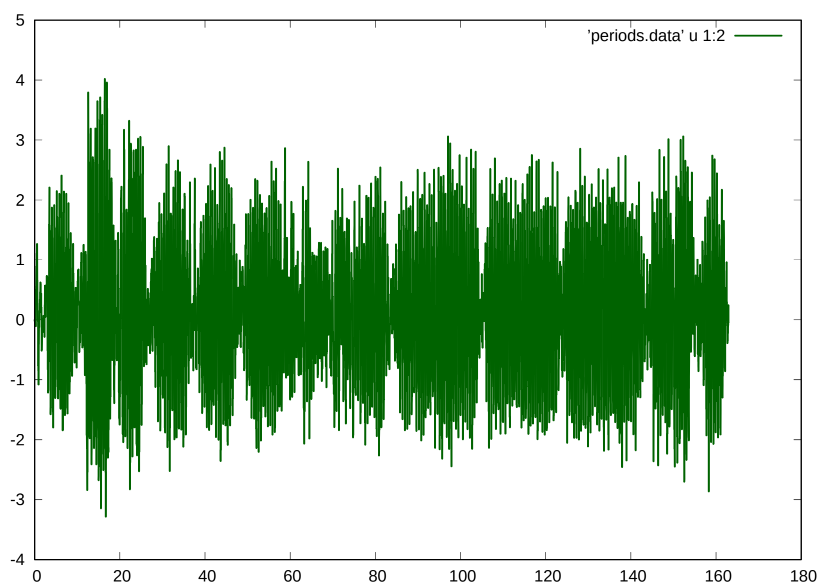
$$\begin{pmatrix} 0 & 0 & 0 & 0 & 0 & 0 & 0 \\ 0.004 & -0.7803 & 0.2999 & -1.799 & -8.0329 & 16.868 & 3.0539 \\ 0.005 & -0.0953 & 0.0762 & -0.0111 & 0.085944 & 21.033 & 12.152 \\ 0.006 & -0.0953 & 0.0762 & -0.0111 & 0.085944 & 21.033 & 12.152 \\ 0.007 & -0.6242 & -0.0203 & -0.9742 & -0.017189 & 42.279 & -1.5986 \\ 0.025 & -0.6242 & -0.0203 & -0.9742 & -0.017189 & 42.279 & -1.5986 \end{pmatrix}$$

```
octave>
```

Interpolate to obtain even-spaced data, and plot the data.

```
octave> t0=data(1,1); t1=data(mu,1); ni = 2^ceil(log2(mu)); dt=(t1-t0)/ni;
        ti=(0:ni-1)*dt; ti=ti';
        ai=interp1(data(:,1),data(:,3),ti);
        fid=fopen('periods.data','w');
        fprintf(fid,'%f %f\n',[ti ai]');
        fclose(fid);
octave>
```

```
GNUplot] cd '/home/student/courses/MATH590/NUMdata'
         set terminal postscript eps enhanced color
         set style line 1 lt 2 lc rgb "0x00006400" lw 3
         plot 'periods.data' u 1:2 w l ls 1
```

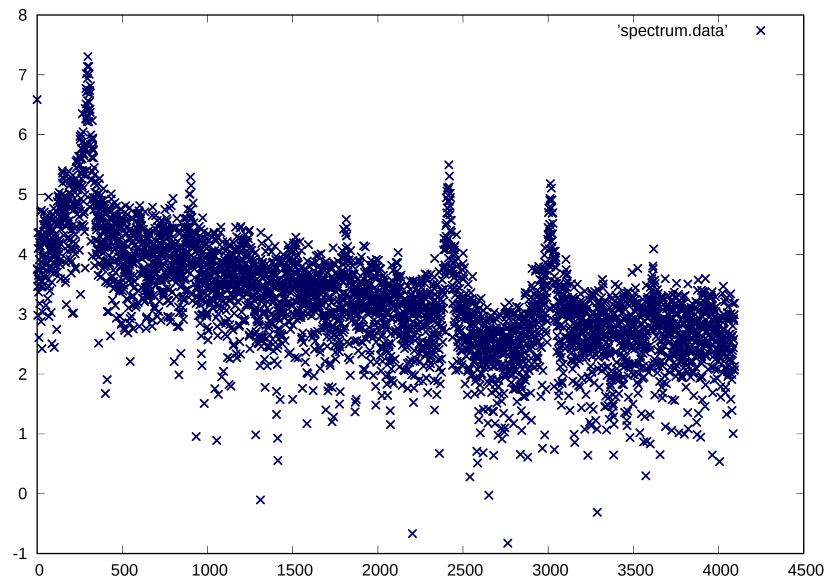



```
GNUplot]
```

Find the Fourier spectrum of the vertical acceleration data.

```
octave> Ai=fft(ai); PAi=log10(Ai.*conj(Ai));
        fid=fopen('spectrum.data','w');
        fprintf(fid,'%f\n',PAi(1:ni/4));
        fclose(fid);
octave>
```

```
GNUplot] cd '/home/student/courses/MATH590/NUMdata'
         set terminal postscript eps enhanced color
         set style line 1 lt 2 lc rgb "0x00000064" lw 3
         plot 'spectrum.data' ls 1
```




```
GNUplot]
```

There are several peaks within the Fourier spectrum. Seek the peak corresponding to a natural step period of approximately $T_s = (t_1 - t_0) / n_{\text{Steps}} \cong 0.5\, s$. This turns out to be close to the global peak as shown in the following calculations

```
octave> ks=floor((t1-t0)/(0.5))
```

325

```
octave> [PAimx imx]=max(PAi);
octave> [PAimx imx]
```

(7.3053 299)

```
octave> N=imx-1;
octave> Ts=(t1-t0)/N
```

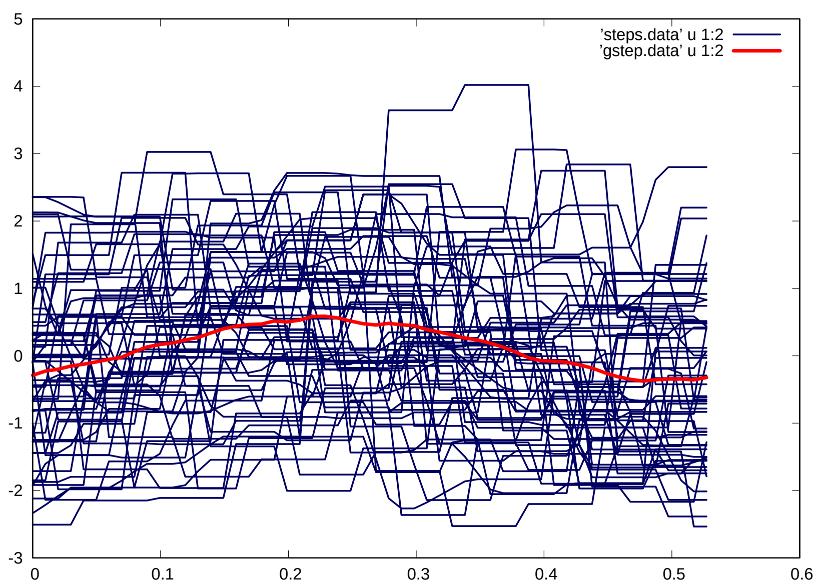
0.54679

```
octave>
```

Isolate the steps, find an average step waveform g, and compute the centered data matrix $\mathbf{X}$

```
octave> m = floor(ni/N); t=(0:m-1)'*dt;
        a = reshape(ai(2:m*N+1),[m,N]); g = mean(a')';
        X = a - repmat(g,1,N);
octave> fid=fopen('steps.data','w');
        i=1; while(i<N)
          fprintf(fid,'%f %f\n',[t a(:,i)]');
          fprintf(fid,'\n');
          i=i+5;
        endwhile;
        fclose(fid);
octave> fid=fopen('gstep.data','w');
        fprintf(fid,'%f %f\n',[t g]');
        fclose(fid);
octave>
```

```
GNUplot] cd '/home/student/courses/MATH590/NUMdata'
         set terminal postscript eps enhanced color
         set style line 1 lt 2 lc rgb "0x00000064" lw 3
         set style line 2 lt 2 lc rgb "0x00ff0000" lw 6
         plot 'steps.data' u 1:2 w l ls 1, 'gstep.data' u 1:2 w l ls 2
```




```
GNUplot]
```

Find the natural basis for the problem by computing the SVD of $\mathbf{X}$, and investigate the dominant singular values

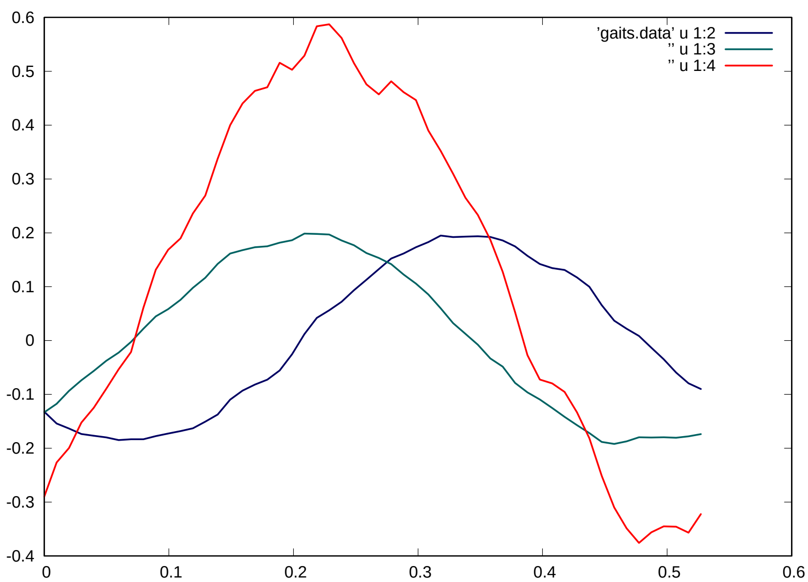
```
octave> [U,S,V]=svd(X,0);
octave> (diag(S)(1:6))'
```

(108.031 93.528 23.645 19.604 17.588 15.726)

```
octave> fid=fopen('gaits.data','w');
        fprintf(fid,'%f %f %f %f\n',[t U(:,1:2) g]');
        fclose(fid);
octave>
```

From the above each step can be characterized by the first two components $\mathbf{u}_1, \mathbf{u}_2$. Plot these modes and compare to the average gait $\mathbf{g}$.

```
GNUplot] cd '/home/student/courses/MATH590/NUMdata'
         set terminal postscript eps enhanced color
         set style line 1 lt 2 lc rgb "0x00000064" lw 3
         set style line 2 lt 2 lc rgb "0x00006464" lw 3
         set style line 3 lt 2 lc rgb "0x00FF0000" lw 3
         plot 'gaits.data' u 1:2 w l ls 1, '' u 1:3 w l ls 2, '' u 1:4 w l
         ls 3
```




```
GNUplot]
```

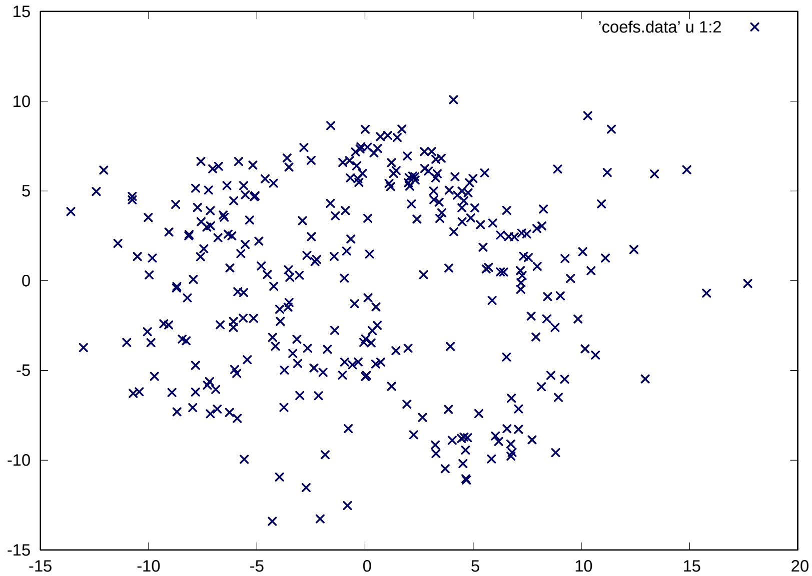
Find the coordinates $\mathbf{c}$ of each step along these directions.

```
octave> c=X'*U(:,1:2);
        fid=fopen('coefs.data','w');
        fprintf(fid,'%f %f\n',c');
        fclose(fid);
octave> size(c)
```

(298 2)

```
octave>
```

```
GNUplot] cd '/home/student/courses/MATH590/NUMdata'
         set terminal postscript eps enhanced color
         set style line 1 lt 2 lc rgb "0x00000064" lw 3
         plot 'coefs.data' u 1:2 w p ls 1
```




```
GNUplot]
```

At this point, an economical representation of each of the N steps has been obtained, and the next question is to classify the observed results. This forms the objective of clustering analysis that relies on the mathematical theory of sets, as presented in the next module.