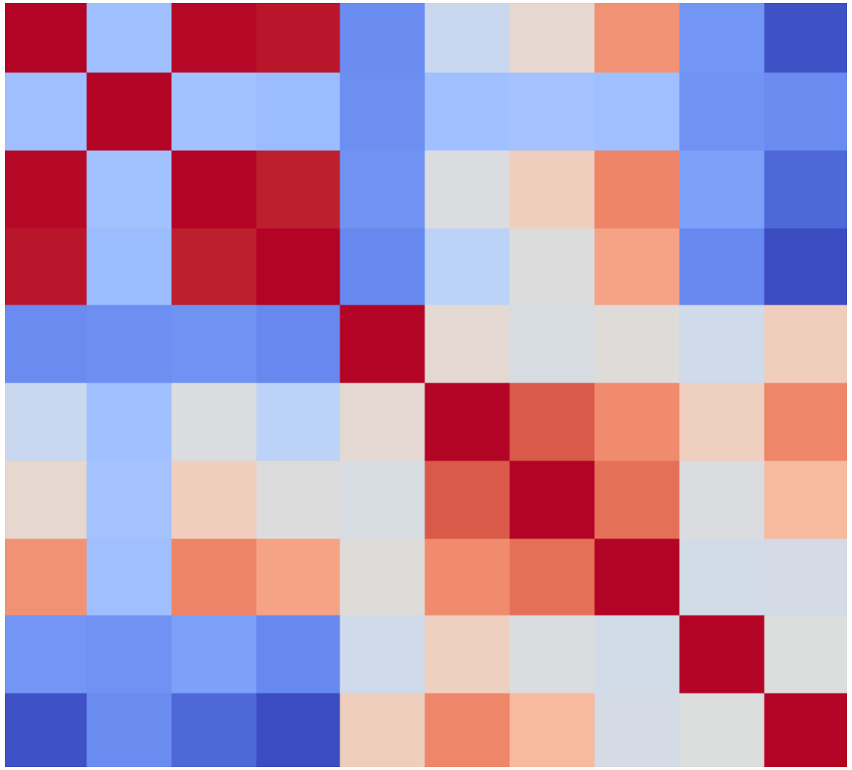
2.160 Identification, Estimation, and Learning

Part 1 Regression

Lecture 5

Principal Component Regression



H. Harry Asada
Department of Mechanical Engineering
MIT

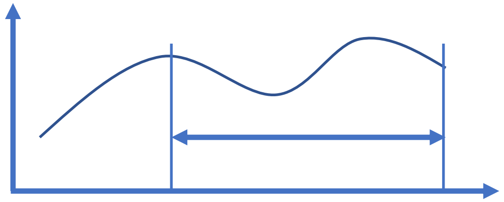
Rank-Deficit Data Matrix

In previous chapters on Least Squares Estimate and Recursive Least Squares, we assumed that

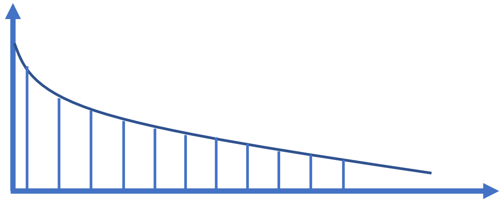
$$\sum_{i=1}^{t} \varphi(i)\varphi^T(i) = \; Non\text{-}singular, \;\; P_0 = Positive - Definite$$

This assumption may not be valid for:

- ❑ High input dimension: $m >> 1$
  - Temporal case, e.g. a wide time window, slowly-decaying FIR model

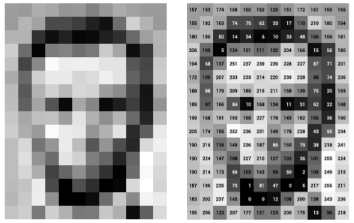


$$\varphi(t) = (y(t), \cdots, y(t-m))^T$$



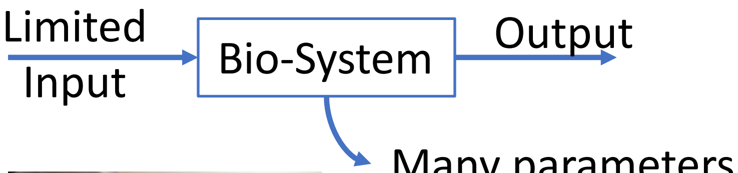
$$\varphi(t) = (u(t-1), \cdots, y(t-m))^T$$

  - Spatial case, e.g. image data



Pixels: m = (# of row) x (# of column)

- ❑ Limited sample size: $N < m$
  - Difficulty in obtaining a large number of data under consistent conditions, e.g. clinical trial data, biological experiment



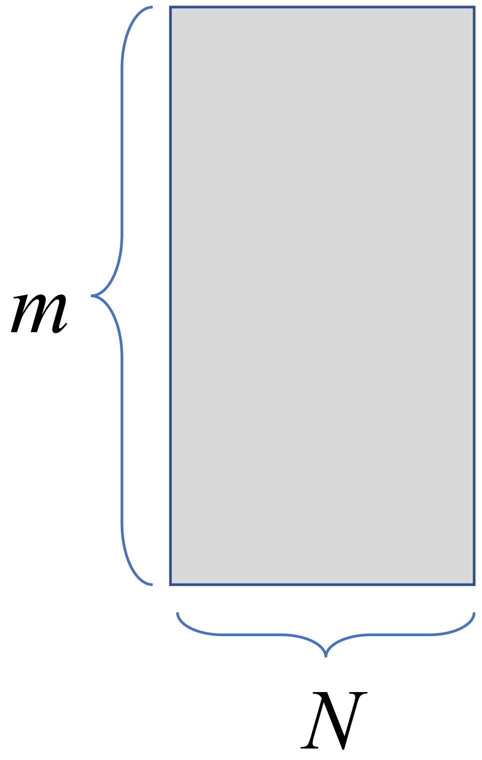
- ❑ When we do not know which input signals are important, we include as many potentially important inputs as possible.

Here we consider the case $N < m$, and/or $\sum_{i=1}^{t} \varphi(i)\varphi^T(i)$ is singular

Approach

$$\Phi = [\varphi(1), \cdots, \varphi(N)] \in \Re^{m \times N}$$

1. Use the pseudo-inverse
2. Regularization
    - Ridge Regression
3. Statistical Multivariate Analysis
    - Principal Component Regression
    - Partial Least Squares Regression



1. Quick Math : Moore-Penrose Pseudo-inverse

Consider a linear simultaneous equation: $Ax = b$ where A: m x n

- If there is no solution to $Ax = b$, then the pseudoinverse of A, denoted $A^{\#}$, minimizes the squared error $|Ax - b|^2$ and is given by

$$A^{\#} = (A^T A)^{-1} A^T$$

With this, the solution is $\hat{x} = A^{\#} b$.
This corresponds to the Least Squares solution:

$$\theta = \left(\sum_{i=1}^{t} \varphi(i) \varphi^T(i) \right)^{-1} \sum_{i=1}^{t} \varphi(i) y(i)$$

- If there are many (infinite) solutions to $Ax = b$, e.g. $n > m$, then the pseudoinverse $A^{\#}$ provides the solution that minimizes the square norm of vector x.

$$\min |x|^2 \quad \textit{subject to } Ax{=}b$$

$$\hat{x} = A^{\#} b \qquad \textit{Such that} \quad A\hat{x} = b \qquad \textit{and} \quad \min |x|^2$$

Definition

Given a matrix $A \in \mathbb{C}^{m \times n}$, the matrix $A^{\#} \in \mathbb{C}^{n \times m}$ that satisfies the following four conditions is called the Moore-Penrose Pseudoinverse:

1. $AA^{\#}A = A$
2. $A^{\#}AA^{\#} = A^{\#}$
3. $(AA^{\#})^* = AA^{\#}$
4. $(A^{\#}A)^* = A^{\#}A$

where $(X)^*$ is conjugate transpose of X.

A pseudoinverse is unique.

Exercise:

Show that the Least Squares Estimate below is given by the pseudoinverse of the matrix concatenating all the regressor vectors

$$\Phi = [\varphi(1), \cdots, \varphi(N)] \in \Re^{m \times N}$$

$$\theta = \left(\sum_{i=1}^{t} \varphi(i)\varphi^T(i) \right)^{-1} \sum_{i=1}^{t} \varphi(i) y(i)$$

$$\Downarrow$$

$$\theta = (\Phi^{\#})^T Y$$

where

$$Y = \begin{pmatrix} y(1) \\ \vdots \\ y(t) \end{pmatrix} \qquad \Phi^T \theta = Y$$

2. Ridge Regularization

Consider a cost functional (penalty): $V_R(\theta) = \left| Y - \Phi^T \theta \right|^2 + \eta \left| \theta \right|^2$

where $\eta > 0$ is a weight that determines the trade-off between the squared prediction error and the magnitude of the parameter vector.

Find the parameter that minimizes it: $\hat{\theta}_R = \arg\min_{\theta} V_R(\theta)$

$$\frac{dV_R(\theta)}{d\theta} = 0 \Rightarrow \frac{d}{d\theta}\left[\left(Y - \Phi^T \theta \right)^T \left(Y - \Phi^T \theta \right) + \eta \theta^T \theta \right] = \frac{d}{d\theta}\left[Y^T Y - 2 Y^T \Phi^T \theta + \theta^T \Phi \Phi^T \theta + \eta \theta^T \theta \right] = 0$$

This is only positive semi-definite; not invertible.

The identity matrix is positive definite.

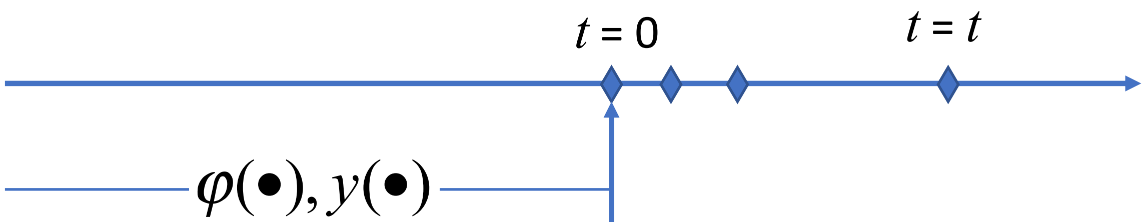
$$\Phi \Phi^T \theta + \eta \theta = \Phi Y \Rightarrow \underbrace{(\Phi \Phi^T + \eta I)}_{\text{Positive-definite; invertible}} \theta = \Phi Y$$

$$\boxed{\therefore\ \hat{\theta}_R = (\Phi \Phi^T + \eta I)^{-1} \Phi Y}$$

As long as $\eta > 0$, a unique solution is obtained.

Regularization is a standard technique when you have some knowledge about an appropriate solution.

Initial Conditions for Recursive Least Squares can be viewed as a type of regularization.



$$P_0^{-1} = \sum \varphi(\bullet)\varphi^T(\bullet),$$

$$B_0 = \sum \varphi(\bullet) y(\bullet),$$

$$\theta_0 = P_0 B_0$$

- Suppose that before $t = 0$, there were some data $\varphi(\bullet), y(\bullet)$.
- P_0, B_0, and θ_0 were computed based on these data.

- Consider the following cost function J_0. The parameter vector θ that minimizes J_0 is the initial condition θ_0.

$$J_0 = \frac{1}{2}(\theta - \theta_0)^T P_0^{-1}(\theta - \theta_0)$$

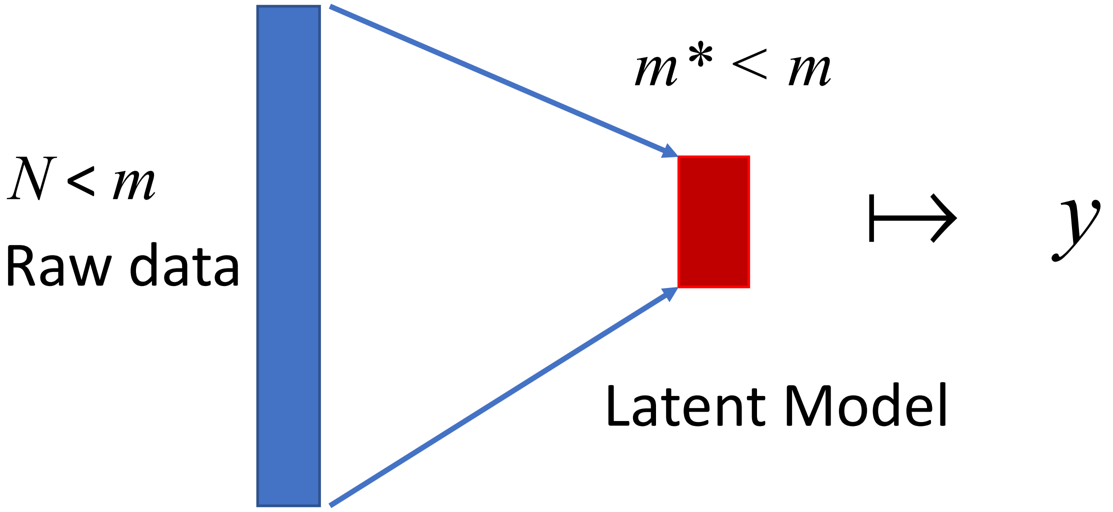
- As shown in Lecture Notes Chapter 2 Section 2.4, we can prove that the recursive least squares algorithm, starting with the initial conditions θ_0 and P_0 , is the optimal value that minimizes the following cost function.

$$J_t = \frac{1}{2}\sum_{i=1}^{t}(y(i) - \varphi^T(i)\cdot\theta)^2 + \underbrace{\frac{1}{2}(\theta - \theta_0)^T P_0^{-1}(\theta - \theta_0)}_{\text{Regularization}}$$

- The second term above can be viewed as a regularization term.

3. Statistical Multivariate Analysis

- ❑ Examine relationships among multiple variables in a high dimensional space;
- ❑ Joint behaviors of multiple variables may indicate some redundancy, and two variables may be linearly correlated: collinear.
- ❑ We are interested in extracting succinct, significant variables from high dimensional raw data;
- ❑ ***Latent*** Variable Method is to find hidden or encapsulated variables in the raw data that represent the raw data in a low dimensional space: Latent Model.
- ❑ Predict the output from the low-dimensional latent model
  - Principal Components Analysis and Regression
  - Partial Least Squares Regression

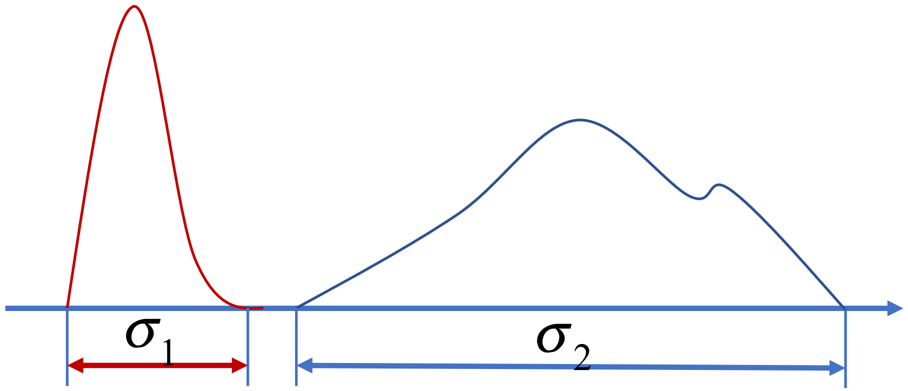


Prediction of output (dependent variables) from the low-dimensional latent model

$$y = \varphi^T \theta$$

Pre-Processing

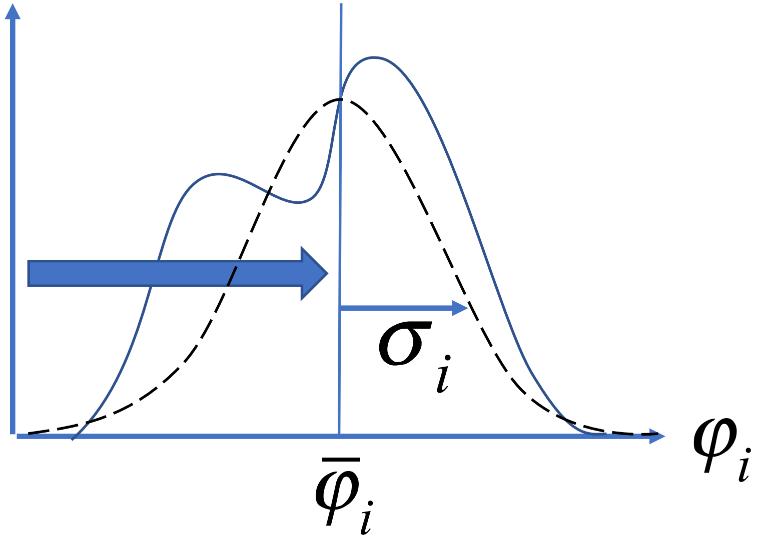
- ❑ A regressor vector contains multiple variables having different units and scales; Example: φ = (10 kg, 5 m/s, 2 cm)^{T}.
- ❑ To make an apple-to-apple comparison, the variables must be normalized.



❖ Mean-centering

Shift the origin to the mean $\bar{\varphi}_i$

$$\varphi_i \quad \mapsto \quad \varphi_i - \bar{\varphi}_i$$



❖ Normalization with some reference value

$$x_i = \frac{\varphi_i - \bar{\varphi}_i}{\sigma_i}$$

reference value: σ_i

- Standard deviation
- Max – Min (dynamic range)

$$x = \begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix} \in \Re^{m\times 1}$$

Input Data Matrix

$$X = \begin{bmatrix} x(1) & \cdots & x(N) \end{bmatrix} \in \Re^{m\times N}$$

Output Data Matrix

$$Y = \begin{bmatrix} y(1) \\ \vdots \\ y(N) \end{bmatrix} \in \Re^{N\times 1}$$

4.1 Principal Component Regression: A Multi-Input, Single-Output Case

- Examine how input data are distributed in the m-dim space and reduce the space to a low dimensional space;
- The distribution can be characterized with the covariance of preprocessed input data:

$$X = \begin{bmatrix} x(1) & \cdots & x(N) \end{bmatrix} \in \Re^{m \times N}$$

Recall we have defined Covariance of two scalar random variables, X and Y, as

$$\operatorname{cov}(X,Y) = E[(X - E[X])(Y - E[Y])]$$

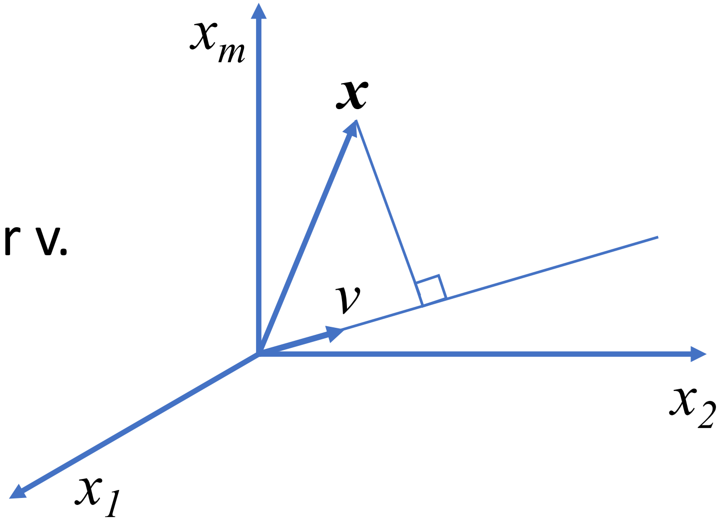
Covariance of a <u>*vector*</u> consists of covariances of all the combinations of the vector components.

$$C = \operatorname{cov}\left[\begin{pmatrix} x_1 \\ \vdots \\ x_m \end{pmatrix}\begin{pmatrix} x_1 & \cdots & x_m \end{pmatrix}\right] = \begin{bmatrix} E[x_1^2] & \cdots & E[x_1 x_m] \\ \vdots & \ddots & \vdots \\ E[x_m x_1] & \cdots & E[x_m^2] \end{bmatrix} = E[XX^T]$$

Note that input data have been mean-centered.

Principal Component Regression - 2

- ❑ What is the physical sense of the covariance matrix?
- ❑ How can we use it for latent modeling?



Recap Examine the strength of signals in the direction of unit vector v.

$$J = \frac{1}{N}\sum_{i=1}^{N}\left|v^T x(i)\right|^2 = \frac{1}{N}\sum v^T x(i)\cdot x^T(i)v = \frac{1}{N}v^T\left(\sum x(i)\cdot x^T(i)\right)v$$

$$= \frac{1}{N}v^T \underbrace{\begin{bmatrix} x(1) & \cdots & x(N) \end{bmatrix}}_{X} \begin{bmatrix} x^T(1) \\ \vdots \\ x^T(N) \end{bmatrix} v = \frac{1}{N}v^T X X^T v$$

The j-k component of matrix XX^T $\frac{1}{N}\left\{x_j(1)x_k(1)+\cdots+x_j(N)x_k(N)\right\} \cong E[x_j x_k]$ $\therefore J \cong v^T C v$

Find the strongest direction of signal in the input space : an eigenvalue problem.

$$v = \arg\max_v \left(v^T X X^T v - \lambda(v^T v - 1)\right) \Rightarrow XX^T v = \lambda v$$

Eigenvalues: $\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$

Principal Component Regression - 3

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_m \geq 0$$

We are considering the case where $XX^T \cong C$: singular.

Suppose that the last $(m - m^*)$ eigenvalues are zero.

$$\lambda_{\max} = \lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_{m^*} > 0 \qquad \lambda_m = 0, \lambda_{m-1} = 0, \lambda_{m^*+1} = 0$$

$$XX^T = [v_1, \cdots, v_m] \begin{bmatrix} \lambda_1 & \cdots & 0 & \cdots & 0 \\ \vdots & \ddots & & & \vdots \\ 0 & & \lambda_{m^*} & & 0 \\ \vdots & & & 0 & \vdots \\ 0 & \cdots & 0 & \cdots & 0 \end{bmatrix} \begin{bmatrix} v_1^T \\ \vdots \\ v_m^T \end{bmatrix}$$

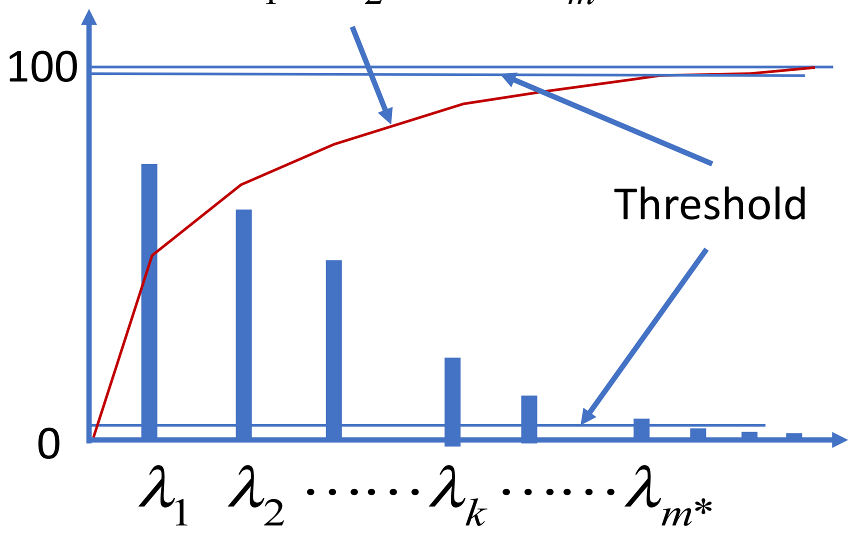
$$\therefore XX^T = \lambda_1 v_1 v_1^T + \cdots + \lambda_{m^*} v_{m^*} v_{m^*}^T + 0 + \cdots + 0$$

The first m* components can generate and fully explain the data.

In practice, some eigenvalues are small but not exactly zero. We can truncate them with a threshold.

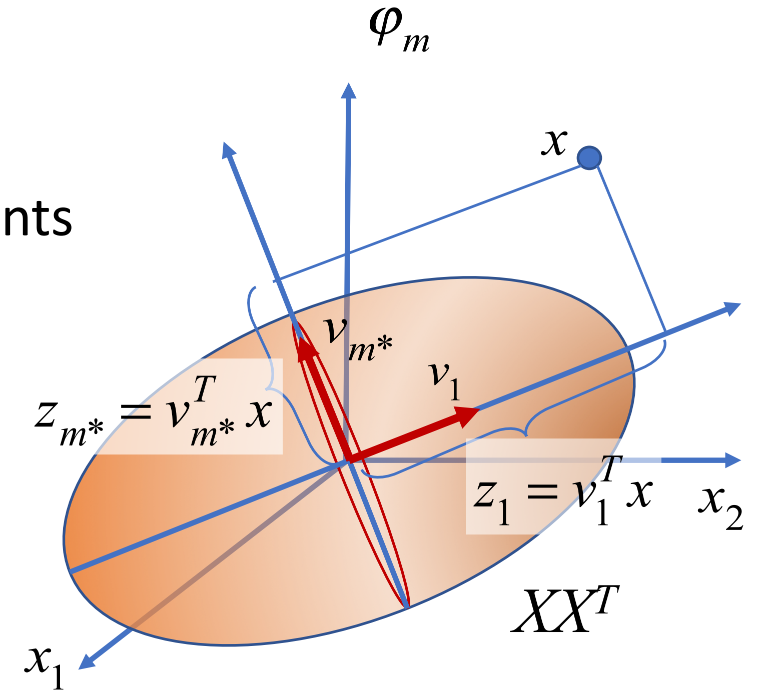
Percent Accuracy: a measure for truncation

$$\mu_k = \frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_m} \times 100\%$$



Truncate the series at $m^* < m$.

Principal Component Regression – 4 (PCR)



Step 1. Reduce the input space and represent it with Principal Components

$$\therefore XX^T = \lambda_1 v_1 v_1^T + \cdots + \lambda_{m*} v_{m*} v_{m*}^T + 0 + \cdots + 0$$

Define $m^* < m$ Latent Variables using eigenvectors associated with the top m^* most significant eigenvalues

$$z_1 = v_1^T x,\ z_2 = v_2^T x,\ \cdots,\ z_{m*} = v_{m*}^T x$$

Step 2. Construct a low-dimensional regression on the Principal Components

$$\hat{y} = b_1 z_1 + b_2 z_2 + \cdots + b_{m*} z_{m*} \qquad m* < m$$

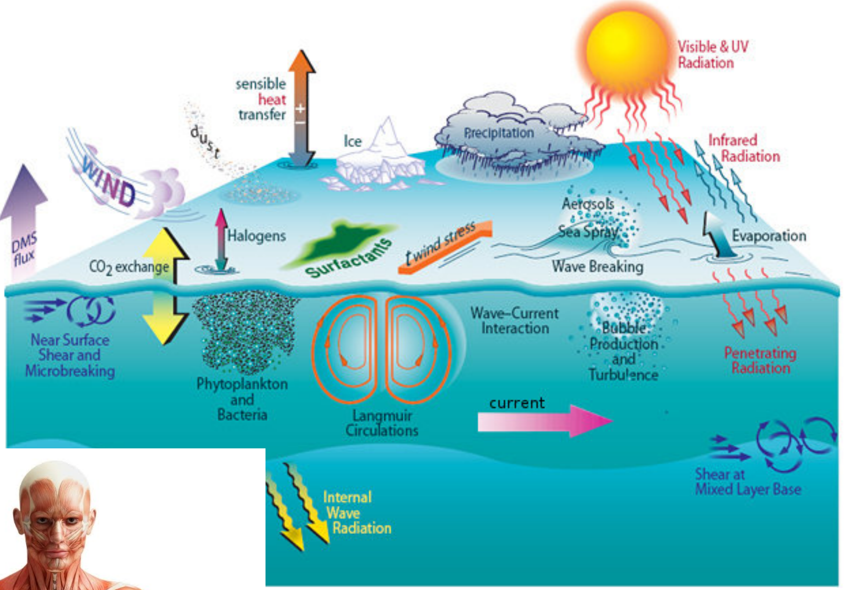
$$\theta = \begin{pmatrix} b_1 \\ b_2 \\ \vdots \\ b_{m*} \end{pmatrix} = \Bigg[\underbrace{\sum_i \begin{pmatrix} z_1(i) \\ \vdots \\ z_{m*}(i) \end{pmatrix} \begin{pmatrix} z_1(i) & \cdots & z_{m*}(i) \end{pmatrix}}_{\text{Non-Singular}} \Bigg]^{-1} \Bigg[\sum_i \begin{pmatrix} z_1(i) \\ \vdots \\ z_{m*}(i) \end{pmatrix} y(i) \Bigg]$$

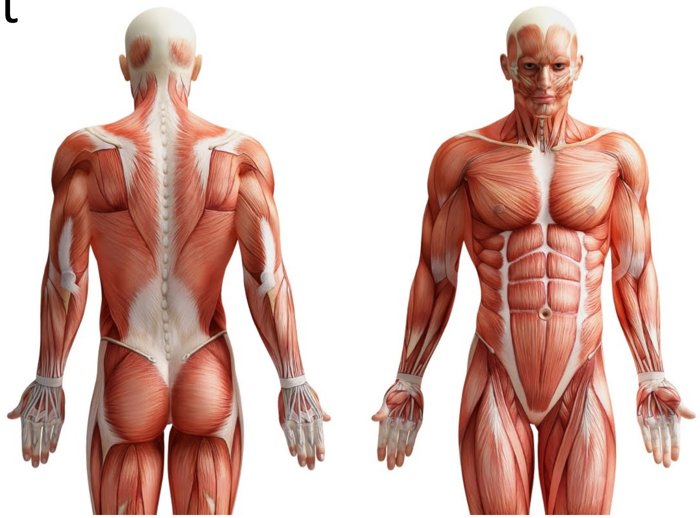
Applications

Where is PCR needed (other than COVID testing)?
These days we deal with large data.

- ❑ High input dimension: m >> 1. Initially we do not know which variables are significant. We include many possible variables in the input space.
- ❑ Grid sampling of distributed systems
    - e.g. Ocean monitoring
    - e.g. Weather forecast
- ❑ Human body movements
    - The human has over 200 muscles.
    - A single hand has at least 19 joints.
- ❑ Bio informatics
    - ▪ e.g. Proteome
    - The entire organism of a yeast fungus contains 4,399 proteins (2006).

Woods Hole Oceanographic Institution



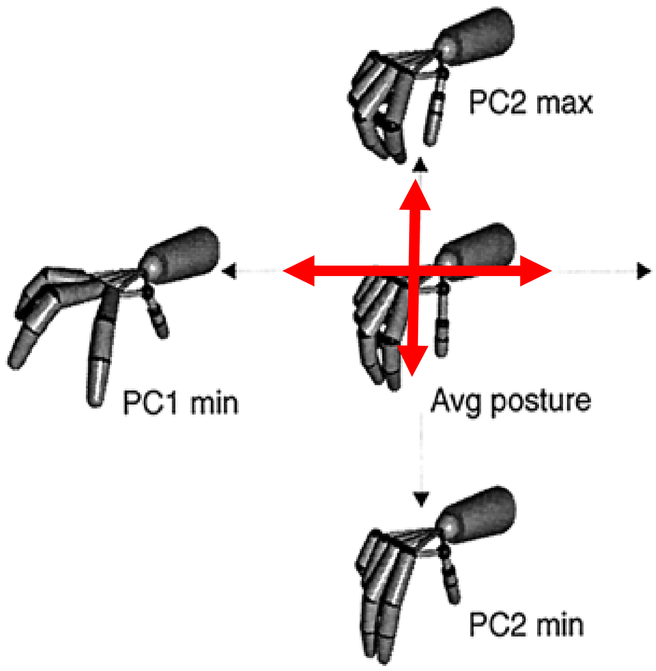


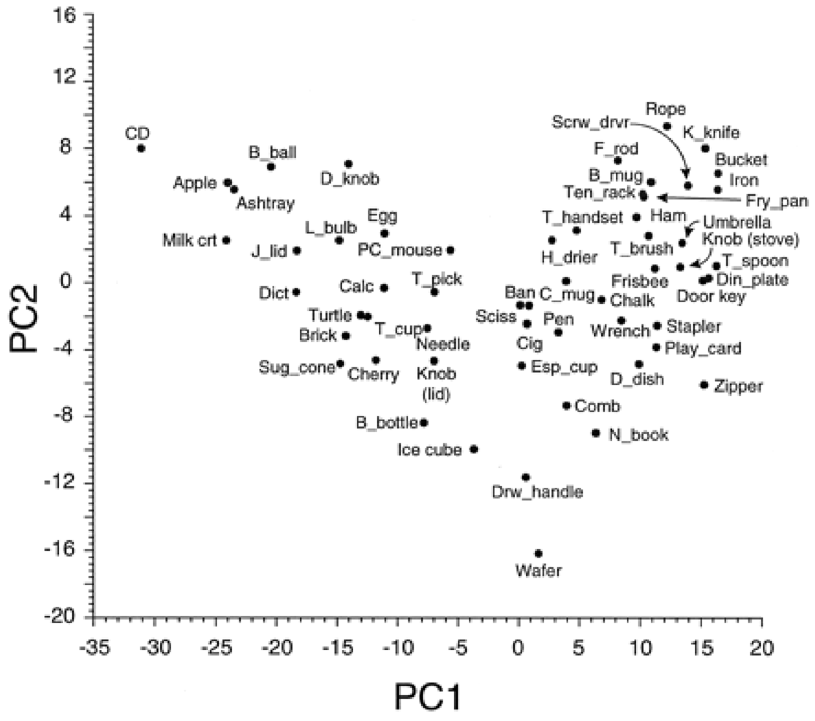
North West Training Vocational College

Max Plunk Institute

Context-Oriented Project of 2019 (Not an assignment for this year)

- At least 19 joints are involved in a single human hand, but the movements of the five fingers are highly coordinated, following some functional relationship. The central nerve system does not control each muscle individually, but control them as a group. In biomechanics and motor control literature, this is called “Synergy”.
- This means that 19-dimensional data of the joint angles, when grasping daily choir items, for example, are distributed in a much smaller subspace. This can be analyzed by using Principal Components.
- Santello, et al measured the posture of five fingers when grasping 100 daily chore items. Two principal components explain over 80 % of data.





M.Santello, M.Flanders, J.F.Soechting, “Postural synergies for tool use,” Journal of Neuroscience. Volume 18: 10105-10115, 1998 Dec.1

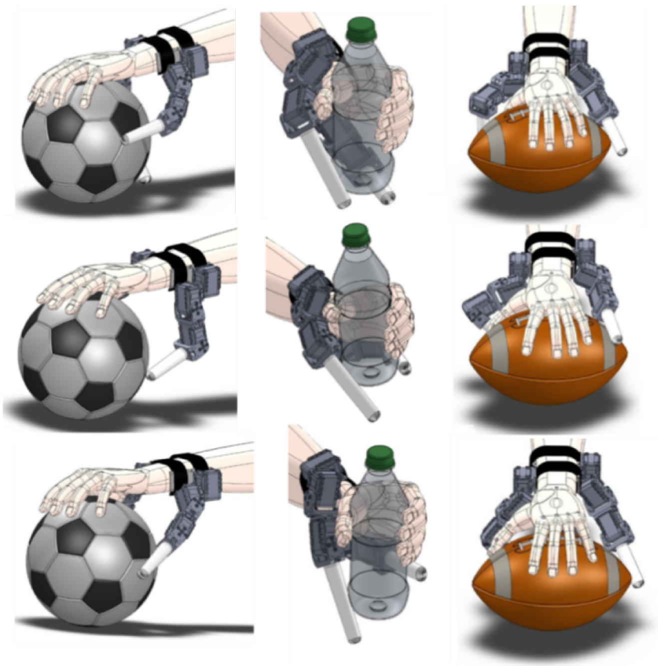
Applying the grasp posture synergy theory to the 7-fingered hand

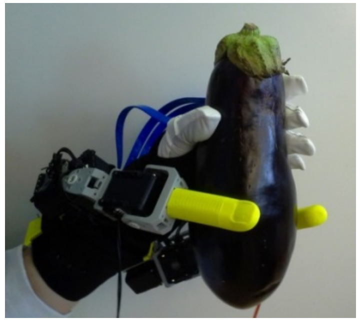
Hybrid Human-Robot Finger System

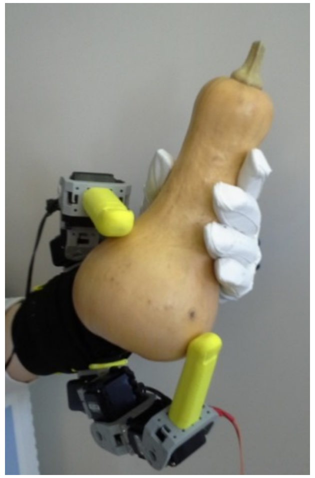
5 + 2 = 7

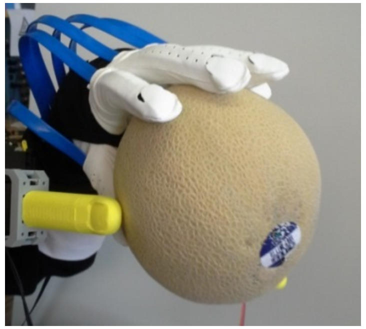
Hypotheses

1. There exists a type of "Synergy" among the 7-fingers.
2. The SR Fingers can be controlled in correlation with the five fingers
   - Control Supernumerary Robotic Finger based on human finger posture measurements



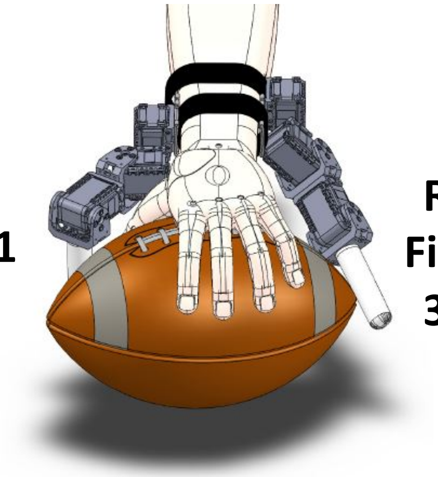






Grasp Posture Data: 7 fingers (5 human fingers + 2 robotic fingers)

Robot Finger 1 3 DOF



Robot Finger 2 3 DOF

$$x = \left.\begin{array}{l}[\text{Joint angles of thumb;} \\ \quad \text{index;} \\ \quad \text{middle;} \\ \quad \text{ring; and} \\ \quad \text{pinky fingers;}\end{array}\right\} \text{19 joints}$$

$$\left.\begin{array}{l}\quad \text{robot finger 1;} \\ \quad \text{robot finger 2]}\end{array}\right\} \text{6 joints}$$

25 dimensional vector

Data Set*

$$\mathbf{X} = \left.\begin{bmatrix} \overbrace{x_1^1, x_2^1, \cdots, x_{19}^1}^{\text{Human}} & \overbrace{x_{20}^1, x_{21}^1, \cdots, x_{25}^1}^{\text{Robot}} \\ x_1^2, \cdots\cdots\cdots\cdots & \cdots\cdots\cdots\cdots, \; x_{25}^2 \\ \cdots\cdots\cdots\cdots & \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots & \cdots\cdots\cdots\cdots \\ x_1^{40}, \cdots\cdots\cdots\cdots & \cdots\cdots\cdots\cdots, x_{25}^{40} \end{bmatrix}\right\} \text{40 observations}$$

$\xrightarrow{\text{Mean-Centering}}$

Data Covariance

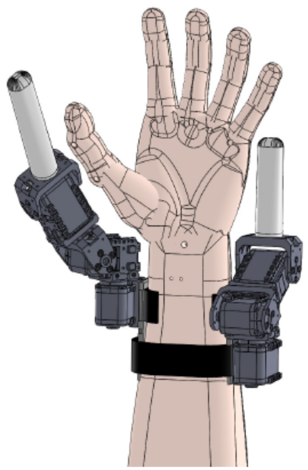
$$\mathrm{cov}(X) = E[(X - \bar{X})(X - \bar{X})]$$

$$\cong \frac{1}{N-1} \sum_{i=1}^{N} (\boldsymbol{x}^i - \bar{\boldsymbol{x}})^T (\boldsymbol{x}^i - \bar{\boldsymbol{x}})$$

* Note that each measurement of 25 joint angles forms a row vector.

PCA of 7-finger grasps

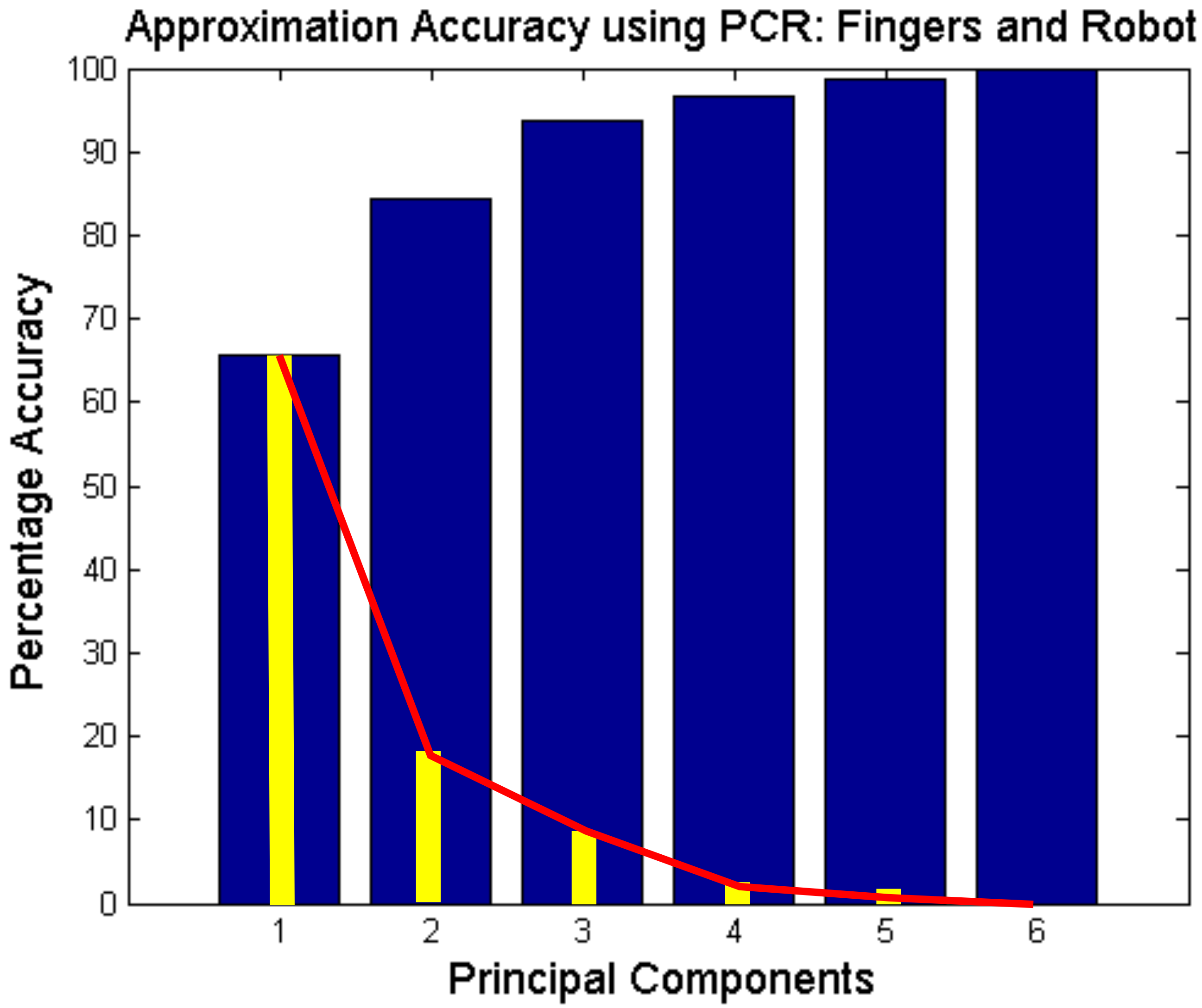
- ❑ Input space = 25 dimensional space
- ❑ Only the first 3~4 components span the data space.



Eigenvalues

$$\lambda_1 \geq \lambda_2 \geq \cdots \geq \lambda_n$$

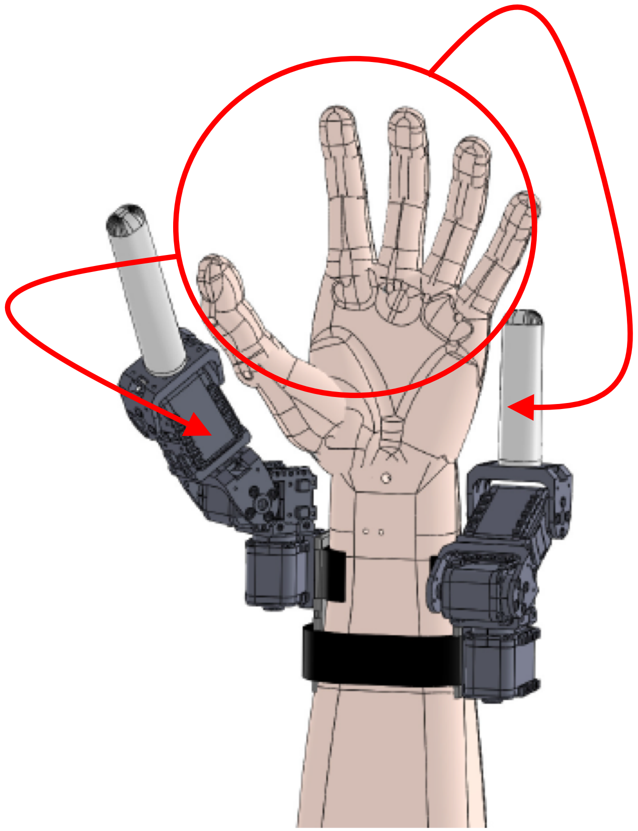
$$\mu = \frac{\lambda_1 + \cdots + \lambda_r}{\lambda_1 + \cdots\cdots + \lambda_n} \times 100\%$$



Now that postural synergies exist for the 7 fingers, can the SR* Fingers be controlled in correlation with the human fingers?

Hypotheses 2

If so, how?



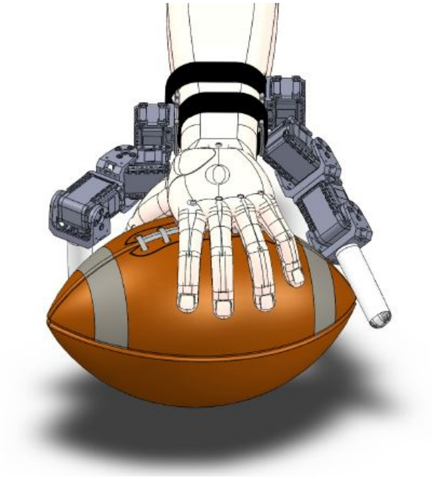
* Supernumerary Robotic (SR) Fingers

A new mathematical tool is needed for the robot finger control.

Data Set

Splitting the data set into Input and Output matrices

$$\mathbf{X} = \left[\begin{array}{c|c} \overbrace{\begin{array}{c} x_1^1, x_2^1, \cdots, x_{19}^1 \\ x_1^2, \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ x_1^{40} \cdots\cdots\cdots\cdots \end{array}}^{\text{Human}} & \overbrace{\begin{array}{c} x_{20}^1, x_{21}^1, \cdots, x_{25}^1 \\ \cdots\cdots\cdots\cdots, \ x_{25}^2 \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots \\ \cdots\cdots\cdots\cdots, x_{25}^{40} \end{array}}^{\text{Robot}} \end{array}\right]$$



Input

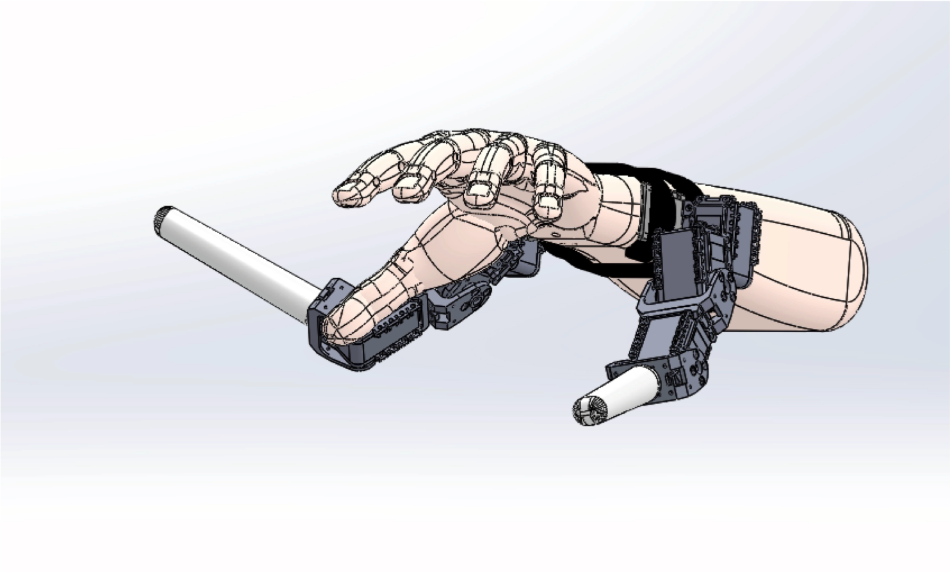
$$\mathbf{X} = \underbrace{\begin{bmatrix} x_1^1, x_2^1, \cdots\cdots\cdots x_{19}^1 \\ x_1^2, \cdots\cdots\cdots\cdots x_{19}^2 \\ \cdots\cdots\cdots\cdots \\ x_1^{40} \cdots\cdots\cdots\cdots x_{19}^{40} \end{bmatrix}}_{\text{Human Fingers: 19 joints}}$$

Principal Component Regression (PCR)
Or
Partial Least Squares Regression

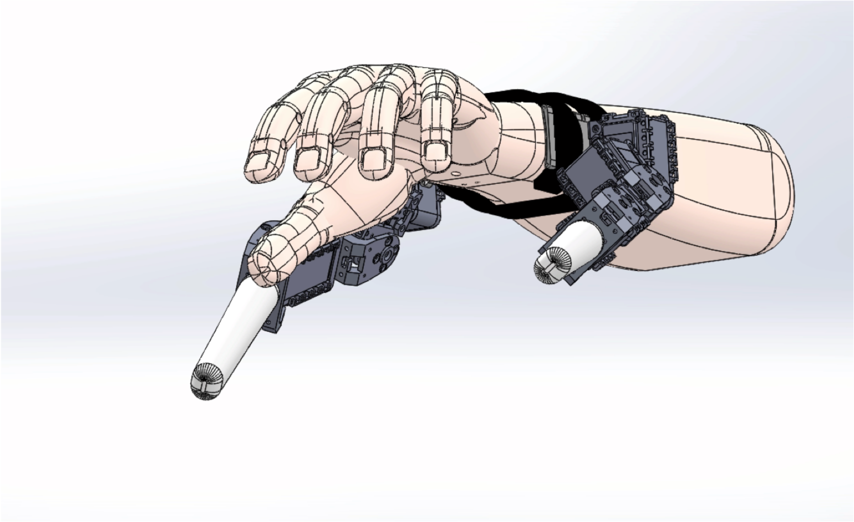
$$\mathbf{Y} = \underbrace{\begin{bmatrix} x_{20}^1, x_{21}^1, \cdots\cdots, x_{25}^1 \\ x_{20}^2 \cdots\cdots\cdots\cdots, x_{25}^2 \\ \cdots\cdots\cdots\cdots \\ x_{20}^{40} \cdots\cdots\cdots\cdots, x_{25}^{40} \end{bmatrix}}_{\text{Robot Fingers: 6 joints}}$$

Output

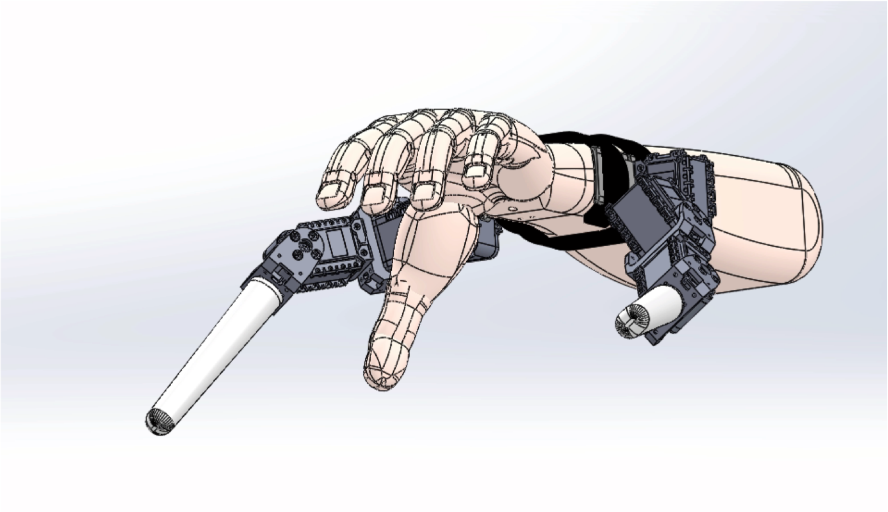
Principal Component Regression



1st component



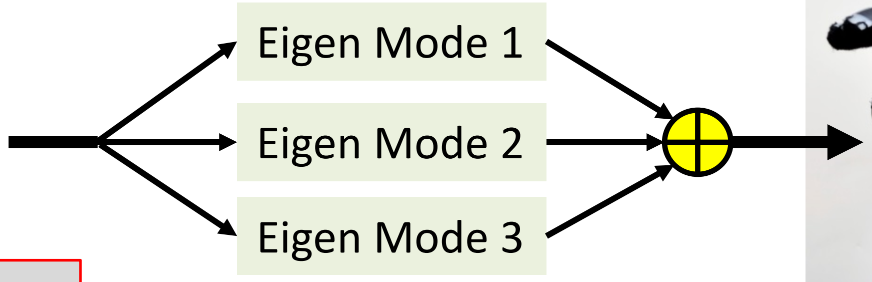
2nd component

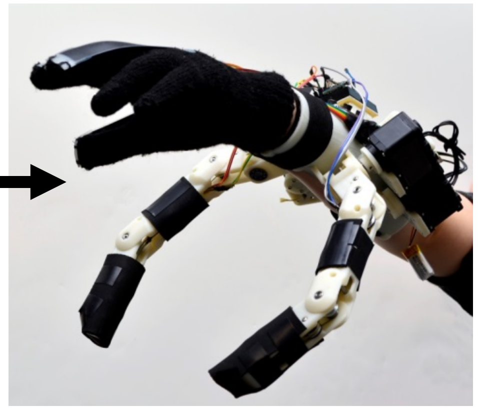


3rd component

Implementation

Intuitive, implicit control of the Extra Fingers

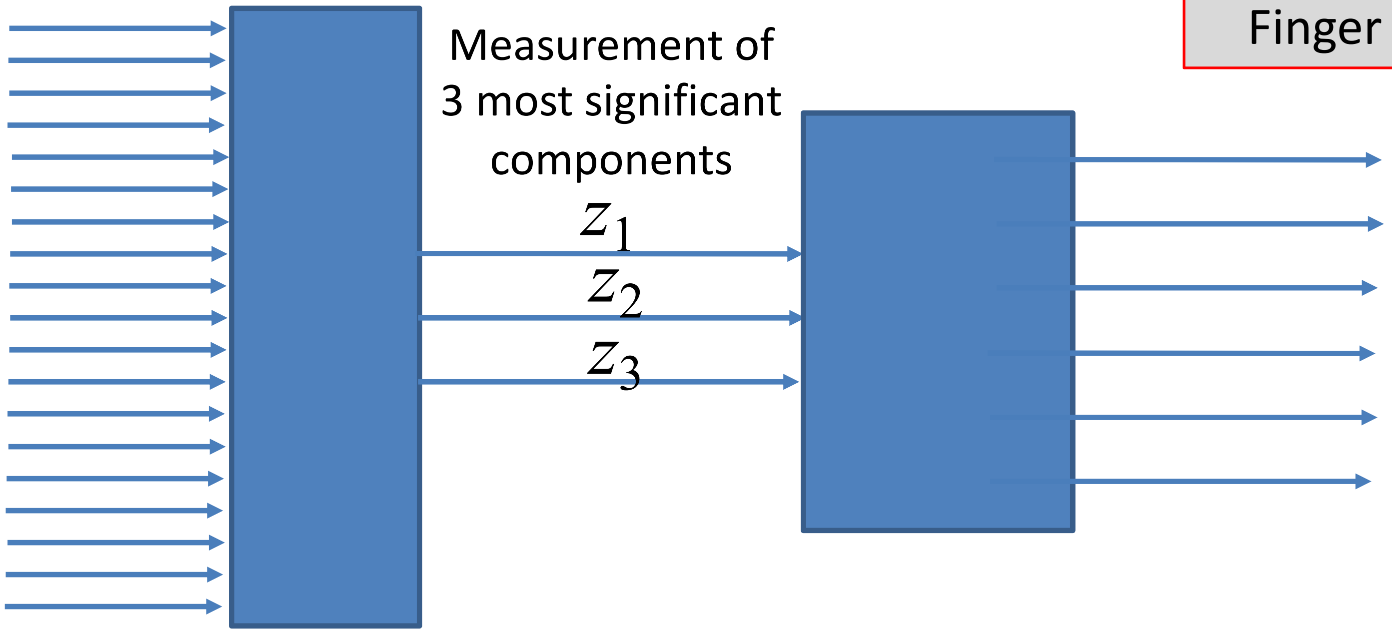




Estimation of Latent Variables

Map to Robot Finger Posture

Human Finger Posture: 19 dim.



Robot Joints 6 DOF