Computational Intelligence Laboratory

Lecture 1

Principal Component Analysis

Thomas Hofmann

ETH Zurich – `cil.inf.ethz.ch`

24 February & March 3, 2017

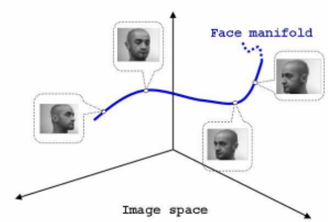
Section 1

Dimension Reduction

Motto

Before **computation** comes **calculation**.

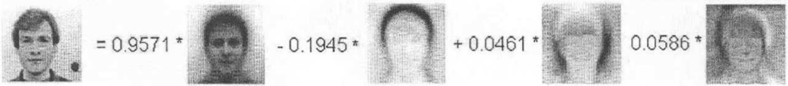
Dimension Reduction



- Dimension reduction
  - given (high-dimensional) data points $\{\mathbf{x}_i \in \mathbb{R}^m : i = 1, \ldots, n\}$
  - map data to **low-dimensional vector space**
  - more generally: recover/learn a **data manifold**

Dimension Reduction: Example

- Example: **face images**
  - 2D pixel fields, e.g. $\mathbf{x}_i \in \mathbb{R}^{100\times100}$;
  - often treated as vectors $\mathbb{R}^{100\times100} \simeq \mathbb{R}^{10000}$
  - approximate image by linear combination of basis images



(from: Turk and Pentland, Eigenfaces for Recognition, 1991)

  - images (on r.h.s.) $=$ basis of 4-dim subspace
  - coefficients $=$ 4-dimensional representation

Dimension Reduction: Motivation

- **Motivation**
  - visualization – e.g. 2D or 3D
  - data compression – fewer coefficients
  - data reconstruction – filling in missing data
  - signal recovery – removing noise
  - discover modes of variation – intrinsic properties of data
  - feature discovery – learn better representations

Section 2

1D Linear Case

Line in $\mathbb{R}^m$

- Parametric form of a line in $\mathbb{R}^m$

$$\boldsymbol{\mu} + \mathbb{R}\mathbf{u} \equiv \{\mathbf{v} \in \mathbb{R}^m : \exists z \in \mathbb{R} \text{ s.t. } \mathbf{v} = \boldsymbol{\mu} + z\mathbf{u}\}$$

  - $\boldsymbol{\mu}$: offset or shift
  - $\mathbf{u}$: direction vector
  - $\mathbf{u}$ "is direction" $\iff \|\mathbf{u}\| = 1$
  - $\|\cdot\|$ or $\|\cdot\|_2$: Euclidean vector norm, $\|\mathbf{v}\|^2 = \sum_j v_j^2 = \langle \mathbf{v}, \mathbf{v} \rangle$
  - $\langle \cdot, \cdot \rangle$: inner or dot product, $\langle \mathbf{u}, \mathbf{v} \rangle = \mathbf{u}^\top \mathbf{v} = \sum_j u_j v_j$

Orthogonal Projection (1 of 2)

- Approximate data point $\mathbf{x} \in \mathbb{R}^m$ by a point on the line
- Minimize (squared) Euclidean distance
- Formally:

$$\text{Dimension Reduction} \quad \leftarrow \arg\min_{z \in \mathbb{R}} \|\boldsymbol{\mu} + z\mathbf{u} - \mathbf{x}\|^2$$

or

$$\text{Reconstruction} \quad \leftarrow \arg\min_{\hat{\mathbf{x}} \in \boldsymbol{\mu} + \mathbb{R}\mathbf{u}} \|\hat{\mathbf{x}} - \mathbf{x}\|^2$$

- We know the answer! ❣

Orthogonal Projection (1 of 2)

- Approximate data point $\mathbf{x} \in \mathbb{R}^m$ by a point on the line
- Minimize (squared) Euclidean distance
- Formally:

$$\text{Dimension Reduction} \quad \leftarrow \arg\min_{z \in \mathbb{R}} \|\boldsymbol{\mu} + z\mathbf{u} - \mathbf{x}\|^2$$

or

$$\text{Reconstruction} \quad \leftarrow \arg\min_{\hat{\mathbf{x}} \in \boldsymbol{\mu} + \mathbb{R}\mathbf{u}} \|\hat{\mathbf{x}} - \mathbf{x}\|^2$$

- We know the answer! ❣ Orthogonal projection.

Orthogonal Projection (2 of 2)

- Warm-up exercise: first order optimality condition

$$\frac{d}{dz}\|\boldsymbol{\mu} + z\mathbf{u} - \mathbf{x}\|^2 = 2\langle \boldsymbol{\mu} + z\mathbf{u} - \mathbf{x}, \mathbf{u}\rangle \overset{!}{=} 0$$

$$\iff \underbrace{\langle \mathbf{u}, \mathbf{u}\rangle}_{\|\mathbf{u}\|^2=1} z \overset{!}{=} \langle \mathbf{x} - \boldsymbol{\mu}, \mathbf{u}\rangle$$

- Solution(s):

$$z = \langle \mathbf{x} - \boldsymbol{\mu}, \mathbf{u}\rangle$$

$$\hat{\mathbf{x}} = \boldsymbol{\mu} + \langle \mathbf{x} - \boldsymbol{\mu}, \mathbf{u}\rangle\, \mathbf{u}$$

- Procedure: (1) shift by $-\boldsymbol{\mu}$, (2) project onto $\mathbf{u}$, (3) shift back by $\boldsymbol{\mu}$

Optimal Line: Formulation

- Assume we are given data points $\{\mathbf{x}_1, \dots, \mathbf{x}_n\} \subset \mathbb{R}^m$.
- What is their optimal approximation by a line?
  - use orthogonal projection result

$$(\mathbf{u}, \boldsymbol{\mu}) \leftarrow \arg\min \left[\frac{1}{n} \sum_{i=1}^{n} \| \underbrace{\boldsymbol{\mu} + \langle \mathbf{x}_i - \boldsymbol{\mu}, \mathbf{u} \rangle \mathbf{u}}_{=\hat{\mathbf{x}}_i} - \mathbf{x}_i \|^2 \right]$$

$$= \left[\frac{1}{n} \sum_{i=1}^{n} \left\| \left(\mathbf{I} - \mathbf{u}\mathbf{u}^\top \right) (\mathbf{x}_i - \boldsymbol{\mu}) \right\|^2 \right]$$

  - some simple algebra
  - exploit identity $\langle \mathbf{v}, \mathbf{u} \rangle \mathbf{u} = (\mathbf{u}\mathbf{u}^\top)\mathbf{v}$

there is a error on the slide <u, v>u = uu’v

I minus U2?

- What does this matrix represent? $\left(\mathbf{I} - \mathbf{u}\mathbf{u}^\top\right)$
  - in general: a matrix represents a linear map (in specific basis)
- Specifically: take argument $\mathbf{v}$, we get

$$\left(\mathbf{I} - \mathbf{u}\mathbf{u}^\top\right)\mathbf{v} = \mathbf{v} - \underbrace{\langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u}}_{\text{projection}}$$

  - so this is the vector itself minus the projection to the line $\mathbb{R}\mathbf{u}$
  - which is the projection to the orthogonal complement $(\mathbb{R}\mathbf{u})^\perp$
  - it is idempotent, because

$$\left(\mathbf{u}\mathbf{u}^\top\right)[\mathbf{v} - \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u}] = \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u} - \langle \mathbf{u}, \mathbf{v} \rangle \mathbf{u} = \mathbf{0}$$

Optimal Line: Solving for μ

- First order optimality condition for $\boldsymbol{\mu}$

$$\nabla_{\boldsymbol{\mu}}[\cdot] \stackrel{!}{=} 0 \iff \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{I}-\mathbf{u}\mathbf{u}^\top\right)(\mathbf{x}_i-\boldsymbol{\mu}) \stackrel{!}{=} 0$$

$$\iff \left(\mathbf{I}-\mathbf{u}\mathbf{u}^\top\right)\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \stackrel{!}{=} \left(\mathbf{I}-\mathbf{u}\mathbf{u}^\top\right)\boldsymbol{\mu}$$

  - does not determine $\boldsymbol{\mu}$ uniquely ❣

Optimal Line: Solving for μ

- First order optimality condition for $\boldsymbol{\mu}$

$$\nabla_{\boldsymbol{\mu}}[\cdot] \stackrel{!}{=} 0 \iff \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{I}-\mathbf{u}\mathbf{u}^{\top}\right)(\mathbf{x}_i-\boldsymbol{\mu}) \stackrel{!}{=} 0$$

$$\iff \left(\mathbf{I}-\mathbf{u}\mathbf{u}^{\top}\right)\frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \stackrel{!}{=} \left(\mathbf{I}-\mathbf{u}\mathbf{u}^{\top}\right)\boldsymbol{\mu}$$

  - does not determine $\boldsymbol{\mu}$ uniquely ❢
  - however, there is a unique (simultaneous) solution for all $\mathbf{u}$:

$$\boldsymbol{\mu} = \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i \quad \equiv \quad \textbf{sample mean}$$

Optimal Line: Conclusion #1

- By **centering** the data:

$$\mathbf{x}_i \leftarrow \mathbf{x}_i - \frac{1}{n}\sum_{i=1}^{n}\mathbf{x}_i$$

  - restrict to **linear** (instead of affine) subspaces
  - identify center of mass of data with origin
  - simplifies derivations and analyses
  - w.l.o.g.: assume data points are **centered**

Optimal Line: Solving for $\mathbf{u}$ (1 of 3)

- We are left with

$$\mathbf{u} \leftarrow \arg\min_{\|\mathbf{u}\|=1} \left[\frac{1}{n} \sum_{i=1}^{n} \|\langle \mathbf{u}, \mathbf{x}_i \rangle \mathbf{u} - \mathbf{x}_i\|^2 \right]$$

- Expanding the squared norm
  - general formula

$$\|\mathbf{v} - \mathbf{w}\|^2 = \langle \mathbf{v} - \mathbf{w}, \mathbf{v} - \mathbf{w} \rangle = \|\mathbf{v}\|^2 + \|\mathbf{w}\|^2 - 2\langle \mathbf{v}, \mathbf{w} \rangle$$

  - yields: const $- \langle \mathbf{u}, \mathbf{x} \rangle^2$ as

$$\begin{aligned} \|\langle \mathbf{u}, \mathbf{x} \rangle \mathbf{u}\|^2 &= \langle \mathbf{u}, \mathbf{x} \rangle^2 \\ \|\mathbf{x}\|^2 &= \text{const.} \\ -2\langle \langle \mathbf{u}, \mathbf{x} \rangle \mathbf{u}, \mathbf{x} \rangle &= -2\langle \mathbf{u}, \mathbf{x} \rangle^2 \end{aligned}$$

Optimal Line: Solving for $\mathbf{u}$ (2 of 3)

- We can equivalently solve

$$\mathbf{u} \leftarrow \underset{\|\mathbf{u}\|=1}{\arg\max} \left[\frac{1}{n} \sum_{i=1}^{n} \langle \mathbf{u}, \mathbf{x}_i \rangle^2 \right]$$

$$= \left[\mathbf{u}^\top \left(\frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top \right) \mathbf{u} \right]$$

- Key matrix: **variance-covariance matrix** of the data sample

$$\mathbf{\Sigma} \equiv \frac{1}{n} \sum_{i=1}^{n} \mathbf{x}_i \mathbf{x}_i^\top = \frac{1}{n} \mathbf{X}\mathbf{X}^\top, \quad \mathbf{X} \equiv [\mathbf{x}_1 \ldots \mathbf{x}_n]$$

Hold on: Baby Step Derivation

$$
\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{u}, \mathbf{x}_i \rangle^2 &= \frac{1}{n}\sum_{i}\left[\left(\sum_{j} u_j x_{ij}\right)\left(\sum_{k} u_k x_{ik}\right)\right] \\
&= \frac{1}{n}\sum_{i}\sum_{j,k} u_j u_k x_{ij} x_{ik} \\
&= \sum_{j,k} u_j u_k \left(\frac{1}{n}\sum_{i} x_{ij} x_{ik}\right) = \mathbf{u}^\top \underbrace{\left(\frac{1}{n}\sum_{i} x_{ij} x_{ik}\right)_{j,k}}_{\text{matrix}} \mathbf{u} \\
&= \mathbf{u}^\top \left(\frac{1}{n}\sum_{i} \mathbf{x}_i \mathbf{x}_i^\top\right) \mathbf{u} = \mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u}
\end{aligned}
$$

Optimal Line: Solving for $\mathbf{u}$ (3 of 3)

- Constrained optimization with Lagrange multiplier λ

$$\mathcal{L}(\mathbf{u}, \lambda) = \mathbf{u}^\top \boldsymbol{\Sigma} \mathbf{u} + \lambda \langle \mathbf{u}, \mathbf{u} \rangle$$

- Minimize over $\mathbf{u} \implies \mathbf{u}$ is an **eigenvector** of $\boldsymbol{\Sigma}$

$$\nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}) = 2(\boldsymbol{\Sigma}\mathbf{u} - \lambda \mathbf{u})$$

$$\nabla_{\mathbf{u}} \mathcal{L}(\mathbf{u}) \overset{!}{=} 0 \iff \boldsymbol{\Sigma}\mathbf{u} = \lambda \mathbf{u}$$

- Maximize over $\lambda \implies \mathbf{u}$ is a **principal** eigenvector of $\boldsymbol{\Sigma}$ (one with the largest eigenvalue λ)

Linear Algebra: Eigen-{Values & Vectors}

- Let $\mathbf{A}$ be a squared matrix, $\mathbf{A} \in \mathbb{R}^{m \times m}$.
- $\mathbf{u}$ is an **eigenvector** of $\mathbf{A}$, if exists $\lambda \in \mathbb{R}$ such that $\mathbf{A}\mathbf{u} = \lambda \mathbf{u}$
- such a λ is called an **eigenvalue**
- if $\mathbf{u}$ is eigenvector with eigenvalue λ, so is any $\alpha \mathbf{u}$ with $\alpha \in \mathbb{R}$
- $\mathbf{A}$ is called **positive semi-definite**, if

$$\mathbf{v}^\top \mathbf{A} \mathbf{v} \geq 0 \quad (\forall \mathbf{v})$$

- If $\mathbf{A} = \mathbf{B}^\top \mathbf{B}$ for some $\mathbf{B} \in \mathbb{R}^{n \times m}$, then $\mathbf{A}$ is p.s.d.

$$\mathbf{v}^\top \left(\mathbf{B}^\top \mathbf{B} \right) \mathbf{v} = (\mathbf{B}\mathbf{v})^\top (\mathbf{B}\mathbf{v}) = \|\mathbf{B}\mathbf{v}\|^2 \geq 0$$

Optimal Line: Conclusion #2

- Optimal direction = **principal eigenvector** of the sample variance-covariance matrix

- Extremal characterization

$$\mathbf{u} \leftarrow \arg\max_{\mathbf{v}:\|\mathbf{v}\|=1} \left[\mathbf{v}^\top \mathbf{\Sigma} \mathbf{v}\right]$$

Variance Maximization

- Re-interpret in term of variance maximization in 1d representation

$$\mathsf{Var}[z] = \frac{1}{n}\sum_{i=1}^{n} z_i^2 = \frac{1}{n}\sum_{i=1}^{n}\langle \mathbf{x}_i, \mathbf{u}\rangle^2 = \mathbf{u}^\top \mathbf{\Sigma} \mathbf{u}$$

  - remember: we subtracted the mean
  - same objective as before

- Direction of **smallest reconstruction error** $\Longleftrightarrow$

  Direction of **largest data variance**

Section 3

Principal Component Analysis

Residual Problem

- Residual: projection to $(\mathbb{R}\mathbf{u})^{\perp}$

$$\mathbf{r}_i := \mathbf{x}_i - \tilde{\mathbf{x}}_i = \left(\mathbf{I} - \mathbf{u}\mathbf{u}^{\top}\right)\mathbf{x}_i$$

- Variance-covariance matrix of residual vectors

$$\begin{aligned}
\frac{1}{n}\sum_{i=1}^{n}\mathbf{r}_i\mathbf{r}_i^{\top} &= \frac{1}{n}\sum_{i=1}^{n}\left(\mathbf{I} - \mathbf{u}\mathbf{u}^{\top}\right)\mathbf{x}_i\mathbf{x}_i^{\top}\left(\mathbf{I} - \mathbf{u}\mathbf{u}^{\top}\right)^{\top} \\
&= \left(\mathbf{I} - \mathbf{u}\mathbf{u}^{\top}\right)\boldsymbol{\Sigma}\left(\mathbf{I} - \mathbf{u}\mathbf{u}^{\top}\right)^{\top} \\
&= \boldsymbol{\Sigma} - 2\underbrace{\boldsymbol{\Sigma}\mathbf{u}}_{=\lambda\mathbf{u}}\mathbf{u}^{\top} + \mathbf{u}\underbrace{\mathbf{u}^{\top}\boldsymbol{\Sigma}\mathbf{u}}_{=\lambda}\mathbf{u}^{\top} = \boldsymbol{\Sigma} - \lambda\mathbf{u}\mathbf{u}^{\top}
\end{aligned}$$

Iterative View

- What does this mean? Note that

$$\left(\boldsymbol{\Sigma} - \lambda \mathbf{u}\mathbf{u}^\top\right) \mathbf{u} = \lambda \mathbf{u} - \lambda \mathbf{u} = 0$$

  - so $\mathbf{u}$ is now an eigenvector with eigenvalue 0

- Because $\boldsymbol{\Sigma}$ is p.s.d., all eigenvalues are non-negative

- Repeating the above procedure:
  - we find the principal eigenvector of $\left(\boldsymbol{\Sigma} - \lambda \mathbf{u}\mathbf{u}^\top\right)$
  - which is the 2nd principal eigenvector of $\boldsymbol{\Sigma}$
  - we keep iterating to identify the d principal eigenvectors of $\boldsymbol{\Sigma}$
  - eigenvectors are guaranteed to be pairwise orthogonal

Diagonalization

- Let us take a matrix view (to complement the iterative one ...)
- $\boldsymbol{\Sigma}$ can be **diagonalized** by **orthogonal matrices**

$$\boldsymbol{\Sigma} = \mathbf{U}\boldsymbol{\Lambda}\mathbf{U}^\top, \quad \boldsymbol{\Lambda} = \mathrm{diag}(\lambda_1, \dots, \lambda_m), \quad \lambda_1 \geq \cdots \geq \lambda_m$$

  where $\mathbf{U}$ is an orthogonal matrix (unit length, orthogonal columns)

$$\mathbf{U} = \begin{pmatrix} \mathbf{u}_1 & \mathbf{u}_2 & \dots & \mathbf{u}_m \end{pmatrix},$$
$$\mathbf{U}^\top \mathbf{u}_i = \mathbf{e}_i, \quad \boldsymbol{\Sigma}\mathbf{u}_i = \lambda_i \mathbf{u}_i$$

  i.e. the columns are eigenvectors (form an eigenvector basis).

Results from Linear Algebra

- $\mathbf{\Sigma}$ is symmetric, $\mathbf{\Sigma} = \mathbf{\Sigma}^\top$
  - obvious as $\sigma_{jk} = \frac{1}{n}\sum_i x_{ij}x_{ik}$
- **Spectral Theorem**: Matrix $\mathbf{A}$ is diagonalizable by an orthogonal matrix if and only if it is symmetric
  - $\mathbf{U}$ orthogonal: $\mathbf{U}^\top\mathbf{U} = \mathbf{U}\mathbf{U}^\top = \mathbf{I}$ (i.e. transpose = inverse)
    - columns are normalized and orthogonal: $\langle \mathbf{u}_j, \mathbf{u}_k \rangle = \delta_{jk}$
- **Theorem**: Distinct eigenvalues of symmetric matrices have orthogonal eigenvectors

$$\mathbf{u}_1^\top \mathbf{A}\mathbf{u}_2 = \langle \mathbf{u}_1, \lambda_2 \mathbf{u}_2 \rangle \overset{\text{symm}}{=} \mathbf{u}_2^\top \mathbf{A}\mathbf{u}_1 = \langle \mathbf{u}_2, \lambda_1 \mathbf{u}_1 \rangle$$

$$\implies (\lambda_1 - \lambda_2)\langle \mathbf{u}_1, \mathbf{u}_2 \rangle = 0 \overset{\lambda_1 \neq \lambda_2}{\implies} \langle \mathbf{u}_1, \mathbf{u}_2 \rangle = 0$$

PCA: Final Answer

- What is the optimal **reduction** to d dimensions?
  - diagonalize $\boldsymbol{\Sigma}$ and pick the d principal eigenvectors

$$\tilde{\mathbf{U}} = \begin{pmatrix} \mathbf{u}_1 & \dots & \mathbf{u}_d \end{pmatrix}, \quad d \leq m$$

  - dimension reduction

$$\mathbf{Z} = \underbrace{\tilde{\mathbf{U}}^\top}_{\in \mathbb{R}^{d \times m}} \underbrace{\mathbf{X}}_{\in \mathbb{R}^{m \times n}} \in \mathbb{R}^{d \times n}$$

- What is the optimal **reconstruction** in d dimensions?
  - use eigenbasis

$$\tilde{\mathbf{X}} = \tilde{\mathbf{U}}\mathbf{Z} = \underbrace{\tilde{\mathbf{U}}\tilde{\mathbf{U}}^\top}_{\text{projection}} \mathbf{X}$$

Section 4

Algorithms and Practicalities

Power Method

- Simple algorithm for finding dominant eigenvector of $\mathbf{A}$
- **Power iteration**

$$\mathbf{v}_{t+1} = \frac{\mathbf{A}\mathbf{v}_t}{\|\mathbf{A}\mathbf{v}_t\|}$$

  - assumptions: $\langle \mathbf{u}_1, \mathbf{v}_0 \rangle \neq 0$ and $|\lambda_1| > |\lambda_j|$ $(\forall j \geq 2)$
- Then it follows:

$$\lim_{t \to \infty} \mathbf{v}_t = \mathbf{u}_1$$

  - recover λ_1 from Rayleigh quotient $\lambda_1 = \lim_{t \to \infty} \|\mathbf{A}\mathbf{v}_t\| / \|\mathbf{v}_t\|$

Power Method: Proof Sketch

- Focus on $\mathbf{\Sigma}$ (p.s.d. and symmetric): eigenbasis $\{\mathbf{u}_1, \ldots, \mathbf{u}_m\}$

$$\mathbf{v}_0 = \sum_{j=1}^{m} \alpha_j \mathbf{u}_j$$

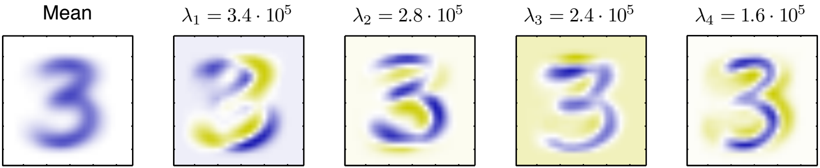
- Evolution equation

$$\mathbf{v}_t = c_t \sum_{j=1}^{m} \alpha_j \lambda_j^t \mathbf{u}_j = c_t \lambda_1^t \sum_{j=1}^{m} \alpha_j \left(\frac{\lambda_j}{\lambda_1} \right)^t \mathbf{u}_j \stackrel{t \to \infty}{\longrightarrow} \mathbf{u}_1$$

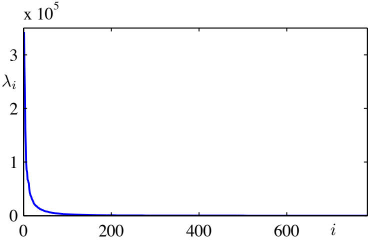
  - as $\lambda_j / \lambda_1 < 1$ and thus $c_t \to 1/\alpha_1$ (as $\|\mathbf{u}_1\| = 1$)

Digits Example

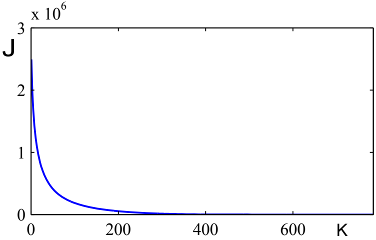
- Mean vector and first four principal directions:



- Eigenvalue spectrum (left), and approximation error (right):

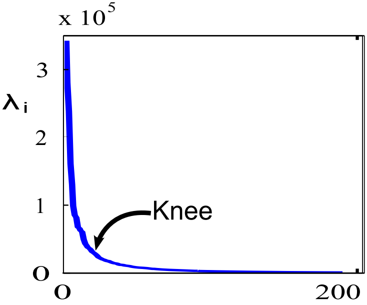


(a)



(b)

Model Selection in PCA



- Eigenvalue spectrum: information source to determine intrinsic dimensionality
- Heuristic: detect "knee" in eigenspectrum (= dimension)