Principal Component Analysis

Rudra Mukhopadhyay

January 21, 2022

“The Curse of Dimensionality”

- Features: *source* of information.
- More features $\implies$ more accurate predictions?
- As the number of variables rises, $\dim(V)$ increases.
- Two complementary hurdles stem from-
    1. $\#$ samples constant.
    2. Density of the samples constant.

“The Curse of Dimensionality”

1. Hughes phenomenon or Peaking phenomenon.
   - Initially accuracy increases with increasing number of features.
   - $L_2^n := \sum_{i=1}^{n} \sqrt{x_i^2} \geq \sum_{i=1}^{n-1} \sqrt{x_i^2} = L_2^{n-1}$
   - The data points become *sparse* and a pattern could hardly be detected in higher dimensions.

Abstract—The overall mean recognition probability (mean accuracy) of a pattern classifier is calculated and numerically plotted as a function of the pattern measurement complexity n and design data set size m. Utilized is the well-known probabilistic model of a two-class, discrete-measurement pattern environment (no Gaussian or statistical independence assumptions are made). The minimum-error recognition rule (Bayes) is used, with the unknown pattern environment probabilities estimated from the data relative frequencies. In calculating the mean accuracy over all such environments, only three parameters remain in the final equation: n, m, and the prior probability p_c of either of the pattern classes.

With a fixed design pattern sample, recognition accuracy can first increase as the number of measurements made on a pattern increases, but decay with measurement complexity higher than some optimum value. Graphs of the mean accuracy exhibit both an optimal and a maximum acceptable value of n for fixed m and p_c. A four-place tabulation of the optimum n and maximum mean accuracy values is given for equally likely classes and m ranging from 2 to 1000.

The penalty exacted for the generality of the analysis is the use of the mean accuracy itself as a recognizer optimality criterion. Namely, one necessarily always has some particular recognition problem at hand whose Bayes accuracy will be higher or lower than the mean over all recognition problems having fixed n, m, and p_c.

Hughes, 1968

"The Curse of Dimensionality"

2. Computational cost and over-fitting
   - m data points/ dimension $\implies m^{\dim(V)}$
   - Computational cost would rise exponentially.
   - Huge training set $\implies$ overly *intricate* patterns.

“The Curse of Dimensionality”

The way out: reducing # dimensions

- Feature selection:
  - $f_S : \mathbb{R}^n \mapsto \mathbb{R}^m$
  - $v = \{v_1, \ldots, v_n\} \in \mathbb{R}^n$
  - $f_S(v) = \{v_{i1}, \ldots, v_{im} | v_{ij} \in v, \forall j\}$
- Feature extraction:
  - $f_E : \mathbb{R}^n \mapsto \mathbb{R}^m$
  - An optimal mapping must not increase $P[\epsilon]$.
  - No *systematic* way to find an optimum nonlinear f_E.

Dimensionality reduction

Representation

Possible most accurate representation.

PCA $\rightarrow$ ensuring maximum variance

Classification

Enhancing the "class discriminatory information".

LDA $\rightarrow$ ensuring maximum separation

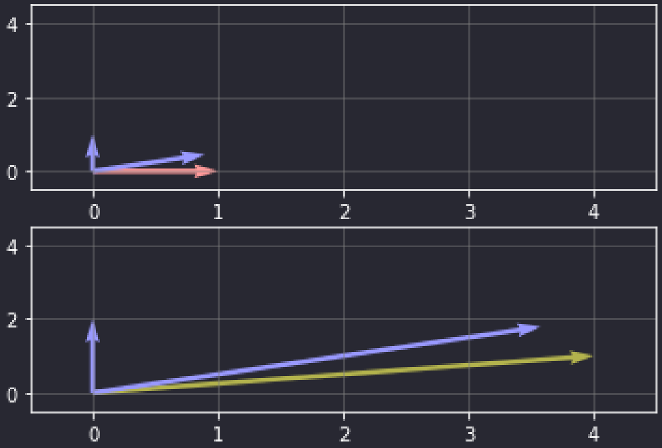
The PCA Algo

Algorithm:

1. Given a $p \times n$ data-set: N samples with n features.
2. Center the data-set.
3. Calculate the co-variance matrix Σ_n.
4. Calculate the eigenvalues (λ_i) and the corresponding eigenvectors (v_i) of Σ.
5. Arrange (λ_i, v_i) following the order of $\{\lambda_i | \lambda_{i+1} < \lambda_i\}$.
6. First m eigenvectors correspond to the first m PA.
7. "Rotate" the data considering $\{v_{i1}, \ldots, v_{im}\}$ as the basis.

Detour: Eigen-world

$$Av_i = \lambda_i v_i$$



Detour: Eigenbasis

Say, A is symmetric
Let $\{v_1, \ldots, v_k\}$ correspond to $\{\lambda_1, \ldots, \lambda_k\}$

Claim: $\{v_1, \ldots, v_k\}$ is LI
Suppose not!
$v_j = \sum_{i=1}^{j-1} \alpha_i v_i$
$\implies \lambda_j v_j = \sum_{i=1}^{j-1} \alpha_i \lambda_i v_i$
$\implies \sum_{i=1}^{j-1} \alpha_i (\lambda_j - \lambda_i) v_i = 0$
λ_i, λ_j are distinct by construction.

Detour: Eigenbasis

Claim: $v_i \perp v_j, \forall i, j$

$Av_i = \lambda_i v_i$

$\implies v_i^t A^t = \lambda_i v_i^t$

$\implies v_i^t A v_j = \lambda_i v_i^t v_j$

$\implies v_i^t \lambda_j v_j = \lambda_i v_i^t v_j$

λ_i, λ_j are distinct by construction.

$\therefore v_i^t v_j = 0$

Detour: Eigenbasis

Real spectral theorem:

$F = \mathbb{R}$, $T \in \mathcal{L}(V)$. Then-
$T = T^* \Leftrightarrow V$ has an orthonormal basis consisting of eigenvectors of T.

Σ *is symmetric, hence Hermitian.*

Detour: The covariance matrix

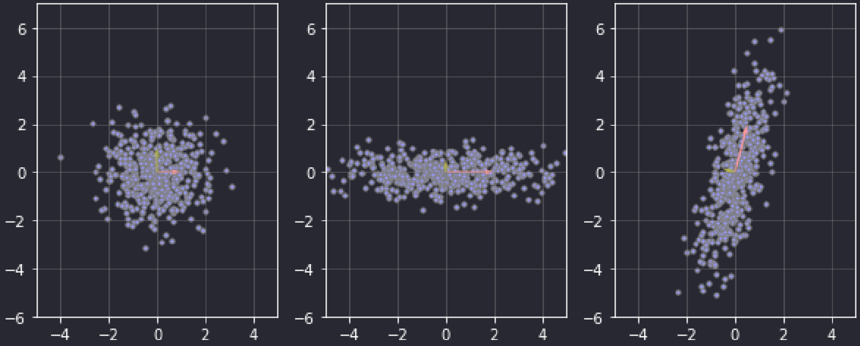
How to *understand* the eigen-decomposition of Σ [1]?

- Take a 2D normally distributed data set, with $\Sigma = I_2$.
- Intuitively, a *tilted, correlated* data-set can be *created* out of it...

  - by scaling and rotation!

[1] Inspired by *Understanding the Covariance matrix* by N. Janakiev, 2018

Detour: The covariance matrix



White data (left), scaled white (middle), correlated data (right)

Detour: The covariance matrix

Is it possible to decompose Σ in terms of L and R?

- Σ is symmetric $\implies k = \dim(\mathbb{R}^n)$ where k is # of distinct eigenvectors.
- $\Sigma v_i = \lambda_i v_i, i = 1, \ldots, n$
- Therefore, $\Sigma V = VL$ where V contains the eigenvectors as columns, $L = \text{diag}[\lambda_1, \ldots, \lambda_n]$
- $\Sigma = VLV^{-1}$.

Detour: The covariance matrix

- V is orthonormal, hence can be treated as a rotation matrix.
- $\Sigma = VLV^{-1} = RSSR^{-1}$ where $R = V, S^2 = L$
- $\Sigma = RSSR^{-1} = CC^t$ (Cholesky decomposition)
  - $C^t = (RS)^t = S^t R^t = SR^{-1}$
  - $x := \mu + Cu, u \sim \mathcal{N}(0, I)$
  - $E[x] = \mu, E[(x - \mu)(x - \mu)^t] = \Sigma$

Detour: The covariance matrix

$$S = \begin{pmatrix} 2 & 0 \\ 0 & 0.5 \end{pmatrix}, \; S = \begin{pmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{pmatrix} \text{ where } \theta = 5\pi/12$$

$$\Sigma = \begin{pmatrix} 0.5595 & 1.0793 \\ 1.0793 & 4.1357 \end{pmatrix}$$

$$RSSR^{-1} = \begin{pmatrix} 0.5012 & 0.9375 \\ 0.9375 & 3.7488 \end{pmatrix}$$

Formalisation of *Why* PCA Works

Representation in a low dimensional space

- $x \in \mathbb{R}^n$ and $\beta := \{\beta_i | i = 1, \dots, n\}$ be an orthonormal basis vector.
  - Every inner product space will have an orthonormal basis, constructed via Gram-Schmidt process.
- $x = \sum_{i=1}^{n} x_i \beta_i$
- x can be represented in $\mathbb{R}^m$ by $\hat{x} = \sum_{i=1}^{m} x_i \beta_i + \sum_{i=m+1}^{n} a_i \beta_i$.
- Thus, $\Delta x = x - \hat{x} = \sum_{i=m+1}^{n} (x_i - a_i) \beta_i$

Formalisation of *Why* PCA Works

Estimation of MSE

- $\epsilon := E[\Delta x^2]$
  $= E[\Delta x^t \Delta x]$
  $= E[\sum_{i=m+1}^{n} \sum_{i=m+1}^{n} (x_i - a_i)(x_j - a_j)\beta_i^t \beta_j]$
  $= \sum_{i=m+1}^{n} E[(x_i - a_i)^2]$
- $\frac{\partial \epsilon}{\partial a_i} = 0$
  $\implies a_i = E[x_i]$

Formalisation of *Why* PCA Works

Estimation of MSE

- $\epsilon = \sum_{i=m+1}^{n} E[(x_i - E[x_i])^2]$
  To be noted: $x_i = \beta_i^t x$
  $\epsilon = \sum_{i=m+1}^{n} E[(\beta_i^t(x - E[x]))^2]$
  $= \sum_{i=m+1}^{n} E[(\beta_i^t(x - E[x])(\beta_i^t(x - E[x])^t]$
  $= \sum_{i=m+1}^{n} \beta_i^t(x - E[x])(x - E[x])^t \beta_i$
  $= \sum_{i=m+1}^{n} \beta_i^t \Sigma \beta_i$

Detour: The Lagrange Multiplier

- $f : D \mapsto \mathbb{R}$ is closed and bounded. Then $P_1, P_2 \in D$ such that $\max_{\mathbb{R}} f = f(P_1), \min_{\mathbb{R}} f = f(P_2)$.
- Constrained optimization: Optimize $f : \mathbb{R}^n \mapsto \mathbb{R}$ given a *constraint* $g : \mathbb{R}^n \mapsto \{0\}$.
    - $\mathcal{L} := f + \lambda g$ (λ being the Lagrange multiplier)
    - Find $x \in \mathbb{R}^n$ s.t. $\mathcal{L}'|_x = 0$. If P is an extremum, $P \in \{x\}$.

Detour: The Lagrange Multiplier

Why should this method work?

- P be an extremum of f.
- $\rho(t) = \langle x(t), y(t), z(t) \rangle$ be any parametric curve on the surface defined by $\mathcal{S} := (g(x) = 0)$, passing through $P = \rho(0)$.
- $h(t) := f(x(t), y(t), z(t))$
- h has an extremum at P.
- $h'|_0 = \langle \nabla f|_{\rho(0)}, \rho'(0) \rangle = 0$
- $\nabla f|_P \perp \rho(t)$, for any ρ.
- $\nabla g|_P \perp \mathcal{S} \implies \nabla f|_P \parallel \nabla g|_P$

Formalisation of *Why* PCA Works

Minimizing the MSE

- Constraint: $\langle \beta_i | \beta_i \rangle = 1$, as an orthonormal basis is considered.
- $\mathcal{L} = \sum_{i=m+1}^{n} \beta_i^t \Sigma \beta_i + \sum_{i=m+1}^{n} \lambda_i (\beta_i^t \beta_i - 1)$
- $\nabla \mathcal{L} = 0$
  - $\dfrac{\partial \sum_{i=1}^{n} \beta_i^t \Sigma \beta_i}{\partial \beta_i} = \dfrac{\partial \beta_i^t \Sigma \beta_i}{\partial \beta_i} = \dfrac{\partial \langle \beta_i | \Sigma \beta_i \rangle}{\beta_i} = \langle 1 | \Sigma \beta_i \rangle + \langle \beta_i | \Sigma \rangle = 2\beta_i^t \Sigma$
  - $\dfrac{\partial \sum_{i=1}^{n} (\beta_i^t \beta_i - 1)}{\partial \beta_i} = 2\beta_i^t$
  - $\therefore 2\beta_i^t \Sigma = 2\lambda_i \beta_i^t$
    $\implies (\beta_i^t \Sigma)^t = (\lambda \beta_i^t)^t$
    $\implies \Sigma \beta_i^t = \lambda_i \beta_i$ - *the necessary condition*

Formalisation of *Why* PCA Works

Minimizing the MSE

- The necessary condition $\xrightarrow{?}$ sufficient condition
- Arrange $\{\lambda_i | i = 1, \ldots, n\} \mapsto \{\lambda_{ij} | j = 1, \ldots, n, \lambda_{i(j+1)} < \lambda_{ij}\}$
- Consider $S = \{\lambda_{i(m+1)}, \ldots, \lambda_{in}\}$
- S minimizes $\mathcal{L} = \sum_{i=m+1}^{n} (2\beta_i^t \lambda_i \beta_i - \lambda_i) = \sum_{i=m+1}^{n} \lambda_i$, considering any collection of $n - m$ tuples (λ_i, β_i).

Formalisation of *Why* PCA Works

An Alternative POV: Maximizing Var

- The implicit goal of PCA is to maximize *spread* of the data-set.
- Assume y is already mean-deducted ($y = y_0 - E[y_i]$).
- $\beta = \{\beta_i | i = 1, \ldots, n\}$ be an orthonormal basis of $\mathbb{R}^n$ (as before).
- The new data, after relevant projection, is given by $\langle y^t | \beta_i \rangle \, \beta_i$.
- The goal is to maximize $\sum_{i=1}^{n} \mathrm{Var}[y^t \beta_i]$.

Formalisation of *Why* PCA Works

An Alternative POV: Maximizing Var

- $\sum_{i=1}^{n} \mathrm{Var}[y^t\beta_i]$
  $= \sum_{i=1}^{n} E[(y^t\beta_i)^t(y^t\beta_i)]$
    - $E[y^t\beta_i] = E[y]^t\beta_i = 0_n$

  $= \sum_{i=1}^{n} E[\beta_i^t yy^t\beta_i]$
  $= \sum_{i=1}^{n} \beta_i^t E[yy^t]\beta_i = \sum_i \beta_i^t\Sigma\beta_i$
- Define an exact same $\mathcal{L}$ to find the exact same set of solutions.
    - To *maximize* the measure, consider $S = \{\lambda_{i1}, \ldots, \lambda_{im}\}$.

Formalisation of *Why* PCA Works

What is the interpretation?

- Estimate Σ_n, calculate its normalized eigenvectors $\{v_i | i = 1, \ldots, n\}$.
- $\{v_i | i = 1, \ldots, n\} \mapsto \{v_{ij} | j = 1, \ldots, n\}$ such that $\{\lambda(v_{ij})\}$ is in decreasing order.
- Say, the intention is to reduce dimension to m, from n.

Formalisation of *Why* PCA Works

What is the interpretation?

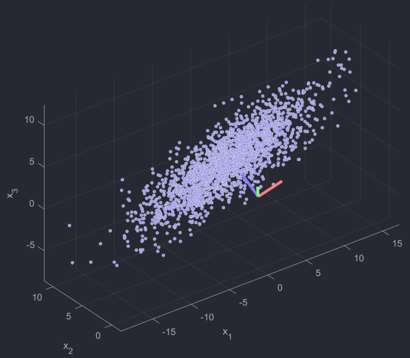
- Consider first m eigenvectors from the rearranged set.
  1. The rest will minimize the MSE after projecting the data onto the eigenbasis and substituting x_i with $E[x_i], \forall i = m+1, \ldots, n$.
  2. Those vectors will maximize the net variance of the projection of each data point y onto the eigenbasis.
- This reasoning justifies the PCA algorithm outlined initially.

Example-1

Data set generated using following parameters:

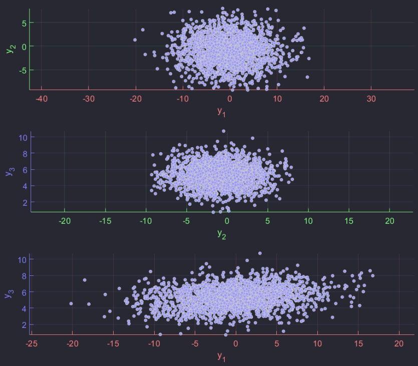
$$\mu = \begin{pmatrix} 0 \\ 5 \\ 2 \end{pmatrix}, \quad \Sigma = \begin{pmatrix} 25 & -1 & 7 \\ -1 & 4 & -4 \\ 7 & -4 & 10 \end{pmatrix}, \quad n = 2000$$

Example-1



Distribution of the data set in $\mathbb{R}^3$

Example-1



Projection onto $\mathbb{R}^2$, spanned by any two eigenbasis

Interpretation of a PCA plot

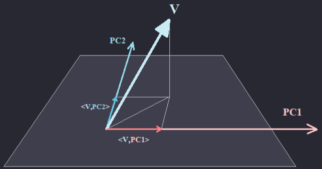
- The dimension is reduced from n to m.
- Any *connection* to the original features, however, is *lost*, in terms of its representation.
- How to bring the original features into the picture?

...Loading and correlation!

Interpretation of a PCA plot

Loading and correlation

- $v \in \mathbb{R}^n$ and PC_1, PC_2 be first two principal components.
- The goal is to represent v on the plane spanned by $\{PC_1, PC_2\}$.
- $\hat{v} = \langle v|PC_1\rangle\, PC_1 + \langle v|PC_2\rangle\, PC_2 = a_{v,1}PC_1 + a_{v,2}PC_2$
  - $a_{v,1}, a_{v,2}$ are the *loadings* for v.

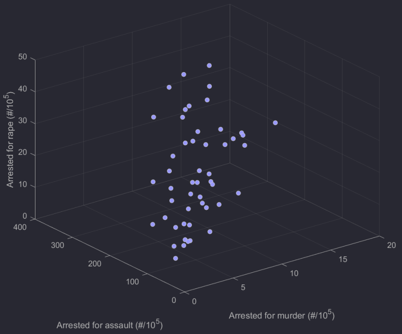


Interpretation of a PCA plot

Loading and correlation

- $u, v \in \mathbb{R}^n$, both centered.
- $\sigma_u^2 = \frac{1}{n-1}\sum_{i=1}^n u_i^2 = \frac{1}{n-1}||u||^2$
- $\sigma_{uv} = \frac{1}{n-1}\sum_{i=1}^n u_i v_i = \frac{1}{n-1}\langle u|v\rangle$
- $\rho_{uv} = \frac{\sigma_{uv}}{\sigma_u \sigma_v} = \frac{\langle u|v\rangle}{||u||||v||} = \cos\phi$
- More interestingly, $a_{v,i} = \langle v|PC_i\rangle = \rho_{v,PC_i}$ when v as well as PC_i is standardized.
  - v can indeed be standardized when the features are measured in different units.

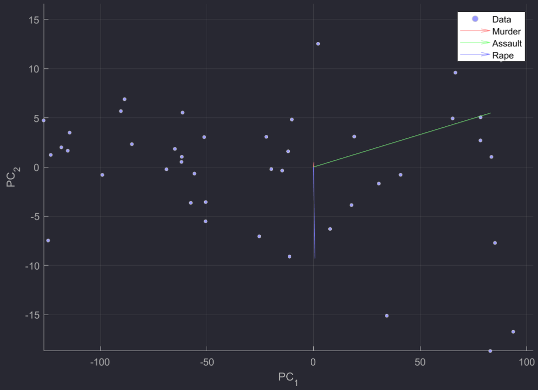
Example-2



Scatter representation of usArrest data-set [2], wrt 3 dimensions

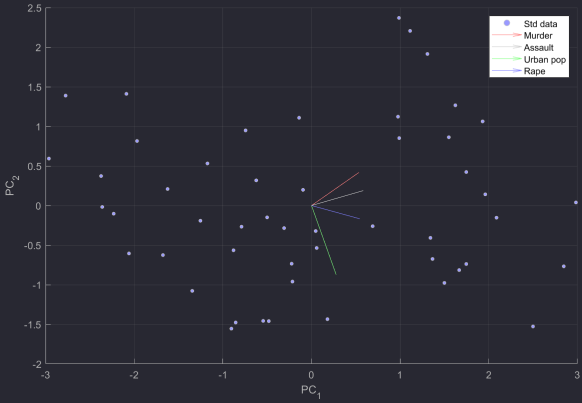
[2]https://www.picostat.com/dataset/usarrests

Example-2



Two PC representation: original features to be noted

Example-2



Two PC representation: original features to be noted

Proportion of Variance Explained

- $PVE_i := \frac{\text{Var}((V^t X)_i)}{\sum_{j=1}^{p} \text{Var}(X_j)}$
- For the last representation, $PVE = (0.62, 0.25, 0.09, 0.04)$