Other properties

- Every single variable has a univariate normal distribution.

$$X_j \sim N(\mu_j, \sigma_{jj}^2) \quad \text{if} \quad \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \sim N(\mu, \Sigma)$$

- Any subset of the variables also has a multivariate normal distribution.

$$\begin{bmatrix} X_1 \\ X_2 \\ X_3 \end{bmatrix} \sim N\left(\begin{bmatrix} \mu_1 \\ \mu_2 \\ \mu_2 \end{bmatrix} ; \begin{bmatrix} & & \end{bmatrix} \right.$$

- Zero covariance terms or a diagonal covariance matrix implies that the variables are independent of each other.

$$\Sigma_{3\times 3} = \begin{bmatrix} \sigma_{11} & 0 & 0 \\ 0 & \sigma_{22} & 0 \\ 0 & 0 & \sigma_{33} \end{bmatrix}$$

$$X_1 \perp\!\!\!\perp X_2 \qquad X_3 \perp\!\!\!\perp X_1$$

$$X_2 \perp\!\!\!\perp X_3$$

- Any conditional distribution for a subset of the variables conditional on known values for another subset of variables is a multivariate distribution.

Partitioned Gaussian Distributions

$$p(\mathbf{x}) = \mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma})$$

$- p \times 1 \quad - p \times p$

$x_a \cap x_b = \phi$

$$\mathbf{x} = \begin{pmatrix} \mathbf{x}_a \\ \mathbf{x}_b \end{pmatrix} \qquad \boldsymbol{\mu} = \begin{pmatrix} \boldsymbol{\mu}_a \\ \boldsymbol{\mu}_b \end{pmatrix} \qquad \boldsymbol{\Sigma} = \begin{pmatrix} \boldsymbol{\Sigma}_{aa} & \boldsymbol{\Sigma}_{ab} \\ \boldsymbol{\Sigma}_{ba} & \boldsymbol{\Sigma}_{bb} \end{pmatrix}$$

$$\boldsymbol{\Lambda} \equiv \boldsymbol{\Sigma}^{-1} \qquad \boldsymbol{\Lambda} = \begin{pmatrix} \boldsymbol{\Lambda}_{aa} & \boldsymbol{\Lambda}_{ab} \\ \boldsymbol{\Lambda}_{ba} & \boldsymbol{\Lambda}_{bb} \end{pmatrix}$$

$x_a \sim \mathcal{N}(\mu_a ; \Sigma_{aa})$

Partitioned Conditionals and Marginals

$a = \{1\}; \quad \{b = \{2\}$

$$p(\mathbf{x}_a|\mathbf{x}_b) = \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_{a|b}, \boldsymbol{\Sigma}_{a|b})$$

$P(x_1 | x_2) = \mathcal{N}(\mu_{1|2}; \Sigma_{1|2})$

$P\left(\begin{bmatrix} x_1 \\ x_2 \end{bmatrix}\right)$

Derive for bi-variate case.

$$\boxed{\boldsymbol{\Sigma}_{a|b} = \boldsymbol{\Lambda}_{aa}^{-1} = \boldsymbol{\Sigma}_{aa} - \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}\boldsymbol{\Sigma}_{ba}}$$

$$\begin{aligned}
\boldsymbol{\mu}_{a|b} &= \boldsymbol{\Sigma}_{a|b}\{\boldsymbol{\Lambda}_{aa}\boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b)\} \\
&= \boldsymbol{\mu}_a - \boldsymbol{\Lambda}_{aa}^{-1}\boldsymbol{\Lambda}_{ab}(\mathbf{x}_b - \boldsymbol{\mu}_b) \\
&= \boldsymbol{\mu}_a + \boldsymbol{\Sigma}_{ab}\boldsymbol{\Sigma}_{bb}^{-1}(\mathbf{x}_b - \boldsymbol{\mu}_b)
\end{aligned}$$

$\Sigma = \begin{bmatrix} \sigma_{11}^2 & \sigma_{12} \\ \sigma_{21} & \sigma_{22}^2 \end{bmatrix}$

$\mu = \begin{bmatrix} \mu_1 \\ \mu_2 \end{bmatrix}$

$$\begin{aligned}
p(\mathbf{x}_a) &= \int p(\mathbf{x}_a, \mathbf{x}_b)\, \mathrm{d}\mathbf{x}_b \\
&= \mathcal{N}(\mathbf{x}_a|\boldsymbol{\mu}_a, \boldsymbol{\Sigma}_{aa})
\end{aligned}$$

$\sigma_{1|2} = \sigma_{11}^2 - \dfrac{\sigma_{12}^2}{\sigma_{22}^2}$ ✓

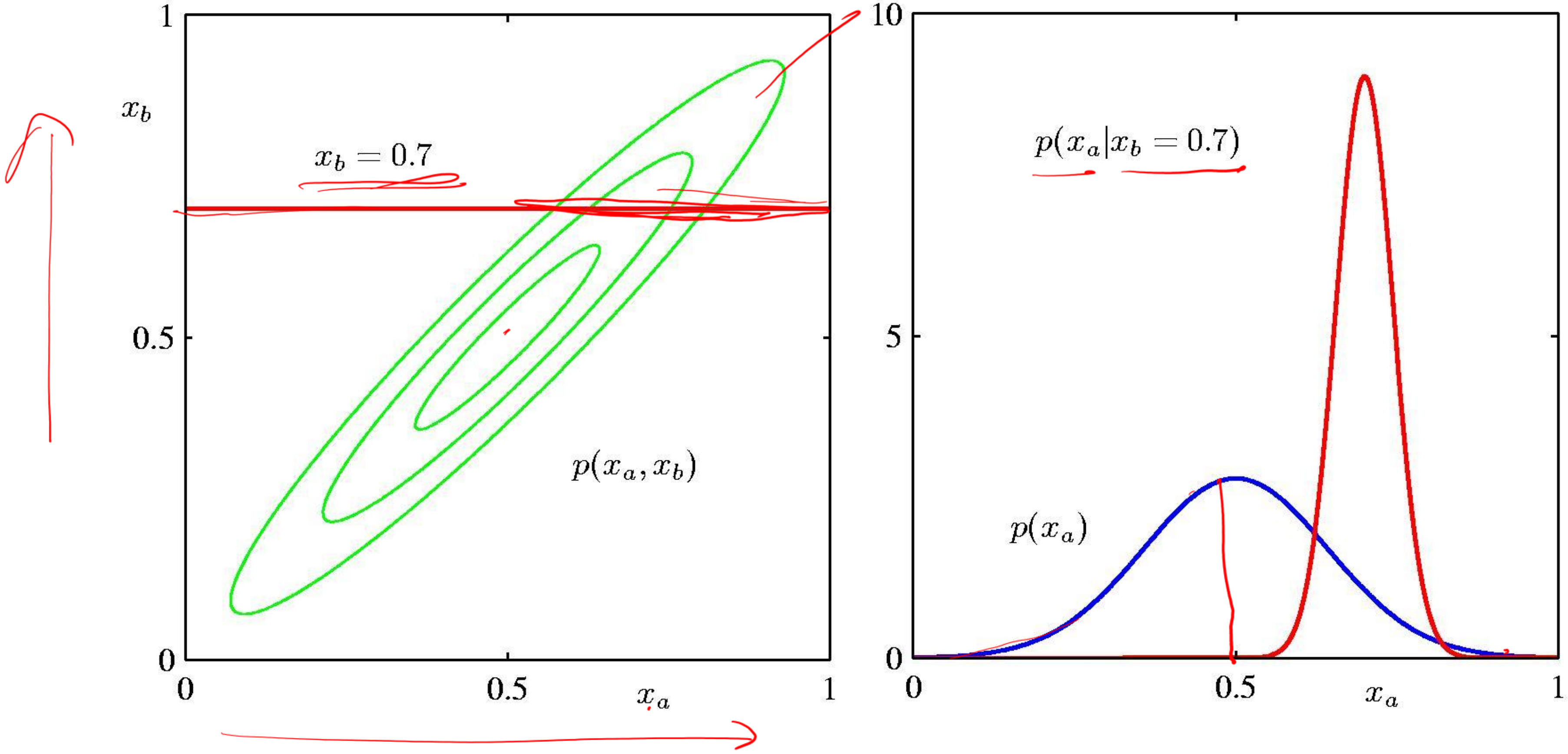
$\mu_{1|2} = \mu_1 + \dfrac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)$

Conditional distribution for bivariate case

$$\text{Mean} = \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2)$$

$$\text{Variance} = \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}}$$

Partitioned Conditionals and Marginals



Demos: https://colab.research.google.com/github/goodboychan/goodboychan.github.io/blob/main/_notebooks/2021-08-11-Multivariate-distribution.ipynb

Example 6-1: Conditional Distribution of Weight Given Height for College Men

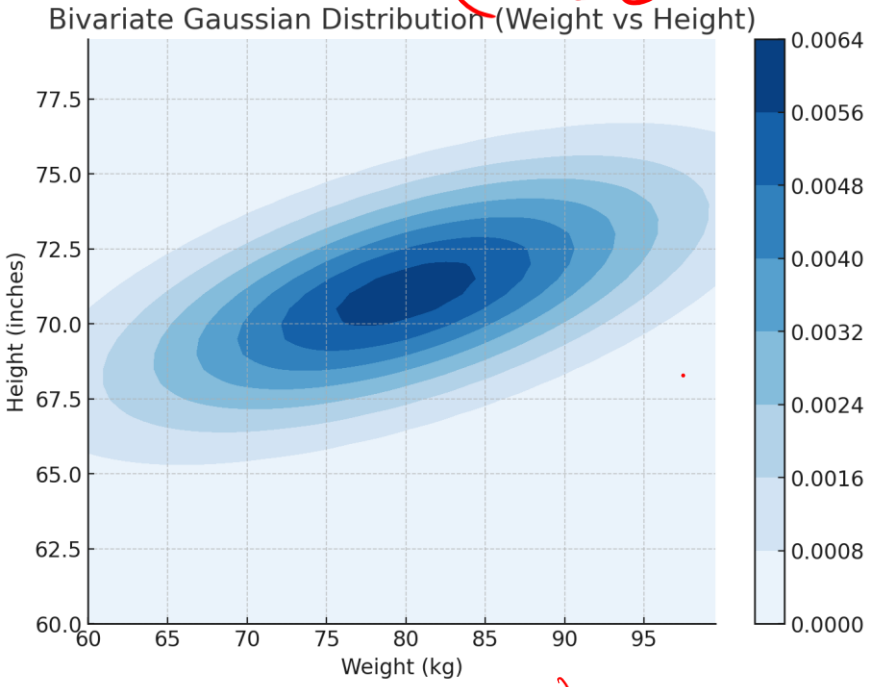
Suppose that the weights (lbs) and heights (inches) of undergraduate college men have a multivariate normal distribution with mean vector $\mu = \begin{pmatrix} 175 \\ 71 \end{pmatrix}$ and covariance matrix $\mathbf{\Sigma} = \begin{pmatrix} 550 & 40 \\ 40 & 8 \end{pmatrix}$.

$\mu_{kg} = \begin{pmatrix} 80 \\ 71 \end{pmatrix}$ $\Sigma_{kg} = \begin{pmatrix} 550/4 & 20 \\ 20 & 8 \end{pmatrix}$

The conditional distribution of X_1 weight given x_2 = height is a normal distribution with

$$\begin{aligned} \text{Mean} &= \mu_1 + \frac{\sigma_{12}}{\sigma_{22}}(x_2 - \mu_2) \\ &= 175 + \frac{40}{8}(x_2 - 71) \\ &= -180 + 5x_2 \end{aligned}$$

$$\begin{aligned} \text{Variance} &= \sigma_{11} - \frac{\sigma_{12}^2}{\sigma_{22}} \\ &= 550 - \frac{40^2}{8} \\ &= 350 \end{aligned}$$



For instance, for men with height = 70, weights are normally distributed with mean = -180 + 5(70) = 170 pounds and variance = 350. (So standard deviation $\sqrt{350} = 18.71$ = pounds)

Notice that we have generated a simple linear regression model that relates weight to height.

Geometry of the Multivariate Normal Distribution

- Can we characterize the shape and orientation of the ellipse that defines that contours of equal density?

$$\textit{Constant probability density contour} = \{\text{all } \mathbf{x} \text{ such that } (\mathbf{x}-\boldsymbol{\mu})'\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu}) = c^2\}$$

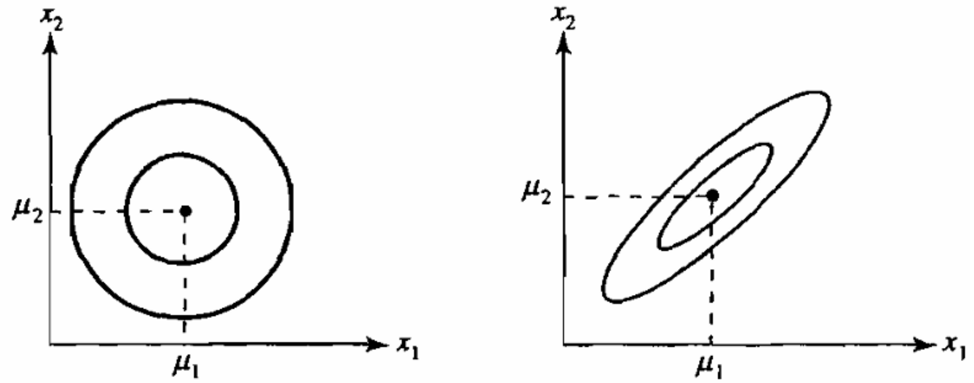


Figure 4.4 The 50% and 90% contours for the bivariate normal distributions in Figure 4.2.

- We will see that these can be characterized using eigen vectors and values of the covariance matrix.

Eigen values and Eigen vectors

- A square matrix A has a eigen value, eigen vector pair $\lambda, e \neq 0$ if $Ae = \lambda e$ where norm of e is 1

Let $\mathbf{A}$ be a $k \times k$ square symmetric matrix. Then $\mathbf{A}$ has k pairs of eigenvalues and eigenvectors namely,

$$\lambda_1, \mathbf{e}_1 \qquad \lambda_2, \mathbf{e}_2 \quad \dots \quad \lambda_k, \mathbf{e}_k \tag{2-15}$$

$e_1^T e_1 = e_1 \cdot e_1 = \langle e_1, e_1 \rangle = \sum_{j=1}^{p} e_{1j}^2$

The eigenvectors can be chosen to satisfy $1 = \mathbf{e}_1'\mathbf{e}_1 = \cdots = \mathbf{e}_k'\mathbf{e}_k$ and be mutually perpendicular. The eigenvectors are unique unless two or more eigenvalues are equal.

$e_j^T e_k = 0 \;\; \forall \; j \neq k$

Spectral decomposition of A

$$\underset{(k\times k)}{\mathbf{A}} = \lambda_1 \underset{(k\times 1)}{\mathbf{e}_1} \underset{(1\times k)}{\mathbf{e}_1'} + \lambda_2 \underset{(k\times 1)}{\mathbf{e}_2} \underset{(1\times k)}{\mathbf{e}_2'} + \cdots + \lambda_k \underset{(k\times 1)}{\mathbf{e}_k} \underset{(1\times k)}{\mathbf{e}_k'}$$

Spectral decomposition of a positive semi-definite matrix

- If A is positive-definite than all eigen-values >= 0

$y^T \underline{A} y \geq 0$

choose $y = e_j$ to show that $\lambda_j \geq 0$

- Example:

$$\mathbf{R} = \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}$$

- First find Eigen values and vectors.
  - [4.5 - Eigenvalues and Eigenvectors | STAT 505 (psu.edu)]

[HW]

$e_1 = \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix}$ for $\lambda_1 = 1 + \rho$ and $\begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} = e_2$ for $\lambda_2 = 1 - \rho$

$e_1 \cdot e_2 = 0$

$$\begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix} = (1+\rho) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ \frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & \frac{1}{\sqrt{2}} \end{pmatrix} + (1-\rho) \begin{pmatrix} \frac{1}{\sqrt{2}} \\ -\frac{1}{\sqrt{2}} \end{pmatrix} \begin{pmatrix} \frac{1}{\sqrt{2}} & -\frac{1}{\sqrt{2}} \end{pmatrix}$$

If $\Sigma = \lambda_1 e_1 e_1^T + \cdots \lambda_p e_p e_p^T$

then

$$\Sigma^{-1} = \frac{1}{\lambda_1} e_1 e_1^T + \cdots \frac{1}{\lambda_p} e_p e_p^T$$

Geometry of the Multivariate Gaussian

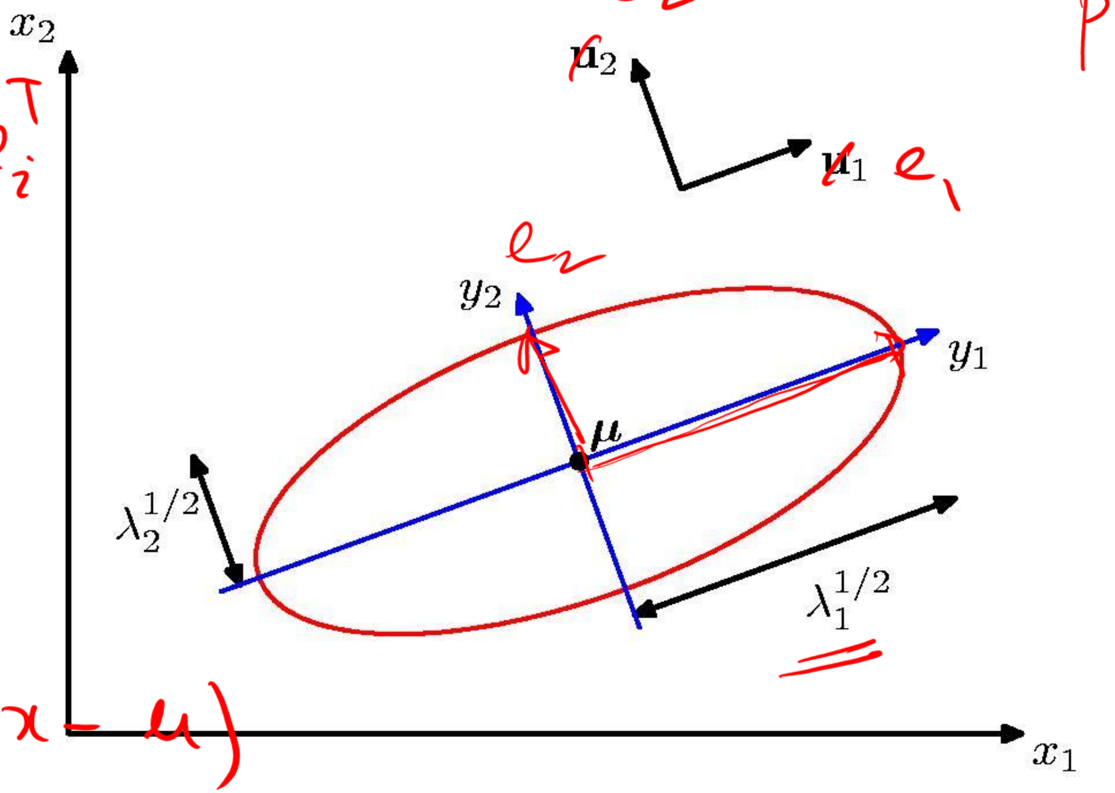
$c^2 = \Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$

$p = 2$

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{p} \cancel{\frac{1}{\lambda_i}\mathbf{u}_i}\mathbf{u}_i^{\mathrm{T}} \qquad \sum_{i=1}^{p} \frac{1}{\lambda_i} e_i e_i^T$$

$$c^2 = \Delta^2 = \sum_{i=1}^{p} \frac{y_i^2}{\lambda_i}$$

$$y_i = \mathbf{u}_i^{\mathrm{T}}\cancel{(\mathbf{x}-\boldsymbol{\mu})} \qquad e_i^T (x-\mu)$$



$$(x-\mu)^T \left[\sum_{i=1}^{p} \frac{1}{\lambda_i} e_i e_i^T \right] (x-\mu)$$

$$\Delta^2 = \sum_{i=1}^{p} \frac{1}{\lambda_i} \left(e_i^T (x-\mu) \right)^2$$

Principal component analysis

Projecting high-dimensional data

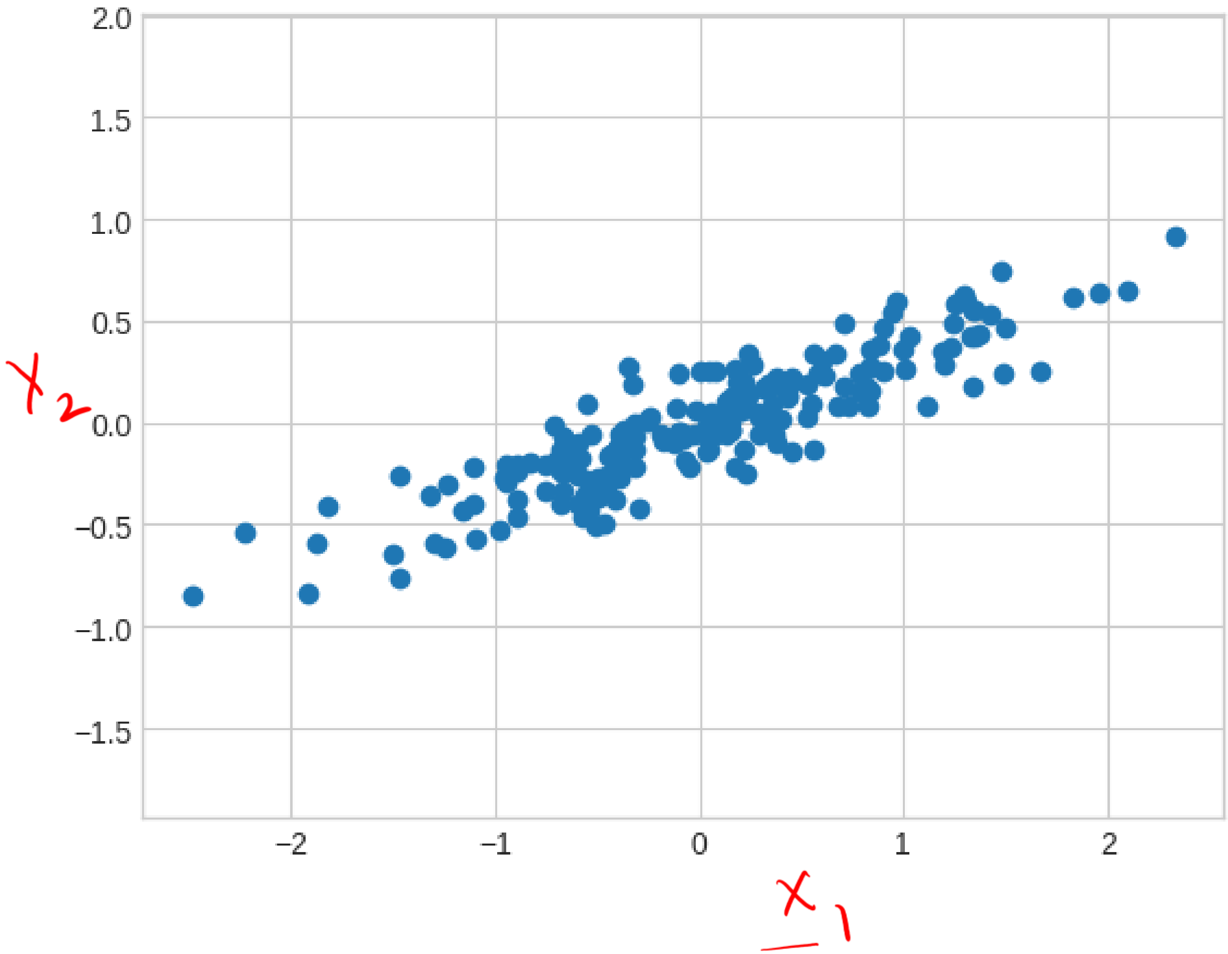
- When multivariate dataset has a large number of variables, analysis and interpretation of the data may be hard.

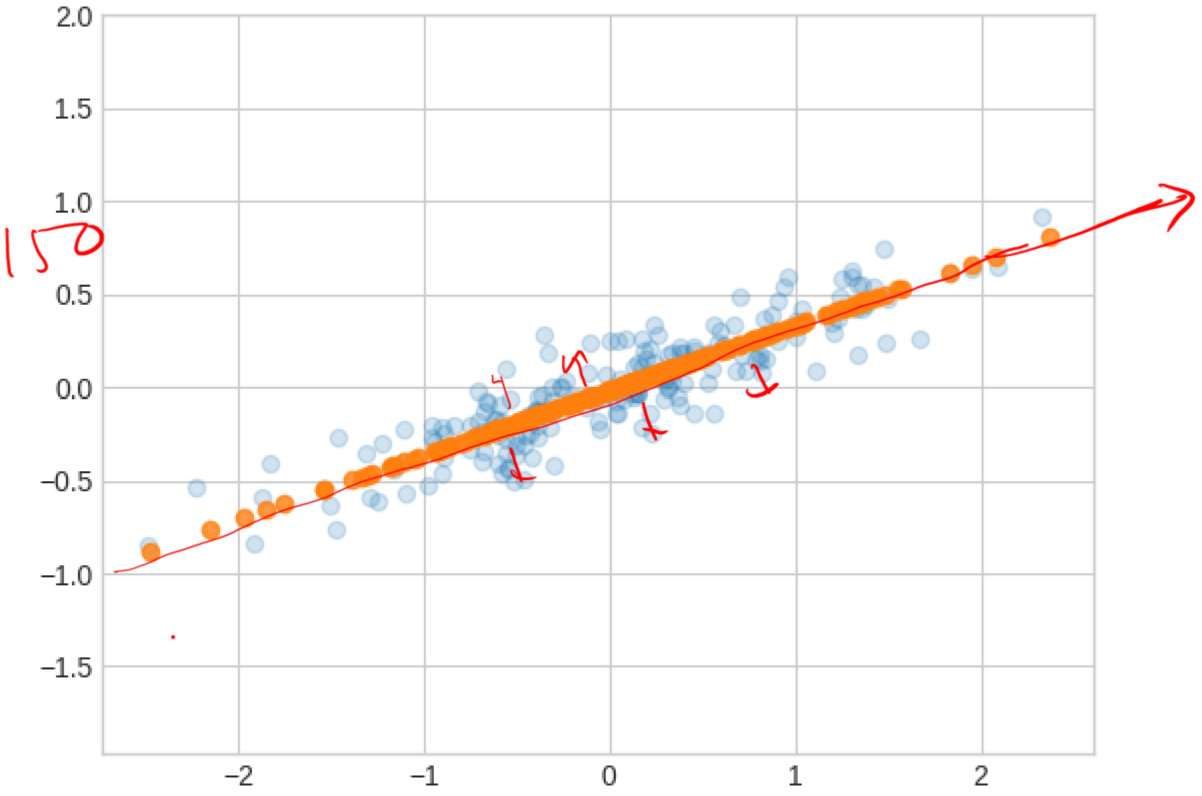
- Too many variables pairs, so pairwise correlation may be hard to grasp.

- For convenient visualization and interpretation
  - Reduce the number of variables.

- How to reduce number of variables while capturing most of the information in the data
  - Information == variance

Example





What is the best way to summarize this two dimensional data into a single dimension without losing much of the dispersion?

How to reduce number of variables: many methods

- Principal component analysis
- Factor analysis
- Other embedding methods
  - Random projection
  - T-SNE

Principal component analysis

- Let original set of p variables be $X_1, X_2, \dots, X_p$
- Define a smaller set of new variables that are linear combinations of existing variables.

$$
\begin{aligned}
Y_1 &= e_{11}X_1 + e_{12}X_2 + \cdots + e_{1p}X_p \\
Y_2 &= e_{21}X_1 + e_{22}X_2 + \cdots + e_{2p}X_p \\
&\vdots \\
Y_p &= e_{p1}X_1 + e_{p2}X_2 + \cdots + e_{pp}X_p
\end{aligned}
$$

Variance and Co-variance of the new variables.

Let

$$\operatorname{var}(\mathbf{X}) = \Sigma = \begin{pmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1p} \\ \sigma_{21} & \sigma_2^2 & \cdots & \sigma_{2p} \\ \vdots & \vdots & \ddots & \vdots \\ \sigma_{p1} & \sigma_{p2} & \cdots & \sigma_p^2 \end{pmatrix}$$

$$Y_i = \begin{bmatrix} e_{i1} \\ \vdots \\ e_{ip} \end{bmatrix}^T \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix}$$

Then:

$$\operatorname{var}(Y_i) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{il} \sigma_{kl} = \mathbf{e}_i' \Sigma \mathbf{e}_i$$

$$\operatorname{cov}(Y_i, Y_j) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{ik} e_{jl} \sigma_{kl} = \mathbf{e}_i' \Sigma \mathbf{e}_j$$

Principal components

- First principal component Y_1 is chosen to maximize the variance among all possible linear combinations such that the norm of coefficients is 1.

More formally, select $e_{11}, e_{12}, \ldots, e_{1p}$ that maximizes

$$\text{var}(Y_1) = \sum_{k=1}^{p} \sum_{l=1}^{p} e_{1k} e_{1l} \sigma_{kl} = \mathbf{e}_1' \Sigma \mathbf{e}_1$$

subject to the constraint that

$$\mathbf{e}_1' \mathbf{e}_1 = \sum_{j=1}^{p} e_{1j}^2 = 1$$

Second principal component

Select $e_{21}, e_{22}, \ldots, e_{2p}$ that maximizes the variance of this new component...

$$\mathrm{var}(Y_2) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{2k}e_{2l}\sigma_{kl} = \mathbf{e}_2'\Sigma\mathbf{e}_2$$

subject to the constraint that the sums of squared coefficients add up to one,

$$\mathbf{e}_2'\mathbf{e}_2 = \sum_{j=1}^{p} e_{2j}^2 = 1$$

along with the additional constraint that these two components are uncorrelated.

$$\mathrm{cov}(Y_1, Y_2) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{1k}e_{2l}\sigma_{kl} = \mathbf{e}_1'\Sigma\mathbf{e}_2 = 0$$

i^{th} Principal Component (PCAi): $\boldsymbol{Y_i}$

We select $\boldsymbol{e}_{i1},\boldsymbol{e}_{i2},\ldots,\boldsymbol{e}_{ip}$ to maximize

$$\text{var}(Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{ik}e_{il}\sigma_{kl} = \mathbf{e}_i'\Sigma\mathbf{e}_i$$

subject to the constraint that the sums of squared coefficients add up to one...along with the additional constraint that this new component is uncorrelated with all the previously defined components.

$$\mathbf{e}_i'\mathbf{e}_i = \sum_{j=1}^{p} e_{ij}^2 = 1$$

$$\text{cov}(Y_1, Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{1k}e_{il}\sigma_{kl} = \mathbf{e}_1'\Sigma\mathbf{e}_i = 0,$$

$$\text{cov}(Y_2, Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{2k}e_{il}\sigma_{kl} = \mathbf{e}_2'\Sigma\mathbf{e}_i = 0,$$

$$\vdots$$

$$\text{cov}(Y_{i-1}, Y_i) = \sum_{k=1}^{p}\sum_{l=1}^{p} e_{i-1,k}e_{il}\sigma_{kl} = \mathbf{e}_{i-1}'\Sigma\mathbf{e}_i = 0$$

For what Y_1 is Variance(Y_1) maximized?

- The coefficient of the first principal component correspond to the Eigen vector with the maximum Eigen value.

More generally

- The i-th principal component corresponds the i-th largest eigen vector.

The variance for the *i*th principal component is equal to the *i*th eigenvalue.

$$var(Y_i) = \mathrm{var}(e_{i1}X_1 + e_{i2}X_2 + \ldots e_{ip}X_p) = \lambda_i$$

$$\mathrm{cov}(Y_i, Y_j) = 0$$

The proportion of variance explained

- The total variance of X
- We can show that sum of p Eigen values equals the total variance

- The fraction of variance explained by the i-th Eigen value $$\frac{\lambda_i}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$$

Reducing number of dimensions

- Variance explained by first k Eigen values $\frac{\lambda_1 + \lambda_2 + \cdots + \lambda_k}{\lambda_1 + \lambda_2 + \cdots + \lambda_p}$

11.3 - Example: Places Rated

Example 11-2: Places Rated

We will use the Places Rated Almanac data (Boyer and Savageau) which rates 329 communities according to nine criteria:

1. Climate and Terrain
2. Housing
3. Health Care & Environment
4. Crime
5. Transportation
6. Education
7. The Arts
8. Recreation
9. Economics

[11.3 - Example: Places Rated | STAT 505 (psu.edu)]

Notes

- The data for many of the variables are strongly skewed to the right.
- The log transformation was used to normalize the data.

More demos

- https://colab.research.google.com/github/jakevdp/PythonDataScienceHandbook/blob/master/notebooks/05.09-Principal-Component-Analysis.ipynb