Machine Learning

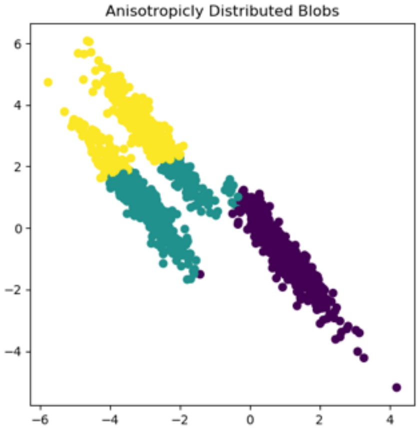
Session 3: Generative Models

1. Introduction to Generative models
2. The generative Gaussian Model: ML estimator
3. Review of linear algebra for the Gaussian Model
4. Properties of the Gaussian Model

Bibliography:

- Bishop: 1.2.4, 2.3 (almost up to 2.3.1, this excluded), Appendix C, Eigenvector equation
- See (up to Applications) the link below, with lot of code: https://towardsdatascience.com/understanding-singular-value-decomposition-and-its-application-in-data-science-388a54be95d

Generative Models for Clustering



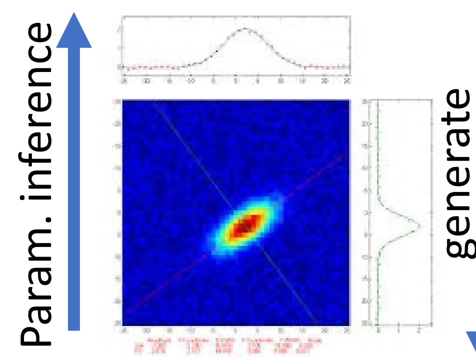
Problems with clustering by **K-means**:

- Elongated (non isotropic) clusters give us problems
- Different priors (different number of points in each cluster)
- Is not clear the meaning of clustering

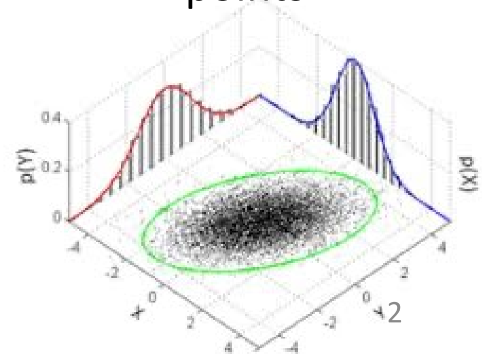
New idea: we imagine the data of each cluster is generated by a model.

- We then adjust the model parameters to maximize the probability that it produce exactly the data we observed (parameter inference).
- This approach give us a method for different priors, a measure of how good are the clusters
- From the model we can predict and generate new data
- We will do in two parts: Gaussian Model (today) and Mixture of Gaussian model

MODEL:Probability: 2D gaussian



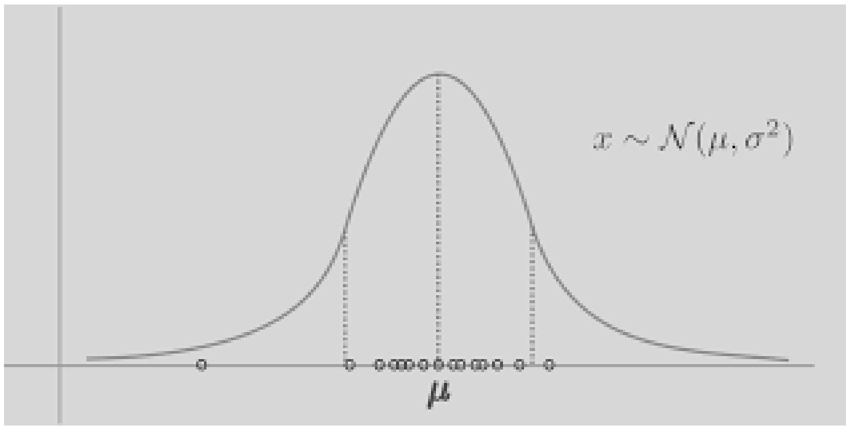
DATA: Sampling: a cluster of points

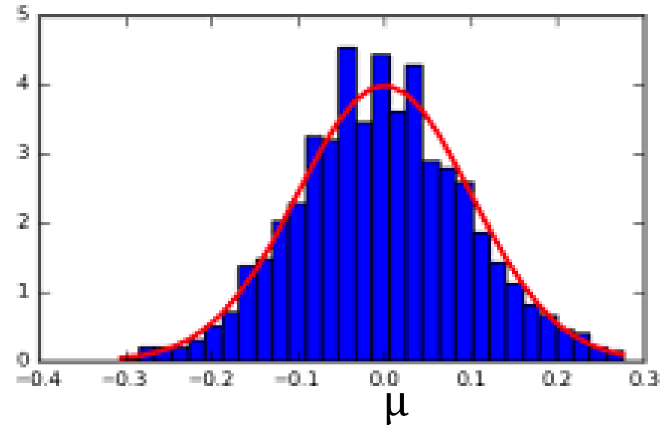


DATA come from a MODEL ->Interpret the model characteristics from the data

Generative Models: From data to the Gaussian Model

1D Gaussian Model: $$\mathcal{N}_1(x|\mu,\sigma) = \frac{1}{\sqrt{2\pi\sigma}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

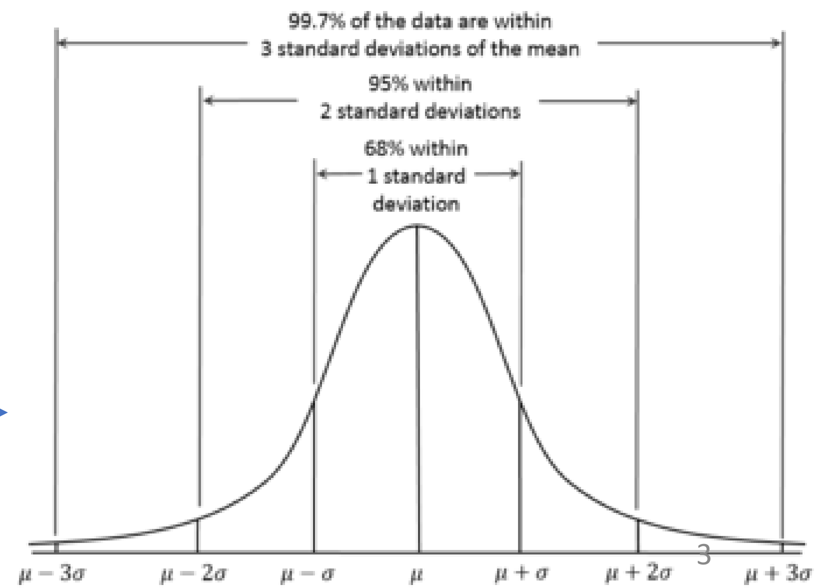




From the data $\mathcal{D} = \left\{x^{(1)}, \ldots, x^{(N)}\right\}$ we can compute the Maximum Likelihood of a 1D Gaussian model:

$$\begin{aligned}\mu_{ML} &= E(x) = \frac{1}{N}\sum_{n=1}^{N} x^{(n)} \\ \sigma^2_{ML} &= E\{(x - E(x))^2\} = \frac{1}{N}\sum_{n=1}^{N}(x^{(n)} - \mu_{ML})^2\end{aligned}$$

Statistic Deviation is also very useful



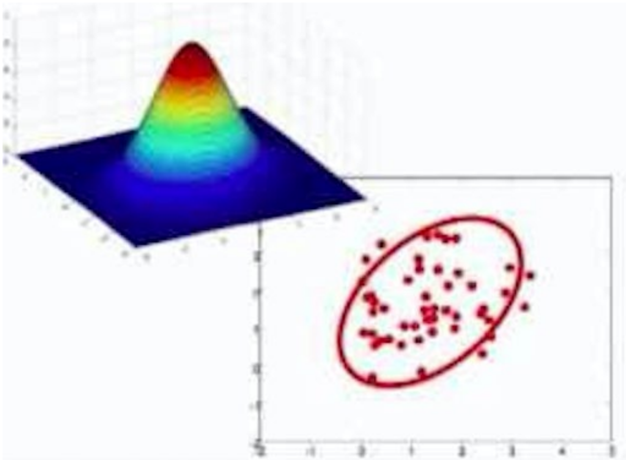
Generative Models: From data to the Gaussian Model

***D*-dimensional Gaussian Model:**

$$\mathcal{N}(\mathbf{x}|\mu, \mathbf{\Sigma}) = \frac{1}{(2\pi)^{D/2}|\Sigma|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)}$$

Determinant

$$\mathcal{N}_1(x|\mu, \sigma) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



From $\mathcal{D} = \{\mathbf{x}^{(1)}, \ldots, \mathbf{x}^{(N)}\}$, each $\mathbf{x}^{(N)} \in \mathbb{R}^D$ we can compute the Maximum Likelihood of a multi-variate Gaussian model:

$$\boldsymbol{\mu} = E\{\mathbf{x}\} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}^{(n)}$$

$$\mathbf{\Sigma} = E\left\{\left(\mathbf{x} - E(\mathbf{x})\right)^2\right\} = E\left\{\left(\mathbf{x} - E(\mathbf{x})\right)\left(\mathbf{x} - E(\mathbf{x})\right)^\top\right\}$$

$$= \frac{1}{N}\sum_{n=1}^{N}[\mathbf{Y}\mathbf{Y}^\top - \boldsymbol{\mu}\boldsymbol{\mu}^\top]$$

Where $\mathbf{Y}$ is a $D \times D$ matrix with $\mathbf{x}^{(n)}$ as columns

Sometimes (if N is small) we correct the bias of variance (covariance) using:

$$\frac{N}{N-1}\mathbf{\Sigma}$$

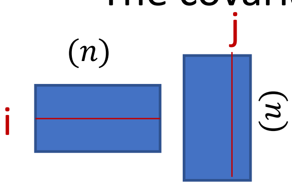
Generative Models: From data to the Gaussian Model

The covariance matrix for 2 or more dimensions:

- Remember than the vectors $\mathbf{x}^{(n)}$ $n = 1, \dots, N$ are column vectors:

$$\mathbf{x}^{(n)} = \begin{pmatrix} x_1^{(n)} \\ \vdots \\ x_i^{(n)} \\ \vdots \\ x_j^{(n)} \\ \vdots \\ x_D^{(n)} \end{pmatrix} \Rightarrow \quad \boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \vdots \\ \mu_i \\ \vdots \\ \mu_j \\ \vdots \\ \mu_D \end{pmatrix} \quad \boldsymbol{\Sigma} = \begin{pmatrix} \sigma_{11} & \sigma_{12} & \sigma_{13} \\ \sigma_{21} & \sigma_{22} & \sigma_{23} \\ \sigma_{31} & \sigma_{32} & \sigma_{33} \end{pmatrix}$$

- The mean of this set of vectors is $\boldsymbol{\mu} = (\mu_1, \dots, \mu_D)^\top$
- The covariance between the i-th and the j-th coordinates measures their (linear) relation:



$$\sigma_{ij} = \mathrm{cov}\left(x_i, x_j\right) = E\{(x_i - \mu_i)\left(x_j - \mu_j\right)\} = \frac{1}{N}\sum_{n=1}^{N}(x_i^{(n)} - \mu_i)\left(x_j^{(n)} - \mu_j\right) = E\{x_i x_j\} - \mu_i \mu_j^T$$

- Let $\mathbf{Y}$ be the matrix with the data points as column vectors (design matrix transposed)
  - We calculate $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu}$
  - $\boldsymbol{\Sigma} = \frac{1}{N}\mathbf{Z}\mathbf{Z}^\top$: to obtain Σ_{ij} multiply all the i-th values for all j-th values
  - $\boldsymbol{\Sigma}$ is a symmetric matrix: $\mathrm{cov}(i, j) = \mathrm{cov}(j, i)$

- We define the correlation corr(i,j)= $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$ with $\sigma_i = \sqrt{\sigma_{ii}}$ $with\ -1 \leq \rho_{ij} \leq 1$

Generative Models: From data to the Gaussian Model

Exercice: Assume N=4 examples in three dimensions

$$\mathbf{Y}=\begin{pmatrix} 2 & 3 & 5 & 6 \\ 2 & 4 & 4 & 6 \\ 4 & 6 & 2 & 4 \end{pmatrix}$$

The mean is: $\boldsymbol{\mu} = \begin{pmatrix} \mu_1 \\ \mu_2 \\ \mu_3 \end{pmatrix} = \begin{pmatrix} 4 \\ 4 \\ 4 \end{pmatrix}$ and $\mathbf{Z} = \mathbf{Y} - \boldsymbol{\mu} = \begin{pmatrix} -2 & -1 & 1 & 2 \\ -2 & 0 & 0 & 2 \\ 0 & 2 & -2 & 0 \end{pmatrix}$

$$\boldsymbol{\Sigma} = \frac{1}{4}\mathbf{Z}\mathbf{Z}^{\top} = \frac{1}{4}\begin{pmatrix} -2 & -1 & 1 & 2 \\ -2 & 0 & 0 & 2 \\ 0 & 2 & -2 & 0 \end{pmatrix}\begin{pmatrix} -2 & -1 & 1 & 2 \\ -2 & 0 & 0 & 2 \\ 0 & 2 & -2 & 0 \end{pmatrix}^{T} = \frac{1}{4}\begin{pmatrix} 10 & 8 & -4 \\ 8 & 8 & 0 \\ -4 & 0 & 8 \end{pmatrix} = \begin{pmatrix} 5/2 & 2 & -1 \\ 2 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

Correlations: $\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$ $\quad \sigma_i = \sqrt{\sigma_{ii}}$

ρ_{11} =1; Why?

$\rho_{12} = \frac{2}{\sqrt{5/2}\sqrt{2}}$=0.89 ;

$\rho_{13} = \frac{-1}{\sqrt{5/2}\sqrt{2}}$= - 0.45

$$\begin{pmatrix} 1 & 0.89 & -0.45 \\ 0.89 & 1 & 0 \\ -0.45 & 0 & 1 \end{pmatrix}$$

Generative Models: From data to the Gaussian Model

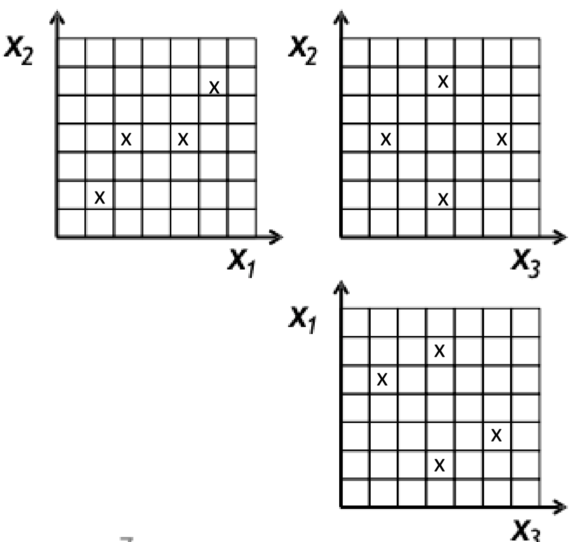
Interpretation of the covariance (and correlation):

Given the following samples from a 3D distribution

- Compute the covariance matrix (done!)
- Generate scatter plots for every pair of variables
- Can you observe any relationships between the covariance and the scatterplots?

Design Matrix: as in DataBases

	Variables (or features)		
Examples	x_1	x_2	x_3
1	2	2	**4**
2	3	4	**6**
3	5	4	**2**
4	**6**	**6**	**4**



$$\mathbf{\Sigma} = \begin{pmatrix} 5/2 & 2 & -1 \\ 2 & 2 & 0 \\ -1 & 0 & 2 \end{pmatrix}$$

(linear) correlations:

$$\rho_{12} = \frac{2}{\sqrt{5/2}\sqrt{2}} = 0.89$$

$$\rho_{32} = \frac{0}{\sqrt{2}\sqrt{2}} = 0$$

$$\rho_{31} = \frac{-1}{\sqrt{5/2}\sqrt{2}} = -0.45$$

Generative Models: From data to the Gaussian Model

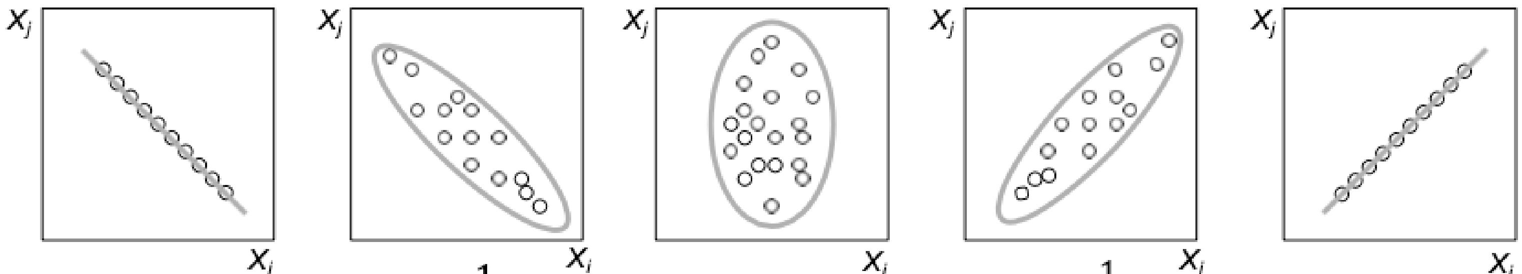
- The Parameters of the Gaussian Model are useful:
  - The covariance matrix indicates the tendency of each pair of features to **co-vary**
  - The covariance matrix has several important properties:

$$\rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$$

$$\sigma_i = \sqrt{\sigma_{ii}}$$

    - If x_i and x_j tend to increase together, then $\sigma_{ij} > 0$
    - If x_i tends to decrease when x_j increases, then $\sigma_{ij} < 0$
    - If x_i and x_j are uncorrelated, then $\sigma_{ij} = 0$
    - $\sigma_{ii} = \sigma_i^2 = \mathrm{Var}(x_i)$
    - The covariance terms can be expressed as: $\sigma_{ij} = \rho_{ij}\, \sigma_i \sigma_j$

- **Data Normalization**: make all features equally centered and on the same scale
  Share mean and variance: $x_i' = \frac{x_i - \mu_i}{\sqrt{\sigma_{ii}}}$



$\sigma_{ij} = -\sigma_i \sigma_j$, $\rho_{ij} = -1$

$\sigma_{ij} = -\frac{1}{2}\sigma_i \sigma_j$, $\rho_{ij} = -\frac{1}{2}$

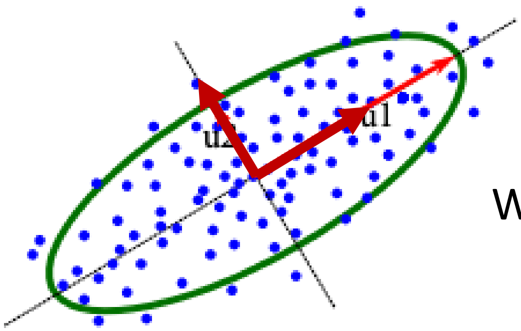
$\sigma_{ij} = 0$, $\rho_{ij} = 0$

$\sigma_{ij} = \frac{1}{2}\sigma_i \sigma_j$, $\rho_{ij} = \frac{1}{2}$

$\sigma_{ij} = \sigma_i \sigma_j$, $\rho_{ij} = 1$

Generative Models: From data to the Gaussian Model

- The parameters of the Gaussian Model (up to now):
  - Mean of the 'cluster'
  - Explain how the features co-variate (important information)
- This is not enough for clustering:
  - We need the **orientation** of the cluster (not only spherical!)
  - Given new points
    - we need to **predict** which cluster they belong to, i.e., probability density at that point under that cluster
  - The model gives us much more than the co-variation of features.



We will work to:

- Estimate the new vector basis from Covariance matrix
- Express the points in the new basis: for example, to say the points out of the ellipse, are rejected

Linear algebra for the Gaussian Model

Eigendecomposition of a covariance matrix

Review: Eigendecomposition: Standard eigenvalue problem

- Given a $n \times n$ matrix $\mathbf{A}$, find a scalar λ and a nonzero vector $\mathbf{v}$ s.t.:

$$\mathrm{A}\mathbf{v} = \lambda\mathbf{v}$$

  λ is the eigenvalue and $\mathbf{v}$ is corresponding eigenvector

- $\mathbf{Av} = \lambda\mathbf{v}$ is equivalent to solve $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0$ ($\mathbf{I}$ identity matrix)

- Nonzero solution of $\mathbf{v}$ if and only if the matrix $(\mathrm{A} - \lambda\mathbf{I})$ is singular

$$\det(\mathbf{A} - \lambda\mathbf{I}) = 0$$

- The eigenvalues are the zeros of the characteristic polynomial $\det(\mathbf{A} - \lambda\mathbf{I}) = 0$ in λ of degree n

- If the matrix $\mathbf{A}$ is symmetric, all the eigenvalues are Real

Linear algebra for the Gaussian Model

Eigendecomposition of a covariance matrix

Example: eigenvalues and eigenvectors of the matrix $\mathbf{A} = \begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix}$

- Characteristic polynomial:

$$\det\left[\begin{pmatrix} 3 & -1 \\ -1 & 3 \end{pmatrix} - \lambda \begin{pmatrix} 1 & 0 \\ 0 & 1 \end{pmatrix}\right] = \det\left[\begin{pmatrix} 3-\lambda & -1 \\ -1 & 3-\lambda \end{pmatrix}\right] = (3-\lambda)(3-\lambda) - 1 = \lambda^2 - 6\lambda + 8$$

Solve $\lambda^2 - 6\lambda + 8 = 0$

- **Eigenvalues** will be: $\lambda = \frac{6 \pm \sqrt{36-32}}{2} = \frac{6 \pm 2}{2}$ so $\boldsymbol{\lambda_1 = 4}$ and $\boldsymbol{\lambda_2 = 2}$
- **Eigenvectors** $\mathbf{x}$, $(\mathbf{A} - \lambda\mathbf{I})\mathbf{v} = 0$,

For $\boldsymbol{\lambda_1 = 4}$: $\begin{pmatrix} -1 & -1 \\ -1 & -1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0$, so $-v_1 - v_2 = 0$; $v_2 = -v_1$

$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ -v_1 \end{pmatrix}$; $\begin{pmatrix} 1 \\ -1 \end{pmatrix}$ making it of unit norm, $\mathbf{v}^{(1)} = \begin{pmatrix} 1/\sqrt{2} \\ -1/\sqrt{2} \end{pmatrix}$

For $\boldsymbol{\lambda_2 = 2}$: $\begin{pmatrix} 1 & -1 \\ -1 & 1 \end{pmatrix}\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = 0$, so $v_1 - v_2 = 0$; $v_2 = v_1$

$\begin{pmatrix} v_1 \\ v_2 \end{pmatrix} = \begin{pmatrix} v_1 \\ v_1 \end{pmatrix}$; $\begin{pmatrix} 1 \\ 1 \end{pmatrix}$ making it of unit norm, $\mathbf{v}^{(2)} = \begin{pmatrix} 1/\sqrt{2} \\ 1/\sqrt{2} \end{pmatrix}$

Note that both vectors are orthonormal

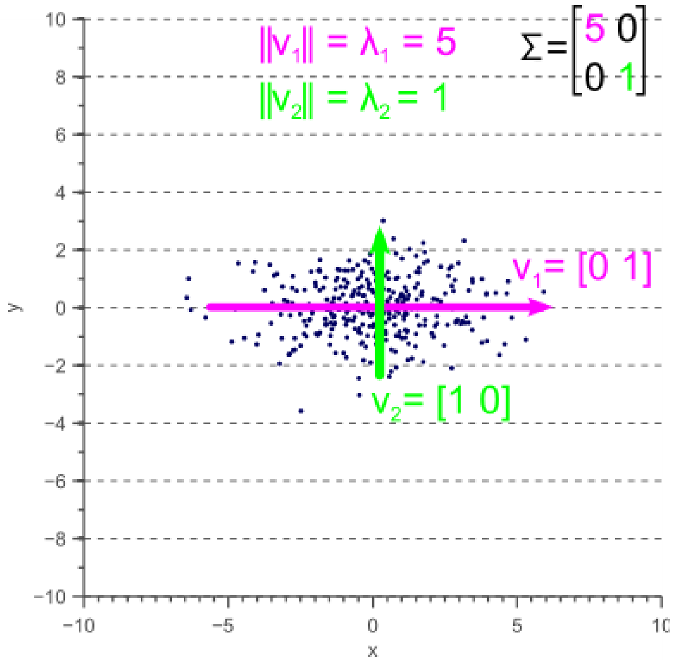
Linear algebra for the Gaussian Model

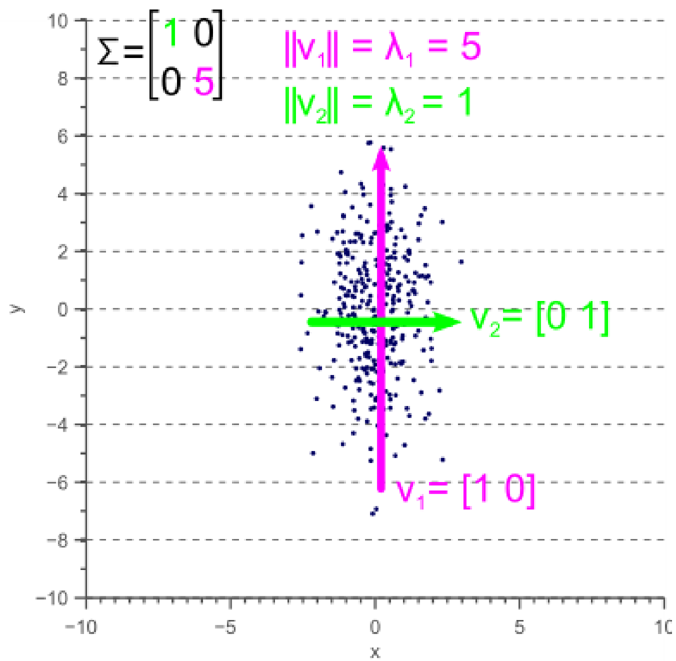
- The vector $\mathbf{v}^{(1)}$ is the "direction of maximum variance"
  If each point $\mathbf{x}$ is *perpendicularly* projected onto this line, the variance of this projected values is maximized
- The vector $\mathbf{v}^{(2)}$ is the "direction of *second* maximum variance"
- The Eigendecomposition of $\boldsymbol{\Sigma}$ gives us:
  - Eigenvectors: the vectors $\mathbf{v}^{(\mathrm{i})}$
  - Eigenvalues: $\lambda_i = \mathrm{Var}\left(\mathbf{u}^{(i)}\right), \mathrm{std}\left(\mathbf{u}^{(\mathrm{i})}\right) = \sqrt{\lambda_i}$

Linear algebra for the Gaussian Model

- The vector $\mathbf{v}^{(1)}$ is the "direction of maximum variance"
  If each point $\mathbf{x}$ is *perpendicularly* projected onto this line, the variance of this projected values is maximized
- The vector $\mathbf{v}^{(2)}$ is the "direction of *second* maximum variance"
- The Eigendecomposition of $\mathbf{\Sigma}$ gives us:
  - Eigenvectors: the vectors $\mathbf{v}^{(\mathrm{i})}$
  - Eigenvalues: $\lambda_i = \mathrm{Var}\left(\mathbf{u}^{(i)}\right), \mathrm{std}\left(\mathbf{u}^{(\mathrm{i})}\right) = \sqrt{\lambda_i}$

Diagonal Covariance Matrix (zero off-diagonal elements)

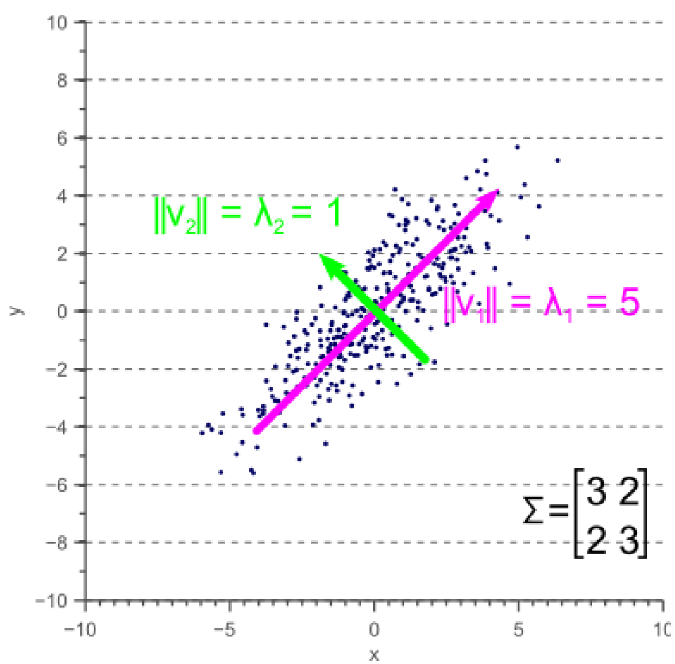


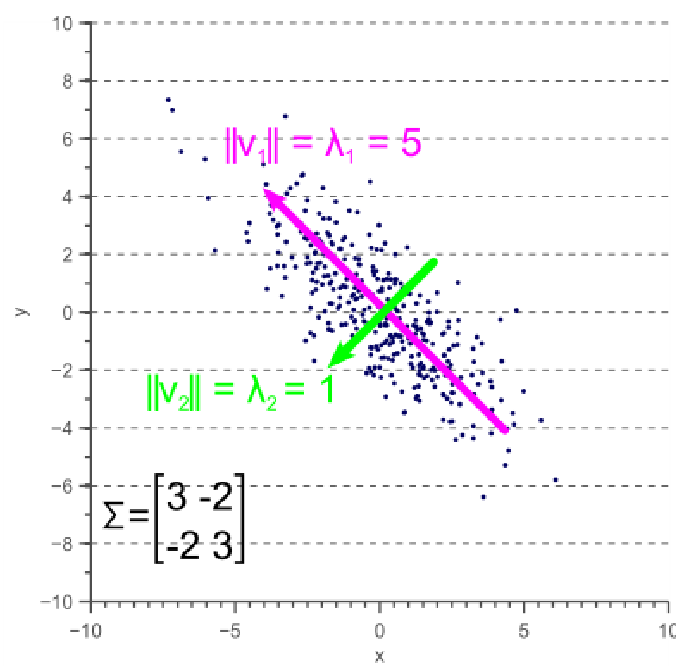


Linear algebra for the Gaussian Model

- The vector $\mathbf{v}^{(1)}$ is the "direction of maximum variance"
  If each point $\mathbf{x}$ is *perpendicularly* projected onto this line, the variance of this projected values is maximized
- The vector $\mathbf{v}^{(2)}$ is the "direction of *second* maximum variance"
- The Eigendecomposition of $\boldsymbol{\Sigma}$ gives us:
  - Eigenvectors: the vectors $\mathbf{v}^{(\mathrm{i})}$
  - Eigenvalues: $\lambda_i = \mathrm{Var}\left(\mathbf{u}^{(i)}\right), \mathrm{std}\left(\mathbf{u}^{(\mathrm{i})}\right) = \sqrt{\lambda_i}$

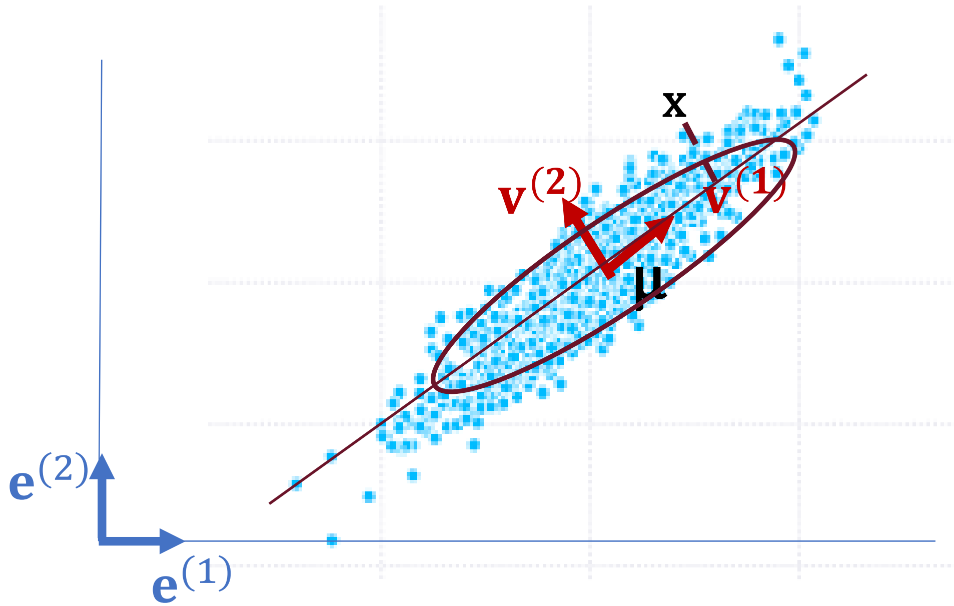
Non-Diagonal Covariance Matrix (not axis aligned anymore)





Linear algebra for the Gaussian Model

Covariance matrix as a linear transformation

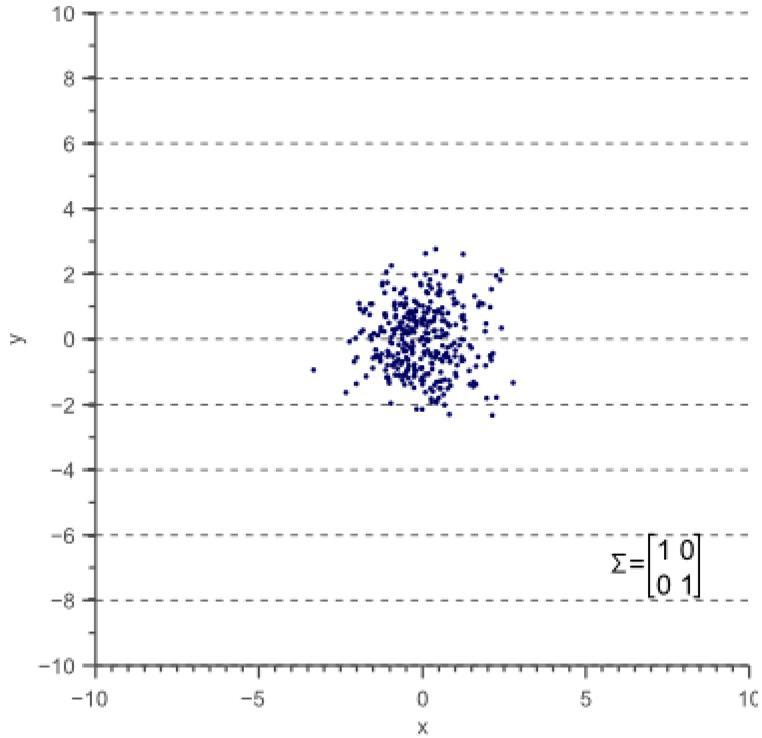


Remember the Gaussian Model: $f_x(\mathbf{x}) = \dfrac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$

- The expression $\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ Is the squared Mahalanobis Distance between $\mathbf{x}$ and $\boldsymbol{\mu}$
- The points at the same **Mahalanobis Distance** to the center have the same probability and form an ellipse (in 2D)

Linear algebra for the Gaussian Model

Covariance matrix as a linear transformation

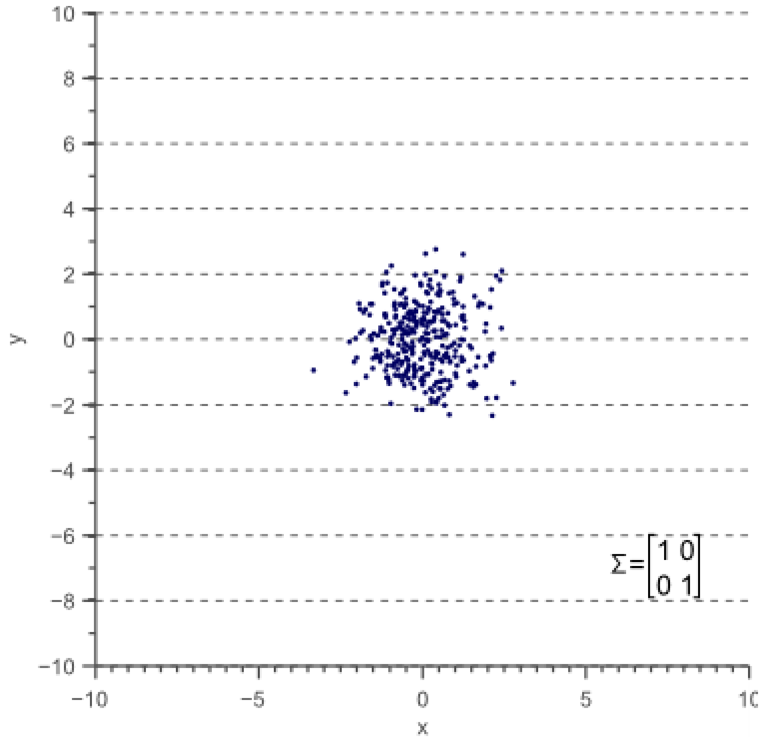


Start from the unit covariance

$$\Sigma = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

Linear algebra for the Gaussian Model

Covariance matrix as a linear transformation



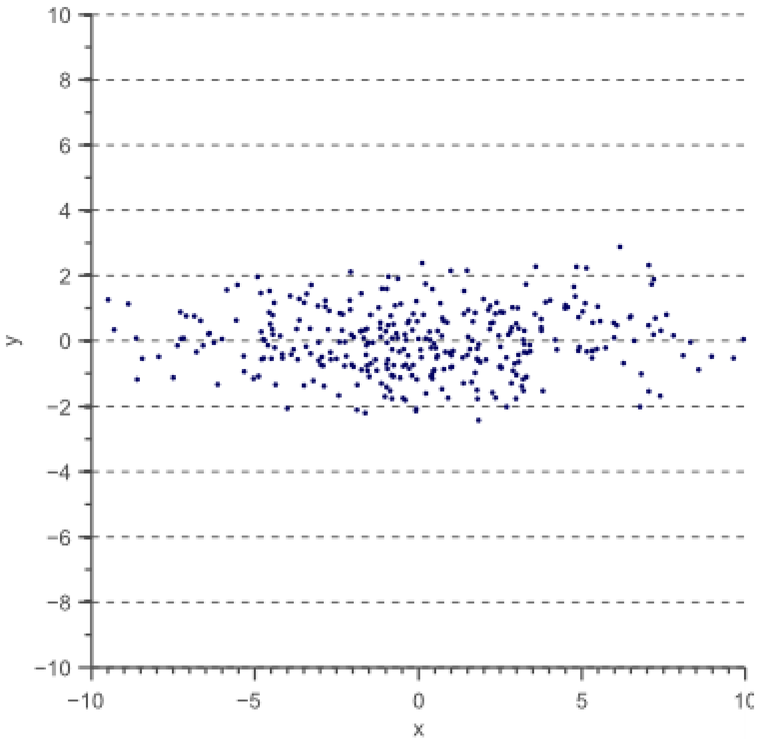
Start from the unit covariance

- Each of the previous examples can be obtained by a linear transformation of the data *D*

$$D' = T\,D \qquad T = R\,S. \qquad R = \begin{bmatrix} \cos(\theta) & -\sin(\theta) \\ \sin(\theta) & \cos(\theta) \end{bmatrix} \qquad S = \begin{bmatrix} s_x & 0 \\ 0 & s_y \end{bmatrix}$$

Linear algebra for the Gaussian Model

Covariance matrix as a linear transformation



- Example: scale the data in the x-direction by a factor 4

No rotation

$$D' = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix} D \qquad \Sigma' = \begin{bmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{bmatrix} = \begin{bmatrix} 16 & 0 \\ 0 & 1 \end{bmatrix} \qquad T = \sqrt{\Sigma'} = \begin{bmatrix} 4 & 0 \\ 0 & 1 \end{bmatrix}.$$

Linear algebra for the Gaussian Model

Covariance matrix as a linear transformation

- In general, Eigendecomposition (in matrix form)

$$\Sigma V = V L$$

Columns of V are eigenvectors of Σ

Diagonal matrix L with eigenvalues

- Covariance matrix Σ can be decomposed as $\Sigma = V L V^{-1}$
- Equivalently $\Sigma = R S S R^{-1}$

R represents a rotation matrix
$S = \sqrt{L}$ represents a scaling matrix

- Linear transformation $T = RS$
  Since S is diagonal, $S = S^\top$
  Since R is orthogonal, $R^{-1} = R^\top$

- Therefore $T^\top = (RS)^\top = S^\top R^\top = SR^{-1}$

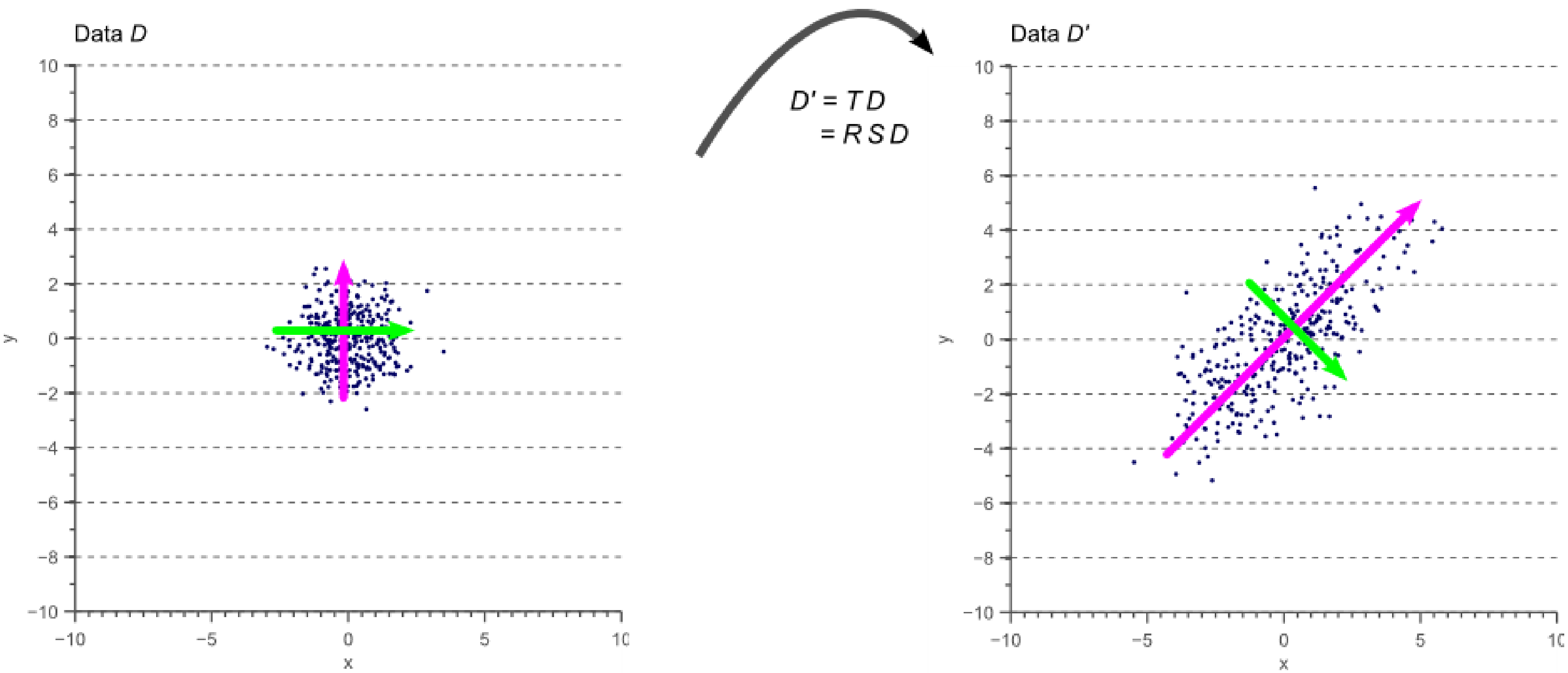
$$\Sigma = RSSR^{-1} = TT^\top$$

Linear algebra for the Gaussian Model

Covariance matrix as a linear transformation

- If we apply the linear transformation defined by $T = RS$ to the original white data, we obtain the rotated and scaled data with covariance

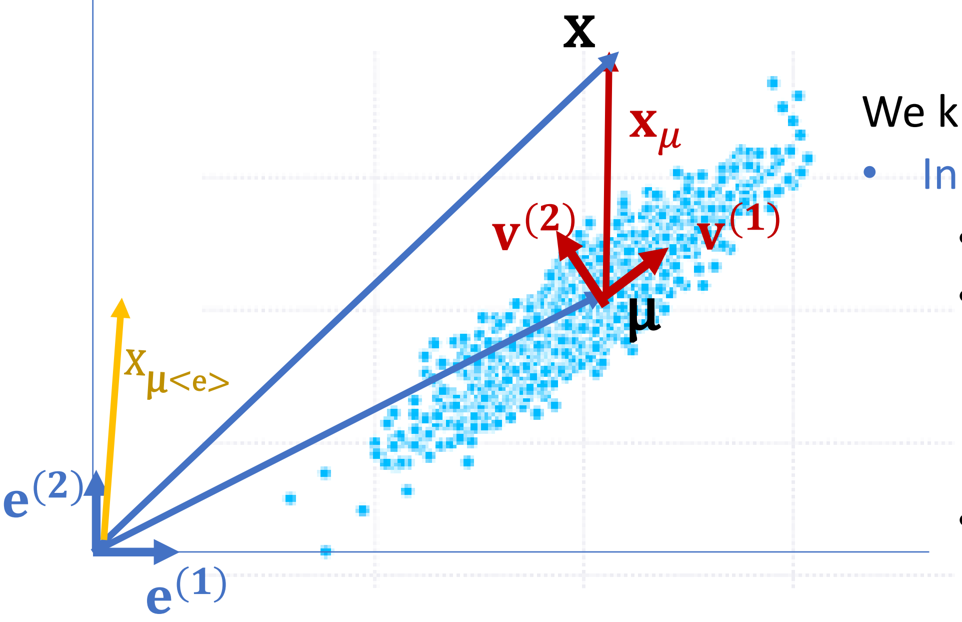
$$TT^\top = \Sigma' = RSSR^{-1}$$



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix} \qquad \Sigma' = \begin{bmatrix} 4.25 & 3.10 \\ 3.10 & 4.29 \end{bmatrix} = RSSR^{-1}$$

Linear algebra for the Gaussian Model

In 2D: we want to see the points from the basis $\langle \boldsymbol{\mu}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)} \rangle$

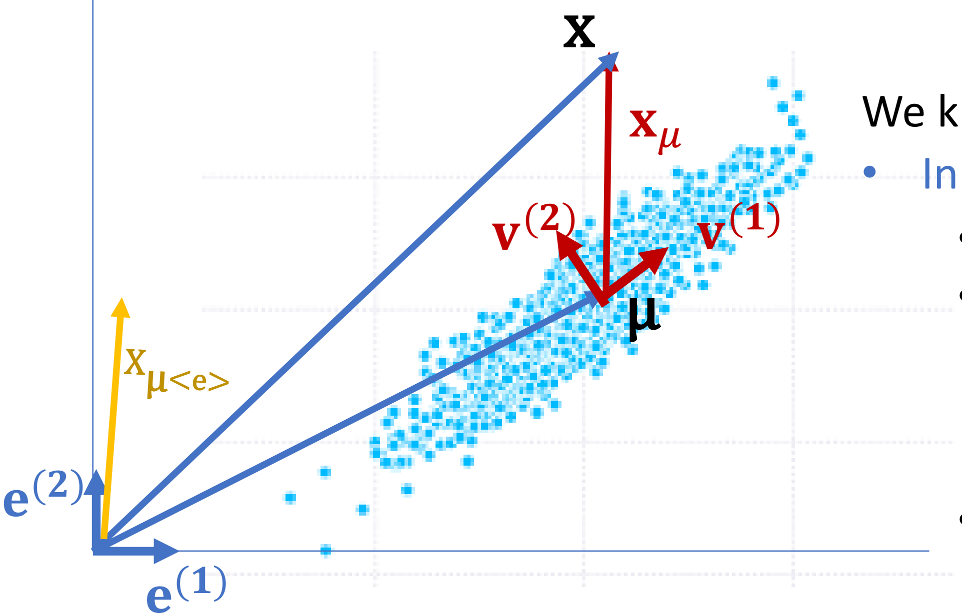


We know $\mathbf{x} = \boldsymbol{\mu} + \mathbf{x}_\mu$ then: $\mathbf{x}_\mu = \mathbf{x} - \boldsymbol{\mu}$, but:

- In which basis is expressed $\mathbf{x}_\mu$?
  - Canonical basis: $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}$
  - $\mathbf{x}_{\langle \mathbf{e} \rangle}$ means the vector $\mathbf{x}$ with respect to the basis $\langle \mathbf{e}^{(1)}, \mathbf{e}^{(2)} \rangle$
    $\mathbf{x}_{\langle \mathbf{e} \rangle} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \mathbf{e}^{(1)} + x_2 \mathbf{e}^{(2)}$
  - The vectors, before doing operations in coordinates, always in the same basis.

Linear algebra for the Gaussian Model

In 2D: we want to see the points from the basis $\langle \boldsymbol{\mu}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)} \rangle$



We know $\mathbf{x} = \boldsymbol{\mu} + \mathbf{x}_\mu$ then: $\mathbf{x}_\mu = \mathbf{x} - \boldsymbol{\mu}$, but:

- In which basis is expressed $\mathbf{x}_\mu$?
  - Canonical basis: $\mathbf{e}^{(1)}, \mathbf{e}^{(2)}$
  - $\mathbf{x}_{\langle \mathbf{e} \rangle}$ means the vector $\mathbf{x}$ with respect to the basis $\langle \mathbf{e}^{(1)}, \mathbf{e}^{(2)} \rangle$
    $\mathbf{x}_{\langle \mathbf{e} \rangle} = \begin{pmatrix} x_1 \\ x_2 \end{pmatrix} = x_1 \mathbf{e}^{(1)} + x_2 \mathbf{e}^{(2)}$
  - The vectors, before doing operations in coordinates, always in the same basis.

Given $\mathbf{x}_{\langle \mathbf{e} \rangle}$, how to calculate $\mathbf{x}_{\mu_{\langle \mathbf{v} \rangle}}$?

To achieve it, we need to express:

$$\mathbf{x}_{\mu_{\langle \mathbf{e} \rangle}} = x_1 \mathbf{e}^{(1)} + x_2 \mathbf{e}^{(2)} \text{ as } \mathbf{x}_{\mu_{\langle \mathbf{v} \rangle}} = w_1 \mathbf{v}^{(1)} + w_2 \mathbf{v}^{(2)}$$

Note that the vector $\mathbf{x}_{\mu_{\langle \mathbf{e} \rangle}}$ can be translated to the origin

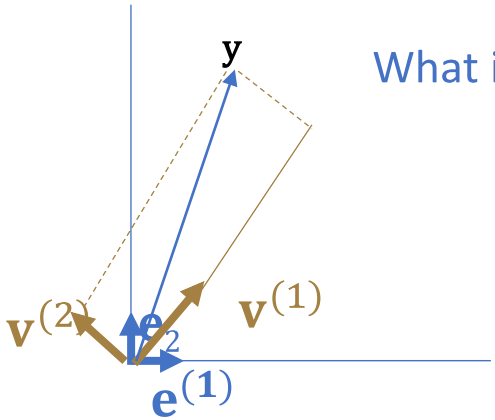
Linear algebra for the Gaussian Model

(simplified): given $\mathbf{y}_{\langle\mathbf{e}\rangle}$, calculate $\mathbf{y}_{\langle\mathbf{v}\rangle}$

To achieve it, we need to express the same point:

$$\mathbf{y}_{\langle\mathbf{e}\rangle} = y_1\mathbf{e}^{(1)} + y_2\mathbf{e}^{(2)} \text{ as } \mathbf{y}_{\langle\mathbf{v}\rangle} = w_1\mathbf{v}^{(1)} + w_2\mathbf{v}^{(2)}$$

Example: Obtain the vector $\mathbf{y}_{\langle\mathbf{e}\rangle} = \begin{pmatrix}1\\7\end{pmatrix}$ in the new basis $\mathbf{v}^{(1)} = \begin{pmatrix}1\\2\end{pmatrix}$, $\mathbf{v}^{(2)} = \begin{pmatrix}-2\\1\end{pmatrix}$



What is the meaning of $\begin{pmatrix}w_1\\w_2\end{pmatrix}$ in $\begin{pmatrix}1\\7\end{pmatrix} = \begin{pmatrix}1 & -2\\2 & 1\end{pmatrix}\begin{pmatrix}w_1\\w_2\end{pmatrix}$?

$$\begin{pmatrix}1 & -2\\2 & 1\end{pmatrix}\begin{pmatrix}w_1\\w_2\end{pmatrix} = \begin{pmatrix}w_1 - 2w_2\\2w_1 + w_2\end{pmatrix} = w_1\begin{pmatrix}1\\2\end{pmatrix} + w_2\begin{pmatrix}-2\\1\end{pmatrix} = w_1\mathbf{v}_1 + w_2\mathbf{v}_2$$

So, we need to solve the system: $\mathrm{A}\mathbf{y}_{\langle\mathbf{v}\rangle} = \mathbf{y}_{\langle\mathbf{e}\rangle}$

- Multiplying each side by the inverse matrix:

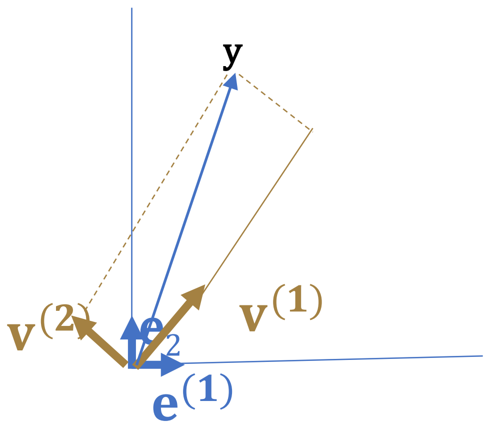
$$\begin{pmatrix}w_1\\w_2\end{pmatrix}_{\langle u\rangle} = \begin{pmatrix}1 & -2\\2 & 1\end{pmatrix}^{-1}\begin{pmatrix}1\\7\end{pmatrix} = \frac{1}{5}\begin{pmatrix}1 & 2\\-2 & 1\end{pmatrix}\begin{pmatrix}1\\7\end{pmatrix} = \begin{pmatrix}3\\1\end{pmatrix}$$

Linear algebra for the Gaussian Model

(simplified): given $\mathbf{y}_{\langle\mathbf{e}\rangle}$, calculate $\mathbf{y}_{\langle\mathbf{v}\rangle}$

To achieve it, we need to express the same point:

$$\mathbf{y}_{\langle\mathbf{e}\rangle} = y_1\mathbf{e}^{(1)} + y_2\mathbf{e}^{(2)} \text{ as } \mathbf{y}_{\langle\mathbf{v}\rangle} = w_1\mathbf{v}^{(1)} + w_2\mathbf{v}^{(2)}$$



$$\mathbf{y} = \begin{cases} \mathbf{y}_{\langle\mathbf{e}\rangle} = \begin{pmatrix}1\\7\end{pmatrix}_{\langle\mathbf{e}\rangle} \\ \mathbf{y}_{\langle\mathbf{v}\rangle} = \begin{pmatrix}3\\1\end{pmatrix}_{\langle\mathbf{v}\rangle} \end{cases}$$

A unique point expressed in two different coordinates systems

Rule:

1. Construct the matrix $\mathbf{A}$ with the basis vectors in columns
2. Calculate $\mathbf{A}^{-1}$
3. The new coordinates are: $\mathbf{y}_{\langle\mathbf{v}\rangle} = \mathbf{A}^{-1}\mathbf{y}_{\langle\mathbf{e}\rangle}$

- Important: If the matrix $\mathbf{A}$ is orthonormal, then $\mathbf{A}^{-1} = \mathbf{A}^{\mathsf{T}}$
  Orthonormal means: $\|\mathbf{v}_i\| = 1$ and $\mathbf{v}_i^{\mathsf{T}}\mathbf{v}_j = \delta_{ij}$ for all i, j

$\mathbf{A} = \begin{pmatrix} \frac{\sqrt{2}}{2} & -\frac{\sqrt{2}}{2} \\ \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$ is orthonormal, $\mathbf{A}^{-1} = \begin{pmatrix} \frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \\ -\frac{\sqrt{2}}{2} & \frac{\sqrt{2}}{2} \end{pmatrix}$

- We take the vector $\mathbf{x}_\mu = \mathbf{x} - \boldsymbol{\mu}$ expressed in the $\langle\mathbf{e}\rangle$ basis and transform it to the $\langle\boldsymbol{\mu}, \mathbf{v}\rangle$ basis

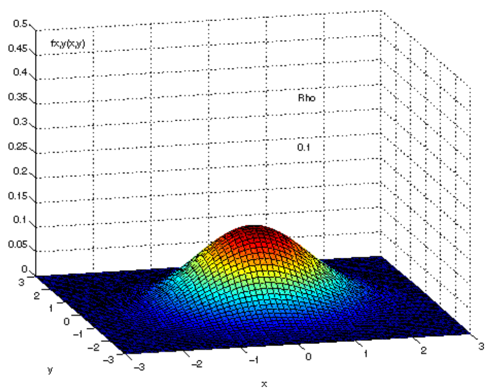
Properties of the Gaussian Model

– The multivariate Normal distribution $\mathcal{N}(\boldsymbol{\mu}, \boldsymbol{\Sigma})$ is defined as

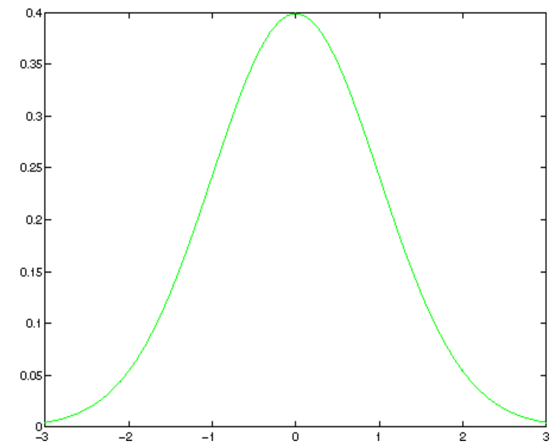
$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

- By Maximum likelihood we can estimate covariance $\boldsymbol{\Sigma}$ and $\boldsymbol{\mu}$ from the data

– For a single dimension, this expression is reduced to

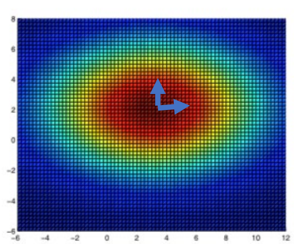


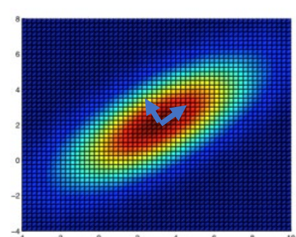
$$f_x(x) = \frac{1}{\sqrt{2\pi}\sigma} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$



Effects of the covariance matrix on the plot:

$$\mu = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 25 & 0 \\ 0 & 9 \end{pmatrix}$$





$$\mu = \begin{pmatrix} 3 \\ 2 \end{pmatrix} \quad \Sigma = \begin{pmatrix} 10 & 5 \\ 5 & 5 \end{pmatrix}$$

λ_1=13.09
λ_2=1.9

Variance in x_1 is 25. Variance in $x_2 = 9$

Variance in 1st direction is λ_1 and λ_2 in 2nd

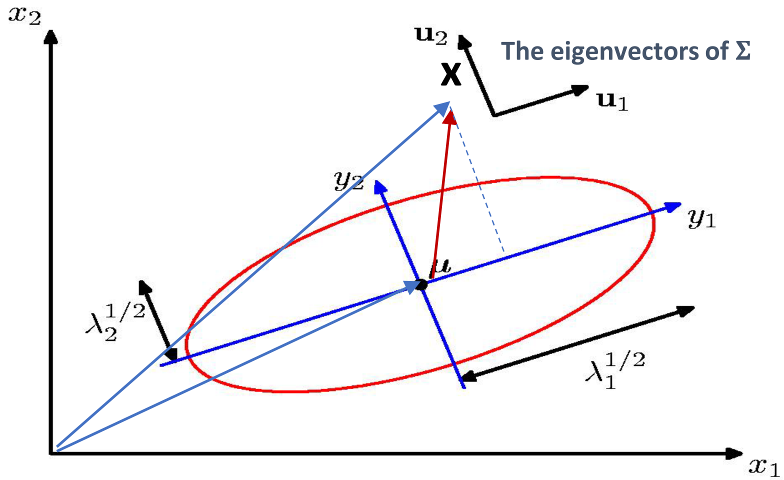
Properties of the Gaussian Model

$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\mathsf{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ **Squared Mahalanobis distance from *x* to $\boldsymbol{\mu}$**

All the points at the same Mahalanobis distance from μ have the same probability

λ_i is the variance along the axis $\mathbf{v}_i$



We know SVD: $\Sigma = \mathrm{VDV}^{\mathrm{T}}$

Properties:

1. $\Sigma = \sum_{i=1}^{D} \lambda_i \, \mathbf{v}^{(i)} \mathbf{v}^{(i)^{\mathsf{T}}}$ (spectral decomposition)

$$\Sigma = \mathrm{VDV}^{\mathrm{T}} = \begin{pmatrix} v_1^{(1)} & v_1^{(2)} \\ v_2^{(1)} & v_1^{(2)} \end{pmatrix} \begin{pmatrix} \lambda_1 & 0 \\ 0 & \lambda_2 \end{pmatrix} \begin{pmatrix} v_1^{(1)} & v_2^{(1)} \\ v_1^{(2)} & v_1^{(2)} \end{pmatrix} =$$

$$= \begin{pmatrix} \lambda_1 v_1^{(1)} v_1^{(1)} + \lambda_2 v_1^{(2)} v_1^{(2)} & \lambda_1 v_1^{(1)} v_2^{(1)} + \lambda_2 v_1^{(2)} v_2^{(2)} \\ \lambda_1 v_2^{(1)} v_1^{(1)} + \lambda_2 v_2^{(2)} v_1^{(1)} & \lambda_1 u_2^{(1)} u_2^{(1)} + \lambda_2 v_2^{(2)} v_2^{(2)} \end{pmatrix} =$$

$$= \lambda_1 \begin{pmatrix} v_1^{(1)} v_1^{(1)} & v_1^{(1)} v_2^{(1)} \\ v_2^{(1)} v_1^{(1)} & v_2^{(1)} v_2^{(1)} \end{pmatrix} + \lambda_2 \begin{pmatrix} v_1^{(2)} v_1^{(2)} & v_1^{(2)} v_2^{(2)} \\ v_2^{(2)} v_1^{(2)} & v_2^{(2)} v_2^{(2)} \end{pmatrix} =$$

$$= \lambda_1 \begin{pmatrix} v_1^{(1)} \\ v_2^{(1)} \end{pmatrix} \begin{pmatrix} v_1^{(1)} & v_2^{(1)} \end{pmatrix} + \lambda_2 \begin{pmatrix} v_1^{(2)} \\ v_2^{(2)} \end{pmatrix} \begin{pmatrix} v_1^{(2)} & v_2^{(2)} \end{pmatrix} =$$

$$= \lambda_1 \mathbf{v}^{(1)} \mathbf{v}^{(1)^{\mathsf{T}}} + \lambda_2 \mathbf{v}^{(2)} \mathbf{v}^{(2)^{\mathsf{T}}}$$

Properties of the Gaussian Model

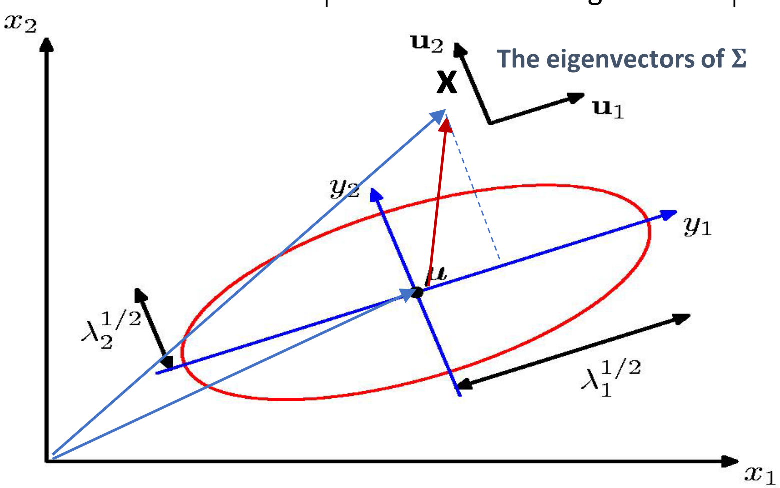
$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

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

Squared Mahalanobis distance from x to μ

All the points at the same Mahalanobis distance from μ have the same probability

λ_i is the variance along the axis $\mathbf{v}_i$



We know SVD: $\Sigma = \mathrm{VDV}^\mathrm{T}$

Properties:

1. $\Sigma = \sum_{i=1}^{D} \lambda_i\, \mathbf{v}^{(i)} \mathbf{v}^{(i)^\top}$ (spectral decomposition)
2. $\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{v}^{(i)} \mathbf{v}^{(i)^\top}$

$\Sigma = \mathrm{VDV}^\top$ so

$$\Sigma^{-1} = (\mathrm{VDV}^\top)^{-1} = \mathrm{V}^{-1}\mathrm{D}^{-1}(\mathrm{V}^\top)^{-1} = \mathrm{V}^\top \mathrm{D}^{-1}\mathrm{V}$$

$$= \mathrm{V}\begin{pmatrix} 1/\lambda_1 & 0 \\ 0 & 1/\lambda_2 \end{pmatrix}\mathrm{V}^\top = \frac{1}{\lambda_1}\mathbf{v}^{(1)}\mathbf{v}^{(1)^\top} + \frac{1}{\lambda_2}\mathbf{v}^{(2)}\mathbf{v}^{(2)^\top}$$

Note that $\begin{pmatrix} a & 0 \\ 0 & b \end{pmatrix}^{-1} = \begin{pmatrix} 1/a & 0 \\ 0 & 1/b \end{pmatrix}$

Properties of the Gaussian Model

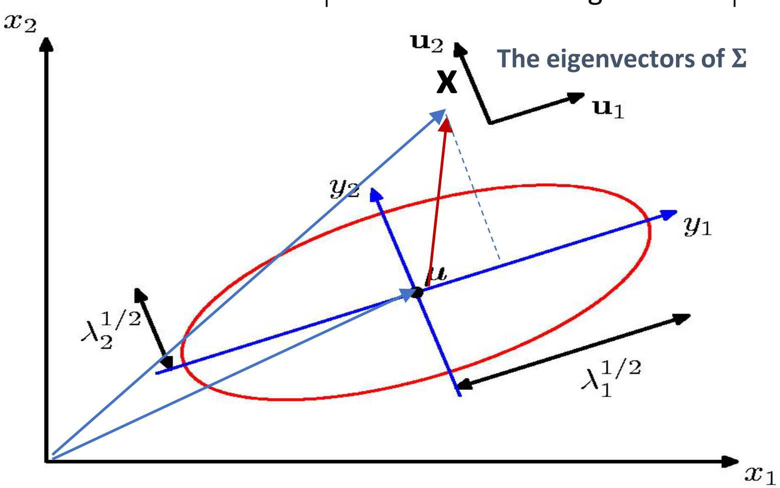
$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

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

Squared Mahalanobis distance from x to μ

All the points at the same Mahalanobis distance from μ have the same probability

λ_i is the variance along the axis $\mathbf{v}_i$



We know SVD: $\Sigma = \mathrm{VDV}^\mathrm{T}$

Properties:

1. $\Sigma = \sum_{i=1}^{D} \lambda_i \, \mathbf{v}^{(i)} \mathbf{v}^{(i)^\top}$ (spectral decomposition)

2. $\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{v}^{(i)} \mathbf{v}^{(i)^\top}$

3. $y_i = \mathbf{v}_i^\top (\mathbf{x}-\boldsymbol{\mu})$ is the projection of $(\mathbf{x}-\boldsymbol{\mu})$ onto the vector $\mathbf{v}^{(i)}$

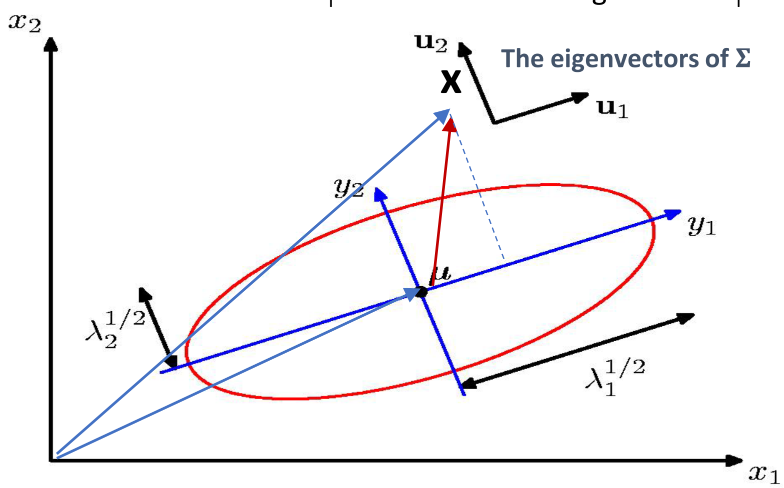
Properties of the Gaussian Model

$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\mathbf{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\top}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

$\Delta^2 = (\mathbf{x}-\boldsymbol{\mu})^{\top}\mathbf{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})$ **Squared Mahalanobis distance from x to μ**

All the points at the same Mahalanobis distance from μ have the same probability

λ_i is the variance along the axis $\mathbf{v}_i$



4. The *squared* Mahalanobis distance: $\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$

$$\Delta^2 = (\mathbf{x}-\mu)^{\top}\Sigma^{-1}(\mathbf{x}-\mu) =$$

$$= (\mathbf{x}-\mu)^{\top}\left(\frac{1}{\lambda_1}\mathbf{v}^{(1)}\mathbf{v}^{(1)\top} + \frac{1}{\lambda_2}\mathbf{v}^{(2)}\mathbf{v}^{(2)\top}\right)(\mathbf{x}-\mu)$$

$$= \frac{1}{\lambda_1}\boxed{(\mathbf{x}-\mu)^{\top}\mathbf{v}^{(1)}}\boxed{\mathbf{v}^{(1)\top}(\mathbf{x}-\mu)}$$

$$+ \frac{1}{\lambda_2}(\mathbf{x}-\mu)^{\top}\mathbf{v}^{(2)}\mathbf{v}^{(2)\top}(\mathbf{x}-\mu)$$

$$= \frac{1}{\lambda_1}y_1y_1 + \frac{1}{\lambda_2}y_2y_2 = \frac{y_1^2}{\lambda_1} + \frac{y_2^2}{\lambda_2}$$

Note that: $\boxed{(\mathbf{x}-\mu)^{\top}\mathbf{v}^{(1)}} = (\mathbf{v}^{(1)\top}(\boldsymbol{x}-\mu))^{\top} = y_i^{\top} = y_i$

Remember:

1. Ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 1$ are the points at Mahalanobis Distance 1 are the ellipse with axis $\sqrt{\lambda_1}, \sqrt{\lambda_2}$
2. $\Delta^2 = 4$ are the points of the ellipse $\frac{x^2}{a^2} + \frac{y^2}{b^2} = 4$. That is $\frac{x^2}{4a^2} + \frac{y^2}{4b^2} = 1$ with axes $2a, 2b$

Properties of the Gaussian Model

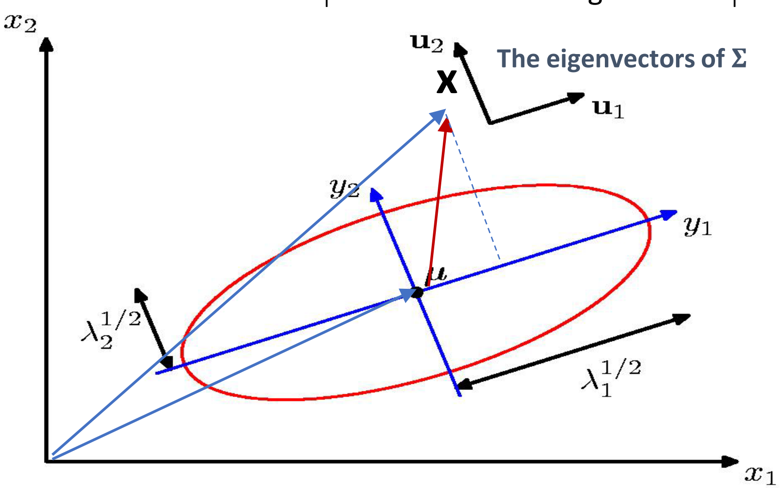
$$f_x(\mathbf{x}) = \frac{1}{(2\pi)^{D/2}|\boldsymbol{\Sigma}|^{1/2}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})}$$

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

Squared Mahalanobis distance from x to μ

All the points at the same Mahalanobis distance from μ have the same probability

λ_i is the variance along the axis $\mathbf{v}_i$



We know SVD: $\Sigma = \mathrm{VDV}^\mathrm{T}$

Properties:

1. $\Sigma = \sum_{i=1}^{D} \lambda_i \, \mathbf{v}^{(i)} \mathbf{v}^{(i)^\top}$ (spectral decomposition)

2. $\Sigma^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{v}^{(i)} \mathbf{v}^{(i)^\top}$

3. $y_i = \mathbf{v}_i^\top (\mathbf{x}-\boldsymbol{\mu})$ is the projection of $(\mathbf{x}-\boldsymbol{\mu})$ onto the vector $\mathbf{v}^{(i)}$

4. The *squared* Mahalanobis distance

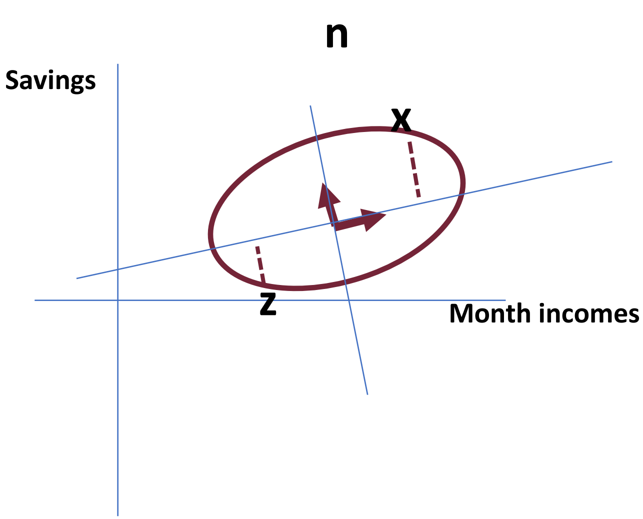
$$\Delta^2 = \sum_{i=1}^{D} \frac{y_i^2}{\lambda_i}$$

- Ellipse with all the points with the same probability.
- Length axis $i = \sqrt{\lambda_i}$ λ_i =Var($\mathbf{v}^{(i)}$)

Applications of the Gaussian Model

Up to now the algebraic part

- Imagine you are client of a bank. Imagine also that each month you have a money assignation and you perhaps finish the month with a small saving.
- To the Data Scientist of the bank, you are interesting for three reasons (at least)
    - your each month savings
    - the money you do not use during the month
    - the cluster you belong to
        - with month savings: possible consumer of new products
        - No savings: special offers, discount tickets
        - Not only those, see the plot



- First, covariance positive: more incomes, more savings
- Clustering offer more:
    - x and z have the same probability different sign projection (to $\mathbf{u}_1$)
        - **x** don't waste money (offer products!, new features: age?)
        - **z** few savings, Special Offers
        - **n** outlier: consider apart

Summary

- Data in the space appears generating some clouds
- Covariance Matrix helps to extract the geometrical properties of these structures
- We can better understand the covariance matrix using eigenvectors, eigenvectors and the Mahalanobis distance.
- The Gaussian Model is very useful:
  - for its clear geometrical properties
  - with few parameters we can summarize lot of points
  - gives a probability, so we can exploit that for prediction (new cases, not seen before).
  - It is a generative model.

Exercises

Example (Exam June 2018)

1. Consider a Gaussian centered in $\vec{\mu}_1 = \binom{0}{4}$ and with covariance matrix $\Sigma = \begin{pmatrix} 4 & 0 \\ 0 & 9 \end{pmatrix}$. Another Gaussian with the same covariance has the center in $\vec{\mu}_2 = \binom{4}{0}$.

   (a) **(1pt)** What is the inverse of the matrix Σ?. Write the expression that give us the Mahalanobis distance from one point $\vec{x} = \binom{x}{y}$ to the first Gaussian.

   (b) **(0.5pt)** Plot the curves that are at a Mahalanobis distance of 1 and 2 for each Gaussian.

   (c) **(0.5pt)** Give a point that is equidistant (in the Euclidean sense) from the two centers and closest to the first Gaussian using the Mahalanobis distance.

Solution

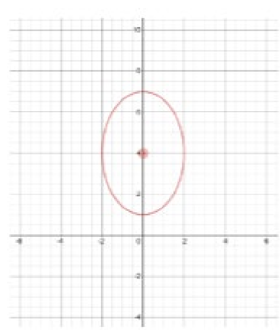
a)

$$\Sigma^{-1} = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/9 \end{pmatrix}$$

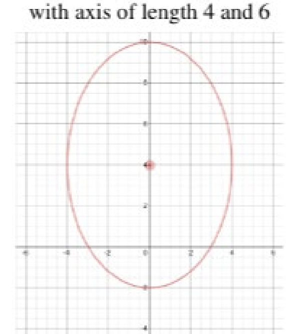
$$D_1(\vec{x}) = \sqrt{(\vec{x}-\mu_1)^\top \Sigma^{-1} (\vec{x}-\mu_1)^\top} = \sqrt{\frac{x^2}{4} + \frac{(y-4)^2}{9}}$$

b)

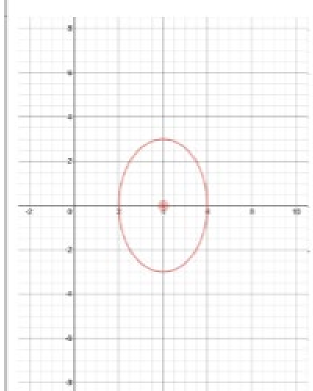
$D_1(\vec{x}) = 1 \rightarrow 1 = \frac{x^2}{4} + \frac{(y-4)^2}{9}$



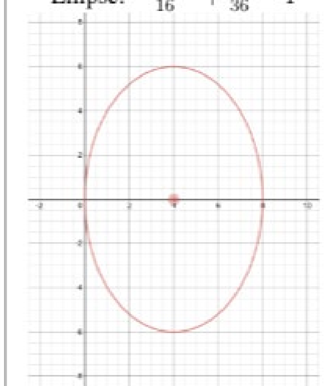
$D_1(\vec{x}) = 2 \rightarrow 4 = \frac{x^2}{4} + \frac{(y-4)^2}{9}$.
The points belong to the ellipse:
$1 = \frac{x^2}{16} + \frac{(y-4)^2}{36}$
with axis of length 4 and 6



$D_2(\vec{x}) = 1 :\rightarrow \frac{(x-4)^2}{4} + \frac{y^2}{9}$



$D_2(\vec{x}) = 2 \rightarrow 4 = \frac{(x-4)^2}{4} + \frac{y^2}{9}$
Ellipse: $\frac{(x-4)^2}{16} + \frac{y^2}{36} = 1$



c) The point (0,0), since it is at distance 4/3 from the first Gaussian and 2 from the second one. (0,0) is closer to N_1 in the Mahalanobis sense.

Exercises

Example (Exam June 2019)

We have a random variable $\mathbf{x} = \binom{x_1}{x_2}$ normally distributed according to $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ with mean $\boldsymbol{\mu} = \binom{2}{3}$ and covariance matrix Σ given by the eigenvectors $\mathbf{u}_1 = \begin{pmatrix} 1/2 \\ \frac{\sqrt{3}}{2} \end{pmatrix}$ and $\mathbf{u}_2 = \begin{pmatrix} -\frac{\sqrt{3}}{2} \\ 1/2 \end{pmatrix}$, with eigenvalues 4 and 2 respectively.

1. **(0.5pt)** Calculate the covariance matrix Σ.

> **Solution:** We give three solutions
>
> (a) $\Sigma = UDU^T = \begin{pmatrix} 1/2 & -\frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & 1/2 \end{pmatrix} \begin{pmatrix} 4 & 0 \\ 0 & 2 \end{pmatrix} \begin{pmatrix} 1/2 & \frac{\sqrt{3}}{2} \\ -\frac{\sqrt{(3)}}{2} & 1/2 \end{pmatrix} = \begin{pmatrix} 5/2 & \frac{\sqrt{3}}{2} \\ \frac{\sqrt{3}}{2} & 7/2 \end{pmatrix}$
>
> (b) $\Sigma = 4\mathbf{u}_1 * \mathbf{u}_1^T + 2\mathbf{u}_2 * \mathbf{u}_2^T$
>
> (c) Σ is a symmetric matrix with 3 unknows: $\Sigma = \begin{pmatrix} a & b \\ b & c \end{pmatrix}$ and we know the eigenvectors and eigenvalues. Using $\begin{pmatrix} a-4 & b \\ b & c-4 \end{pmatrix} \mathbf{u}_1 = \mathbf{0}$ and $\begin{pmatrix} a-2 & b \\ b & c-2 \end{pmatrix} \mathbf{u}_2 = \mathbf{0}$. Then Σ is obtained after solving the resulting four linear equations.

2. **(0.5pt)** Indicate how you can use this to estimate the Σ^{-1}. What are its eigenvalues?

> **Solution:** $\Sigma^{-1} = (UDU^T)^{-1} = UD^{-1}U^T = \begin{pmatrix} 0.44 & -0.10 \\ 0.10 & 0.31 \end{pmatrix}$ where $D^{-1} = \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix}$.
>
> From the last expression, the eigenvectors are the same as for Σ and the eigenvalues are $\lambda_1 = 1/4$ and $\lambda_2 = 1/2$.

Exercises

Example (Exam June 2019)

...

3. **(0.5pt)** Draw the points that are at Mahalanobis distance from $\boldsymbol{\mu}$ equal to 1 according to the distribution $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$. The same at distance 2. [Remember the Mahalanobis distance is defined as $d(\mathbf{x}, \mathbf{y}) = \sqrt{(\mathbf{x}-\mathbf{y})^T \Sigma^{-1} (\mathbf{x}-\mathbf{y})}$].

Solution: The points $\mathbf{x}$ such that $d(\mathbf{x}, \boldsymbol{\mu}) = 1$ are
$1^2 = (\mathbf{x}-\boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x}-\boldsymbol{\mu}) = (\mathbf{x}-\boldsymbol{\mu})^T U D^{-1} U^T (\mathbf{x}-\boldsymbol{\mu})$.
Then $U^T(\mathbf{x}-\boldsymbol{\mu}) = \begin{pmatrix} x' \\ y' \end{pmatrix}$ defines a vector with coordinates of $\mathbf{x}$ in the new coordinate system $\langle \boldsymbol{\mu}, \mathbf{u}_1, \mathbf{u}_2 \rangle$ then $1^2 = \begin{pmatrix} x' & y' \end{pmatrix} \begin{pmatrix} 1/4 & 0 \\ 0 & 1/2 \end{pmatrix} \begin{pmatrix} x' \\ y' \end{pmatrix}$, so $1 = \frac{x'^2}{4} + \frac{y'^2}{2}$.

The solution is given by the points of the ellipse centered in $\boldsymbol{\mu}$ and long axis the direction $\mathbf{u}_1$ with major axis 2, and minor axis $\sqrt{2}$.

The same with $2^2 = \frac{x'^2}{4} + \frac{y'^2}{2}$, so $1 = \frac{x'^2}{16} + \frac{y'^2}{8}$ and the solution is given by the points of the ellipse centered in $\boldsymbol{\mu}$ and long axis the direction $\mathbf{u}_1$ with major axis 4, and minor axis $2\sqrt{2}$.