Mathematics for Machine Learning (AI 512):

Module 4

Amit Chattopadhyay

IIIT-Bangalore

Module 4:

- **K-Means and GMM**: Hard Clustering and Soft Clustering
- **Expectation-Maximization** (EM) framework
- **Expectation-Maximization** (EM) for **Gaussian Mixture Model** (GMM) parameter estimation

Reference books:

1. *Mathematics for Machine Learning*, by Marc Peter Deisenroth, A. Aldo Faisal and Cheng Soon Ong. (Ch 5, **Ch. 11**)
2. *Pattern Recognition and Machine Learning*, by Chrisopher M. Bishop. (Ch 2.3, **Ch. 9**)
3. *Machine Learning - A Probabilistic Perspective* by Kevin Murphy, Ch 11

Multivariate Gaussian

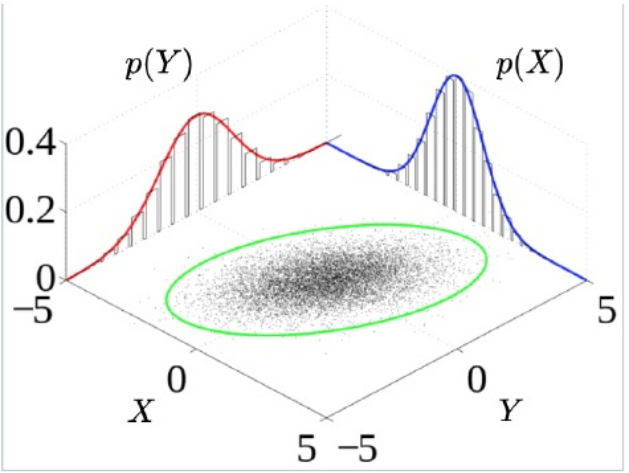
Covariance and Correlation: Bivariate

For two random variables X_1 and X_2:

- joint distribution: $p(x_1, x_2)$

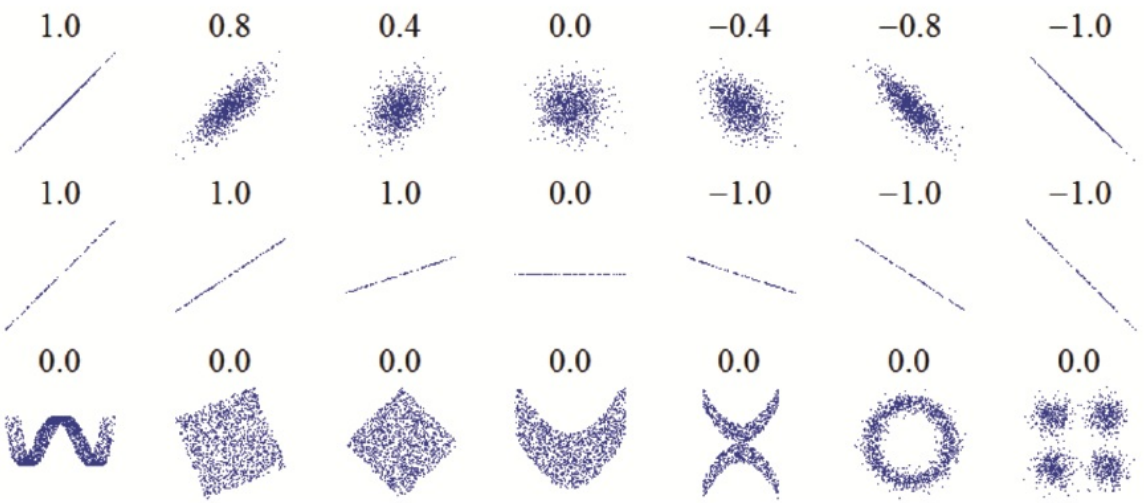
Covariance and Correlation of (X_1, X_2)

1. $\text{Cov}[X_1, X_2] = \mathbb{E}[(X_1 - \mathbb{E}(X_1))(X_2 - \mathbb{E}(X_2))] = \mathbb{E}[X_1 X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$
2. $\rho[X_1, X_2] = \frac{Cov[X_1, X_2]}{\sqrt{Var[X_1]}\sqrt{Var[X_2]}}$



Covariance and Correlation: Bivariate

- $-1 \leq \rho[X_1, X_2] \leq 1$
- $\rho[X_1, X_2] = 1$ iff $X_1 = aX_2 + b$ for some constants a and b, i.e. if there's a linear relationship between X_1 and X_2
- if $\rho[X_1, X_2] = 0$, X_1, X_2 are called **uncorrelated**
- If X_1, X_2 are independent, then X_1, X_2 are uncorrelated. Converse is not true.



Covariance and Correlation: d-Variate

For a random vector $\mathbf{X} = (X_1, \ldots, X_d)$:

- joint distribution: $p(x_1, x_2, \ldots, x_d)$

$$E(\mathbf{X}) = \begin{pmatrix} E(X_1) \\ E(X_2) \\ \vdots \\ E(X_d) \end{pmatrix}$$

Covariance Matrix

$$\mathrm{Cov}(\mathbf{X}) = \mathbb{E}\left[(\mathbf{X} - \mathbb{E}(\mathbf{X}))(\mathbf{X} - \mathbb{E}(\mathbf{X}))^T\right]$$

$$= \begin{bmatrix} \mathrm{Var}[X_1] & \mathrm{Cov}[X_1, X_2] & \ldots & \mathrm{Cov}[X_1, X_d] \\ \mathrm{Cov}[X_2, X_1] & \mathrm{Var}[X_2] & \ldots & \mathrm{Cov}[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{Cov}[X_d, X_1] & \mathrm{Cov}[X_d, X_2] & \ldots & \mathrm{Var}[X_d] \end{bmatrix}$$

Covariance and Correlation: d-Variate

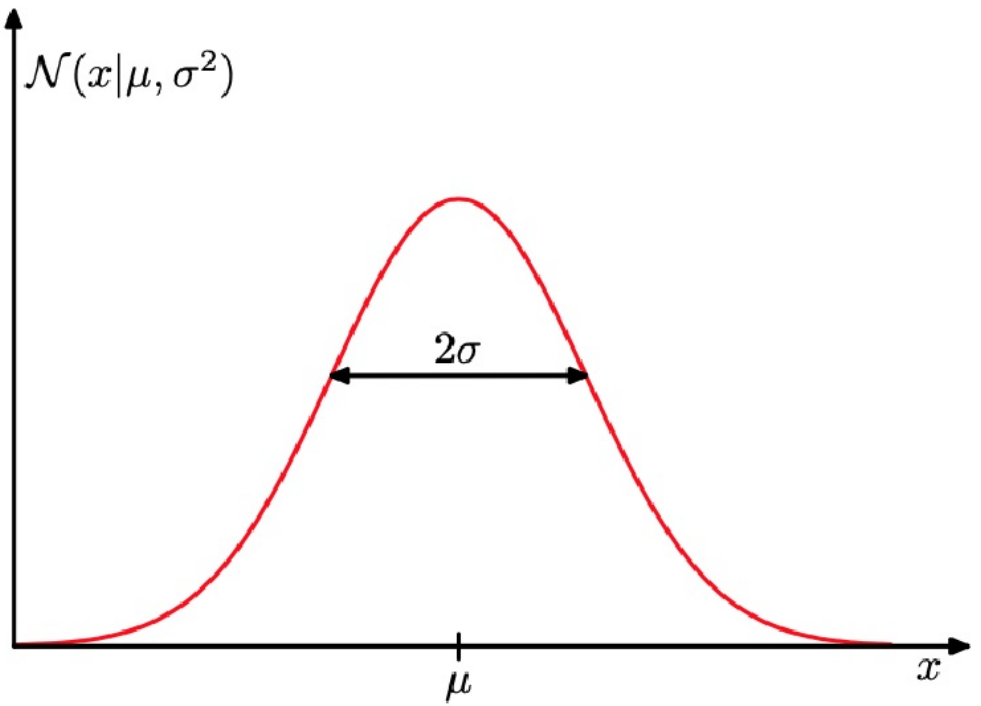
For a random vector $\mathbf{X} = (X_1, \ldots, X_d)$:

(Pearson) Correlation Matrix

$$\rho[\mathbf{X}] = \begin{bmatrix} \rho[X_1, X_1] & \rho[X_1, X_2] & \cdots & \rho[X_1, X_d] \\ \rho[X_2, X_1] & \rho[X_2, X_2] & \cdots & \rho[X_2, X_d] \\ \vdots & \vdots & \ddots & \vdots \\ \rho[X_d, X_1] & \rho[X_d, X_2] & \cdots & \rho[X_d, X_d] \end{bmatrix}$$

Gaussian: Univariate

$$\mathcal{N}(x;\mu,\sigma) = \frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{1}{2}(x-\mu)^2\right\}, \quad -\infty < x < \infty$$



Gaussian: Multivariate

$$\mathcal{N}(\mathbf{x}; \underline{\mu}, \Sigma) = \frac{1}{\sqrt{2\pi}|\Sigma|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\underline{\mu})^T\Sigma^{-1}(\mathbf{x}-\underline{\mu})\right\}, \quad -\infty < x_i < \infty$$

- $\underline{\mu} = \mathbb{E}[\mathbf{X}]$: mean vector in $\mathbb{R}^d$, d parameters
- $\Sigma = \mathrm{Cov}[\mathbf{X}]$: $d \times d$ matrix. In general, $\frac{d(d+1)}{2}$ parameters
- If Σ is diagonal: d parameters
- $|\Sigma|$: Determinant of Σ
- If $\Sigma = \sigma^2\mathbb{I}_d$: spherical/ isotropic covariance, one free parameter

Σ: Properties

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

  Δ: **Mahalanobis Distance**

  For $\Sigma = I_d$, Δ is Euclidean distance

- Σ: real, symmetric matrix
- Eigendecomposition of Σ : $\Sigma = U \Lambda U^T = \sum_{i=1}^{d} \lambda_i \underline{u}_i \underline{u}_i^T$

$$\Delta^2 = \sum_{i=1}^{d} \frac{1}{\lambda_i} (\mathbf{x} - \underline{\mu})^T \underline{u}_i \underline{u}_i^T (\mathbf{x} - \underline{\mu})$$

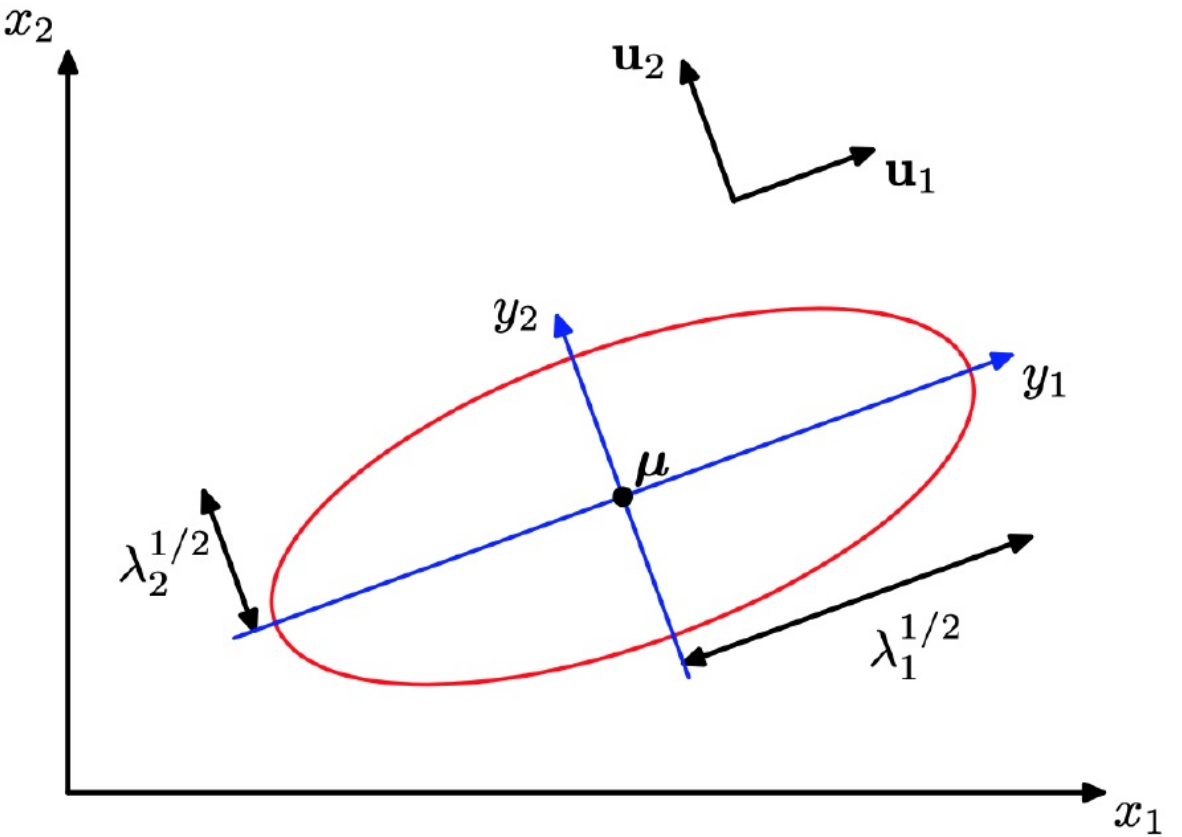
$$= \sum_{i=1}^{d} \frac{y_i^2}{\lambda_i}$$

where $y_i = \underline{u}_i^T (\mathbf{x} - \underline{\mu})$

($\{y_i\}$ form new coordinate system w.r.t. orthonormal vectors: $\{\underline{u}_i\}$)

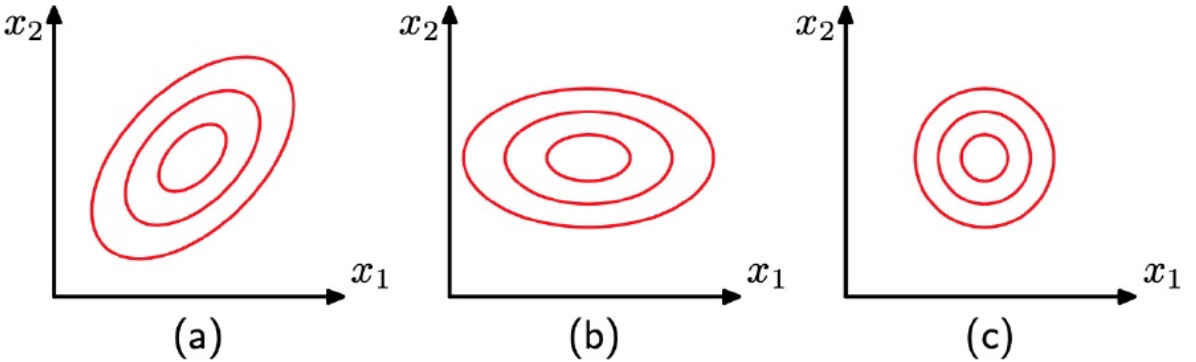
Ploting Isocurve of a Bivariate Gaussian

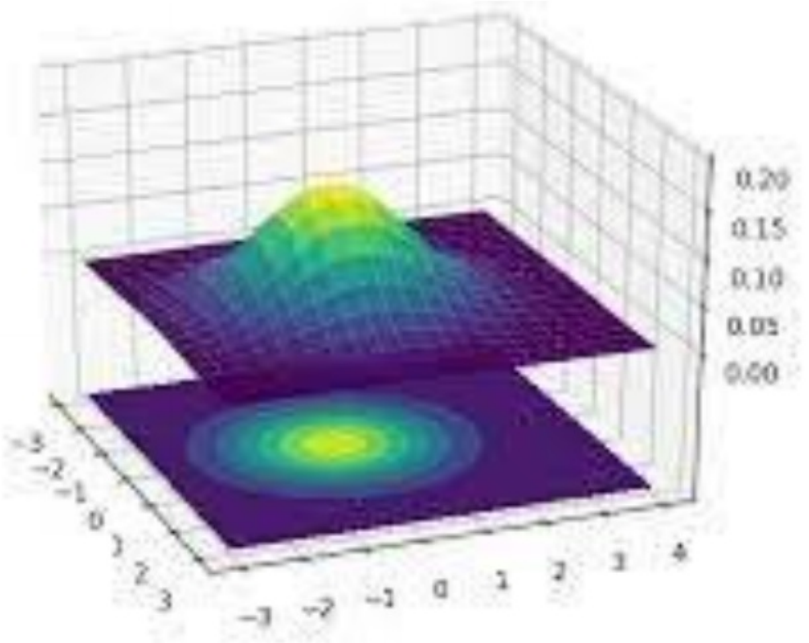
The red curve shows the elliptical surface of constant probability density for a Gaussian in a two-dimensional space $\mathbf{x} = (x_1, x_2)$ on which the density is $\exp(-1/2)$ of its value at $\mathbf{x} = \boldsymbol{\mu}$. The major axes of the ellipse are defined by the eigenvectors $\mathbf{u}_i$ of the covariance matrix, with corresponding eigenvalues λ_i.



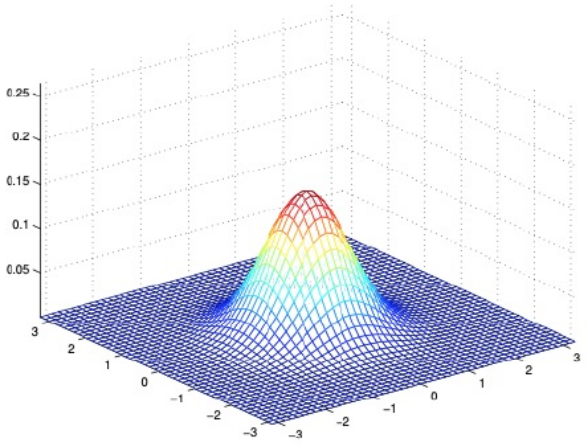
Ploting Isocurves of Bivariate Gaussians

Figure 2.8 Contours of constant probability density for a Gaussian distribution in two dimensions in which the covariance matrix is (a) of general form, (b) diagonal, in which the elliptical contours are aligned with the coordinate axes, and (c) proportional to the identity matrix, in which the contours are concentric circles.

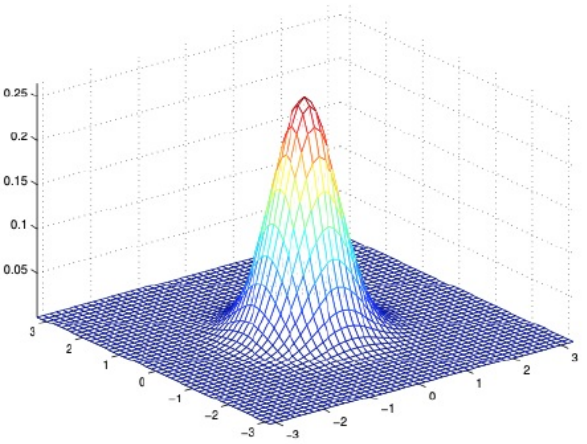




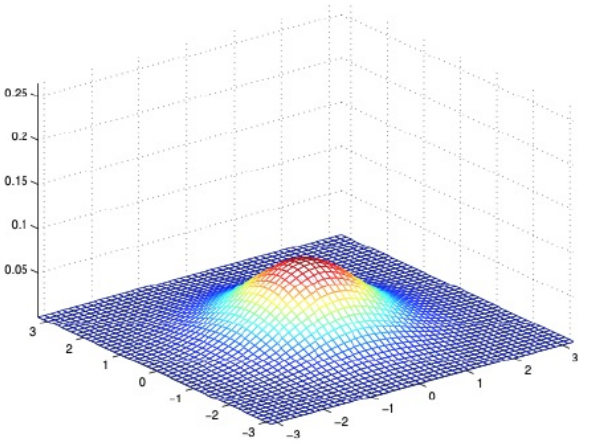
Covariance Matrix: Analysis



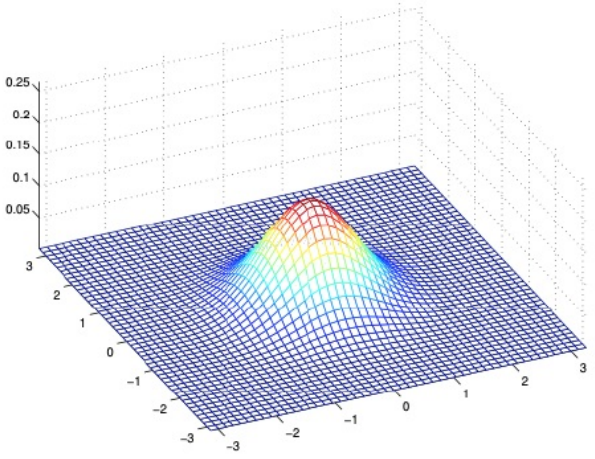
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



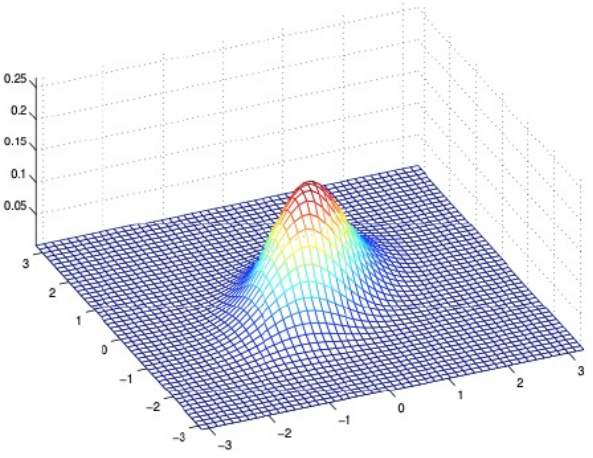
$$\Sigma = 0.6 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



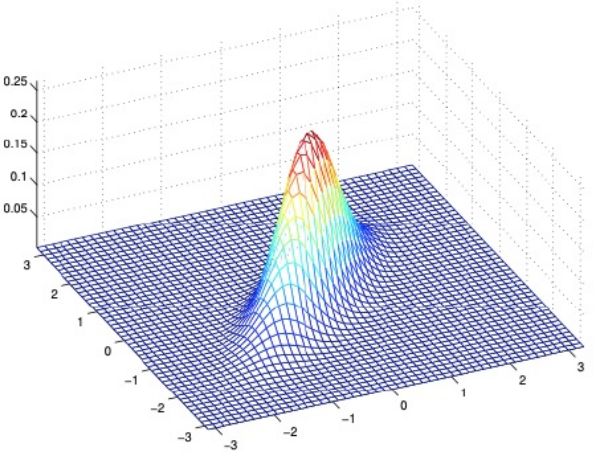
$$\Sigma = 2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$

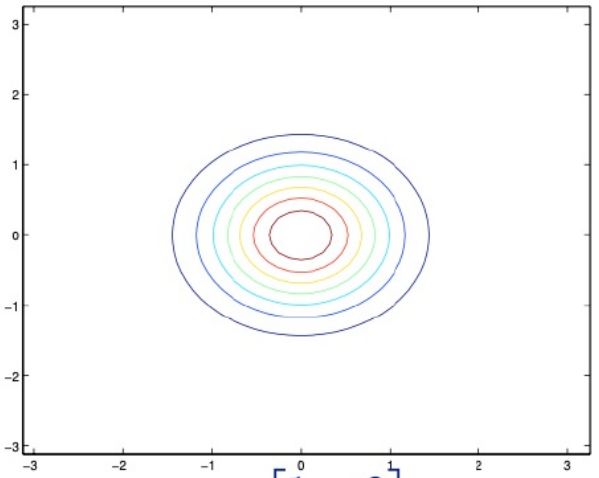


$$\Sigma = \begin{bmatrix} 1 & .5 \\ 0.5 & 1 \end{bmatrix}$$

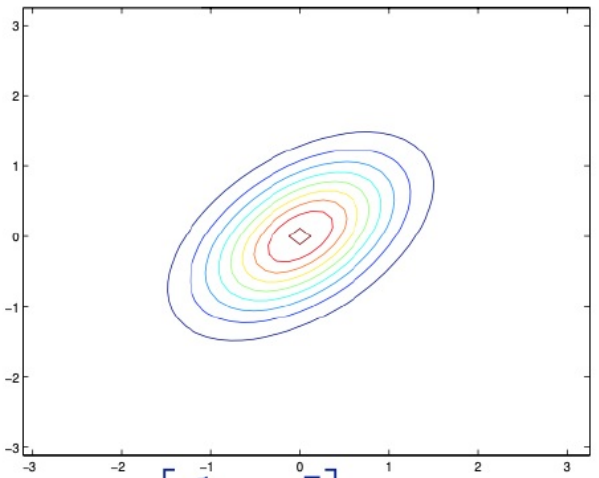


$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$

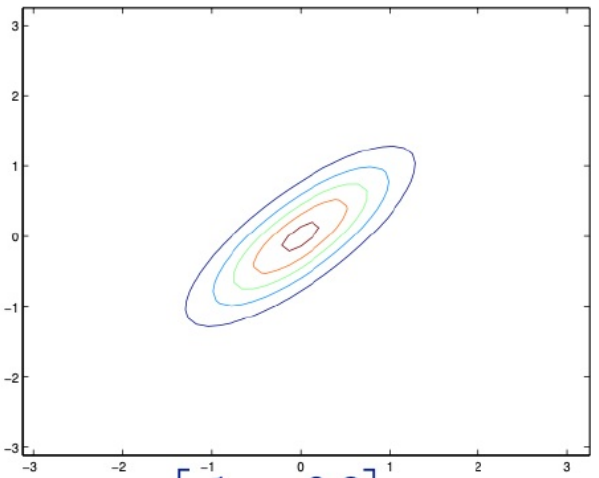
Covariance Matrix: Analysis



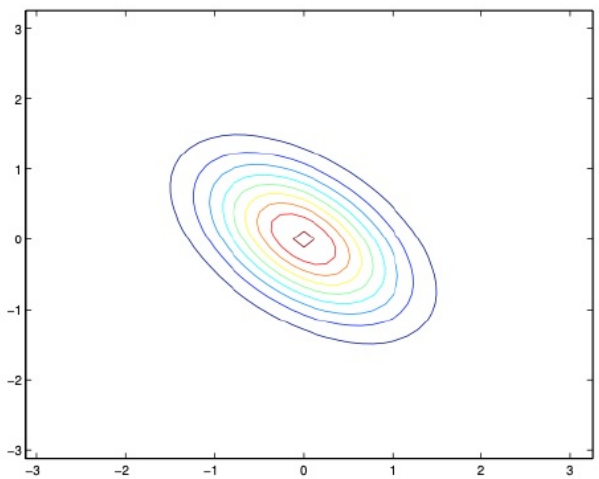
$$\Sigma = \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



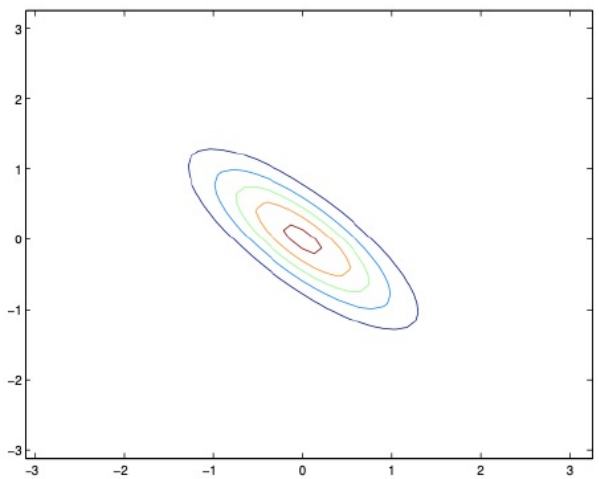
$$\Sigma = \begin{bmatrix} 1 & .5 \\ 0.5 & 1 \end{bmatrix}$$



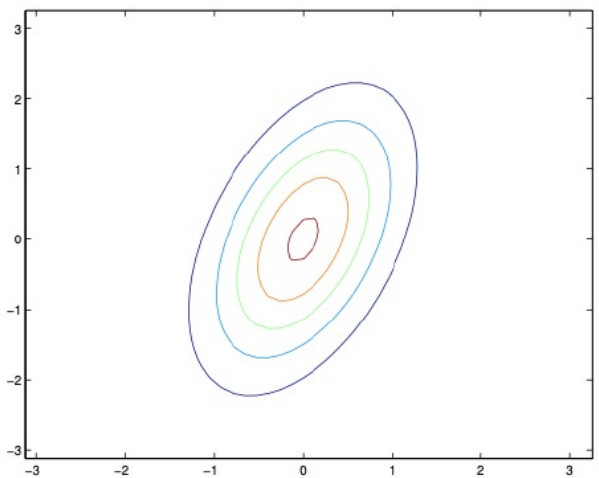
$$\Sigma = \begin{bmatrix} 1 & 0.8 \\ 0.8 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 1 & -0.5 \\ -0.5 & 1 \end{bmatrix}$$

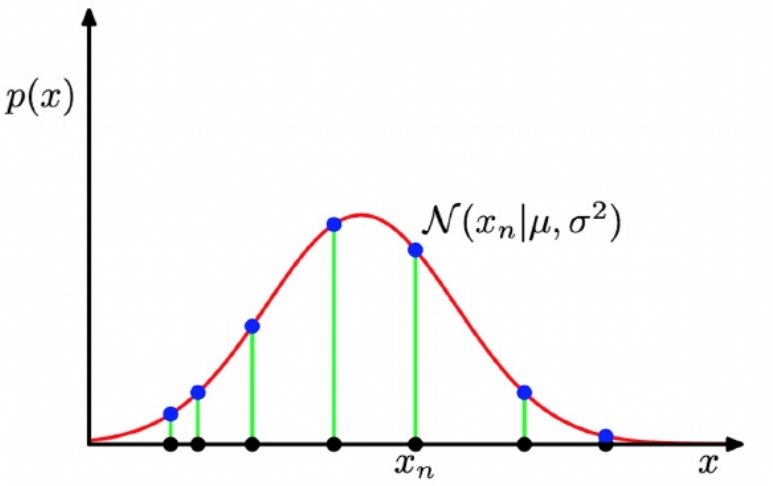


$$\Sigma = \begin{bmatrix} 1 & -0.8 \\ -0.8 & 1 \end{bmatrix}$$



$$\Sigma = \begin{bmatrix} 3 & .8 \\ 0.8 & 1 \end{bmatrix}$$

Maximum Likelihood Estimation



Principle: "We are more likely to observe data D if we are in a world where the appearance of this data is highly probable" - Tom Mitchell

- **Given**: a set of i.i.d. data $D = \{x_n : n = 1, 2, \ldots, N\}$
- Choose statistical model, e.g. $\mathcal{N}(x;\, \mu, \sigma^2)$. Then find the value of $\underline{\theta} = (\mu, \sigma)$ that makes D **most probable**

$$\hat{\theta}^{MLE} := \underset{\underline{\theta}}{\operatorname{argmax}}\, P(D \mid \underline{\theta}) = \underset{\underline{\theta}}{\operatorname{argmax}} \prod_{n=1}^{N} \mathcal{N}(x_n;\, \underline{\theta})$$

Maximum Likelihood Estimation

Note:

1. $P(D \mid \underline{\theta})$ is called (data) **likelihood function**
2. Maximizing $P(D \mid \underline{\theta})$ is equivalent to maximizing $\ell(\underline{\theta}) = \ln P(D \mid \underline{\theta})$
3. $\ell(\underline{\theta})$ is called **log likelihood function**

Maximum Likelihood Estimation

Theorem: If we have N i.i.d. samples $\mathbf{x}_n \sim \mathcal{N}(\underline{\mu}, \Sigma)$, then the MLE for the parameters is given by,

$$\hat{\underline{\mu}} = \frac{1}{N}\sum_{n=1}^{N}\mathbf{x}_n = \bar{\mathbf{x}}$$

$$\hat{\Sigma} = \frac{1}{N}\sum_{n=1}^{N}(\mathbf{x}_n - \bar{\mathbf{x}})(\mathbf{x}_n - \bar{\mathbf{x}})^T = \frac{1}{N}\left(\sum_{n=1}^{N}\mathbf{x}_n\mathbf{x}_n^T\right) - \bar{\mathbf{x}}\bar{\mathbf{x}}^T$$

For univariate case:

$$\hat{\mu} = \frac{1}{N}\sum_{n=1}^{N}x_n = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{N}\sum_{n=1}^{N}(x_n - \bar{x})^2 = \frac{1}{N}\left(\sum_{n=1}^{N}x_n^2\right) - \bar{x}^2 = S^2$$

(**Proof:** Book by Murphy, Page: 99-100)