Statistical Methods in Artificial Intelligence
CSE471 - Monsoon 2016 : Lecture 10

Avinash Sharma

CVIT, IIIT Hyderabad

Lecture Plan

- Revision from Previous Lecture
- Important Concepts
  - Expectation, Covariance Matrix , Entropy, KL Divergence
  - Linear Transformation on Random Variables
- The Normal Density
  - Univariate Normal Density
  - Multivariate Normal Density
  - Statistical Independence
  - Whitening Transform
  - Mahalanobis Distance Metric

Bayesian Decision Theory

- The joint probability density of finding a sample which is in category ω_j and has feature value x is given by:

$$p(\omega_j, x) = P(\omega_j|x)p(x)$$
$$= p(x|\omega_j)P(\omega_j)$$

- **Bayes Formula**

$$P(\omega_j|x) = \frac{p(\omega_j,x)}{p(x)} = \frac{p(x|\omega_j)\,P(\omega_j)}{p(x)}$$

$$posterior = \frac{likelihood \times prior}{evidence}$$

Bayesian Decision Theory

- Evidence act as normalizing term as:

$$p(x) = \sum_{i=1,2} p(\omega_i, x) = \sum_{i=1,2} p(x|\omega_i)P(\omega_i)$$

- Bayes Formula (Two Category)

$$P(\omega_j|x) = \frac{p(x|\omega_j)\,P(\omega_j)}{p(x)} = \frac{p(x|\omega_j)\,P(\omega_j)}{\sum_{i=1,2} p(x|\omega_i)P(\omega_i)}$$

- **Bayes Decision Rule**

  Decide ω_1 if $P(\omega_1|x) > P(\omega_2|x)$; Otherwise decide ω_2

Bayesian Decision Theory

- Bayes Risk

$$R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$$

- Two-Category Classification

$$R(\alpha_1|\mathbf{x}) = \lambda_{11} P(\omega_1|\mathbf{x}) + \lambda_{12} P(\omega_2|\mathbf{x})$$
$$R(\alpha_2|\mathbf{x}) = \lambda_{21} P(\omega_1|\mathbf{x}) + \lambda_{22} P(\omega_2|\mathbf{x})$$

  - Choose ω_1 if $R(\alpha_1|\mathbf{x}) < R(\alpha_2|\mathbf{x})$ or

$$(\lambda_{21} - \lambda_{11})\, P(\omega_1|\mathbf{x}) > (\lambda_{12} - \lambda_{22})\, P(\omega_2|\mathbf{x}) \text{ or}$$

$$(\lambda_{21} - \lambda_{11}) p(\mathbf{x}|\omega_1) P(\omega_1) > (\lambda_{12} - \lambda_{22}) p(\mathbf{x}|\omega_2) P(\omega_2) \text{ or}$$

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \boxed{\frac{(\lambda_{12} - \lambda_{22})}{(\lambda_{21} - \lambda_{11})} \frac{P(\omega_2)}{P(\omega_1)}} \qquad \theta_a$$

Minimum-Error-Rate Classification

- Let $\lambda_{ij} = \lambda(\alpha_i|\omega_j) = \begin{cases} 0 & i = j \\ 1 & i \neq j \end{cases}$

- $R(\alpha_i|\mathbf{x}) = \sum_{j=1}^{c} \lambda(\alpha_i|\omega_j) P(\omega_j|\mathbf{x})$
  $= \sum_{j \neq i} P(\omega_j|\mathbf{x}) = 1 - P(\omega_i|\mathbf{x})$

- If we choose ω_i corresponding to largest $P(\omega_i|\mathbf{x})$ then we minimize $R(\alpha_i|\mathbf{x})$

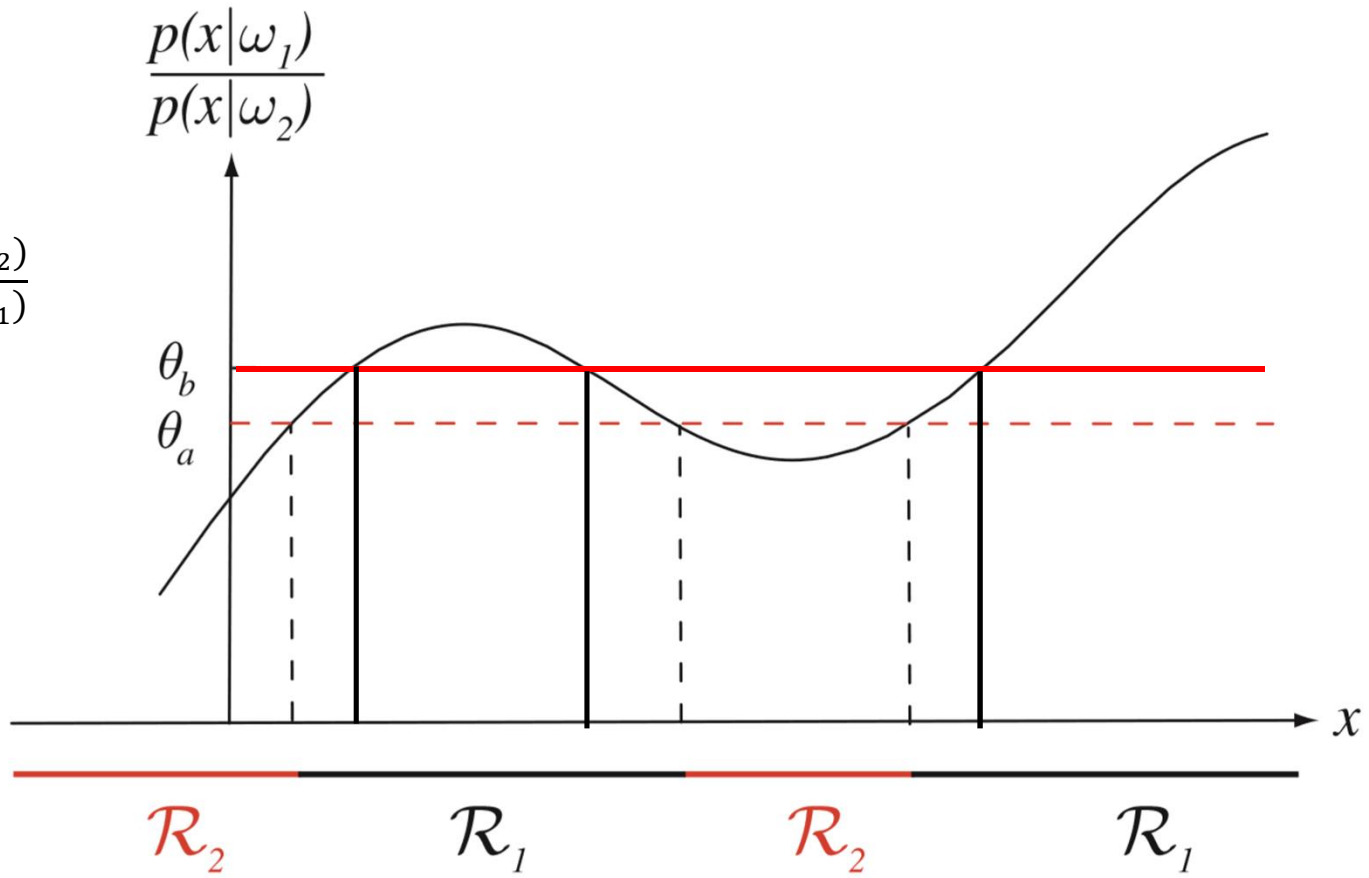
- Decision rule:

  Decide ω_i if $P(\omega_i|x) > P(\omega_j|x) \quad \forall\, j \neq i$

Minimum-Error-Rate Classification

- Let $\lambda_{12} > \lambda_{21}$(penalize miss-categorizing ω_2 as ω_1)

$$\frac{p(\mathbf{x}|\omega_1)}{p(\mathbf{x}|\omega_2)} > \frac{(\lambda_{12}-\lambda_{22})}{(\lambda_{21}-\lambda_{11})}\frac{P(\omega_2)}{P(\omega_1)}$$



Expectation

- The **expected value** of a Random Variable(RV) is long-run average value of repetitions of the experiment it represents.
- In case of discrete RV:

$$\mathcal{E}[f(x)] = \sum_{x \in D} f(x)P(x)$$

- In case of continuous RV:

$$\mathcal{E}[f(x)] = \int_{-\infty}^{\infty} f(x)p(x)\,dx$$

Covariance Matrix

- The covariance between two RV's measures the degree to which both are (linearly) related.

$$cov(X,Y) = \mathcal{E}[(X - \mathcal{E}[X])(Y - \mathcal{E}[Y])]$$
$$= \mathcal{E}[XY] - \mathcal{E}[X]\mathcal{E}[Y]$$

$$\Sigma = \begin{bmatrix} \mathrm{E}[(X_1 - \mu_1)(X_1 - \mu_1)] & \mathrm{E}[(X_1 - \mu_1)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_1 - \mu_1)(X_n - \mu_n)] \\ \mathrm{E}[(X_2 - \mu_2)(X_1 - \mu_1)] & \mathrm{E}[(X_2 - \mu_2)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_2 - \mu_2)(X_n - \mu_n)] \\ \vdots & \vdots & \ddots & \vdots \\ \mathrm{E}[(X_n - \mu_n)(X_1 - \mu_1)] & \mathrm{E}[(X_n - \mu_n)(X_2 - \mu_2)] & \cdots & \mathrm{E}[(X_n - \mu_n)(X_n - \mu_n)] \end{bmatrix}.$$

Entropy

- Entropy is a measure of the randomness or unpredictability.
- Entropy of a discrete RV X with distribution P and finite states K is defined as:

$$\mathbb{H}(P(X)) = -\sum_{k=1}^{K} P(X = k) \log(P(X = k))$$

- Entropy is **maximized** for X if $P(X = k) = \frac{1}{K}$ i.e., **uniform distribution** and **minimized** (i.e., 0) when pdf is a **direc-delta function** i.e., one at one k and zero else where.

- Entropy of a continuous RV x with distribution $p(x)$ is defined as

$$\mathbb{H}(p(x)) = -\int_{-\infty}^{\infty} p(x) \log\big(p(x)\big)\, dx$$

KL Divergence

- Dissimilarity of two probability distributions, P and Q, is measure by using the **Kullback-Leibler** divergence (*KL divergence*) or relative entropy. This is defined as:

$$\mathbb{KL}(P,Q) = \sum_{k=1}^{K} P(X=k)\log(P(X=k)/Q(X=k))$$

$$= \sum_{k=1}^{K} P(X=k)\log\big(P(X=k)\big) - \sum_{k=1}^{K} P(X=k)\log\big(Q(X=k)\big)$$

$$= -\mathbb{H}(P) + H(P,Q)$$

Cross Entropy

- KL divergence is the average number of extra bits needed to encode the data, due to the fact that we used distribution Q to encode the data instead of the true distribution P.

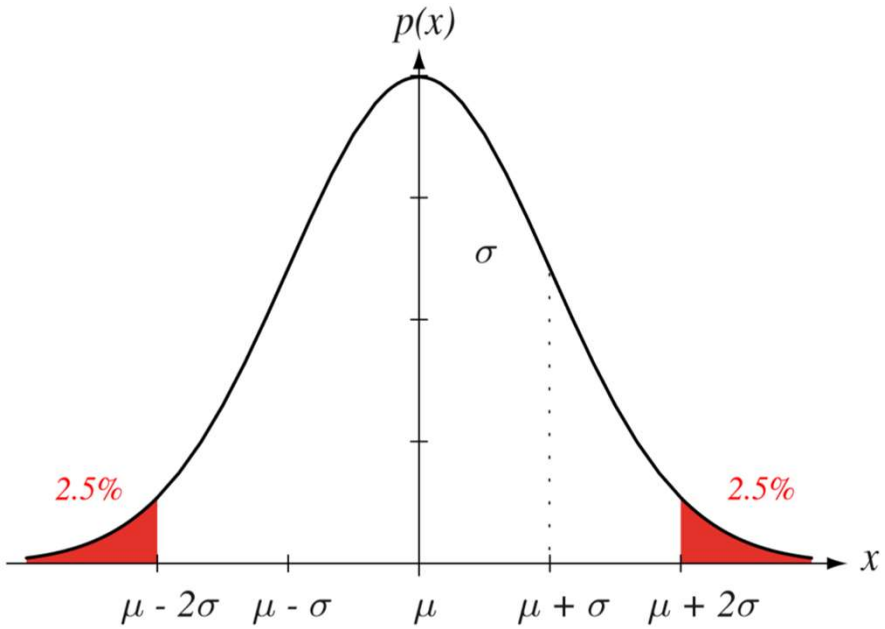
Linear Transformation on RVs

- Let $\mathbf{x}$ be an random variable with density function $p(\mathbf{x})$ with $\mathcal{E}[\mathbf{x}]=\mu$ and $cov[\mathbf{x}] = \Sigma$
- $\mathbf{y} = f(\mathbf{x}) = A\mathbf{x} + b$ for a linear $f(\cdot)$
- Mean: $\mathcal{E}[\mathbf{y}] = \mathcal{E}[A\mathbf{x} + b] = A\mu + b$
- Covariance: $cov[\mathbf{y}] = cov[A\mathbf{x} + b] = A\Sigma A^T$

Univariate Normal Density

- Gaussian or normal density function is most popular due to:
  - Simple model with only two parameters
  - Central limit theorem
  - Closer to real world sampling of data
  - Makes less number of assumptions (maximum entropy)
  - Analytical tractability of mathematical form

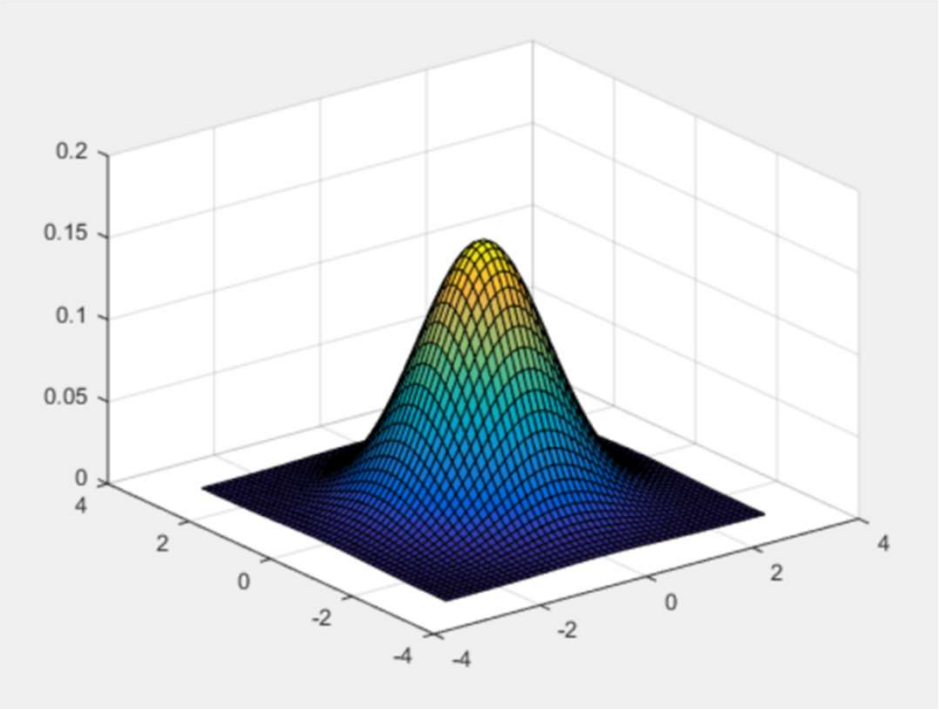
$$p(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left[-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2\right] = \mathcal{N}(\mu, \sigma)$$

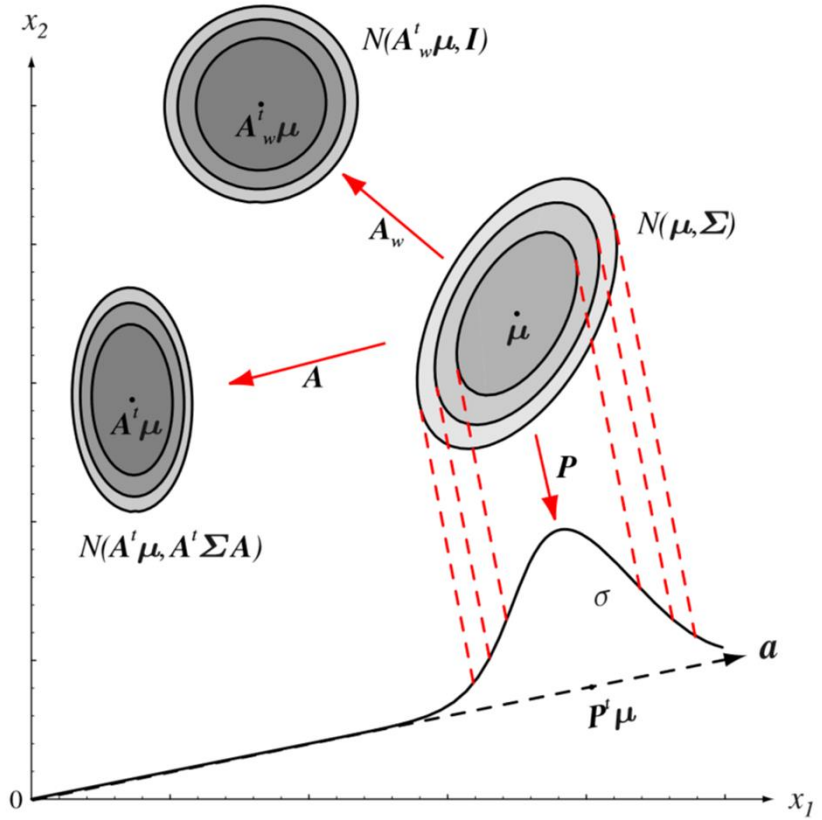


Multivariate Normal Density

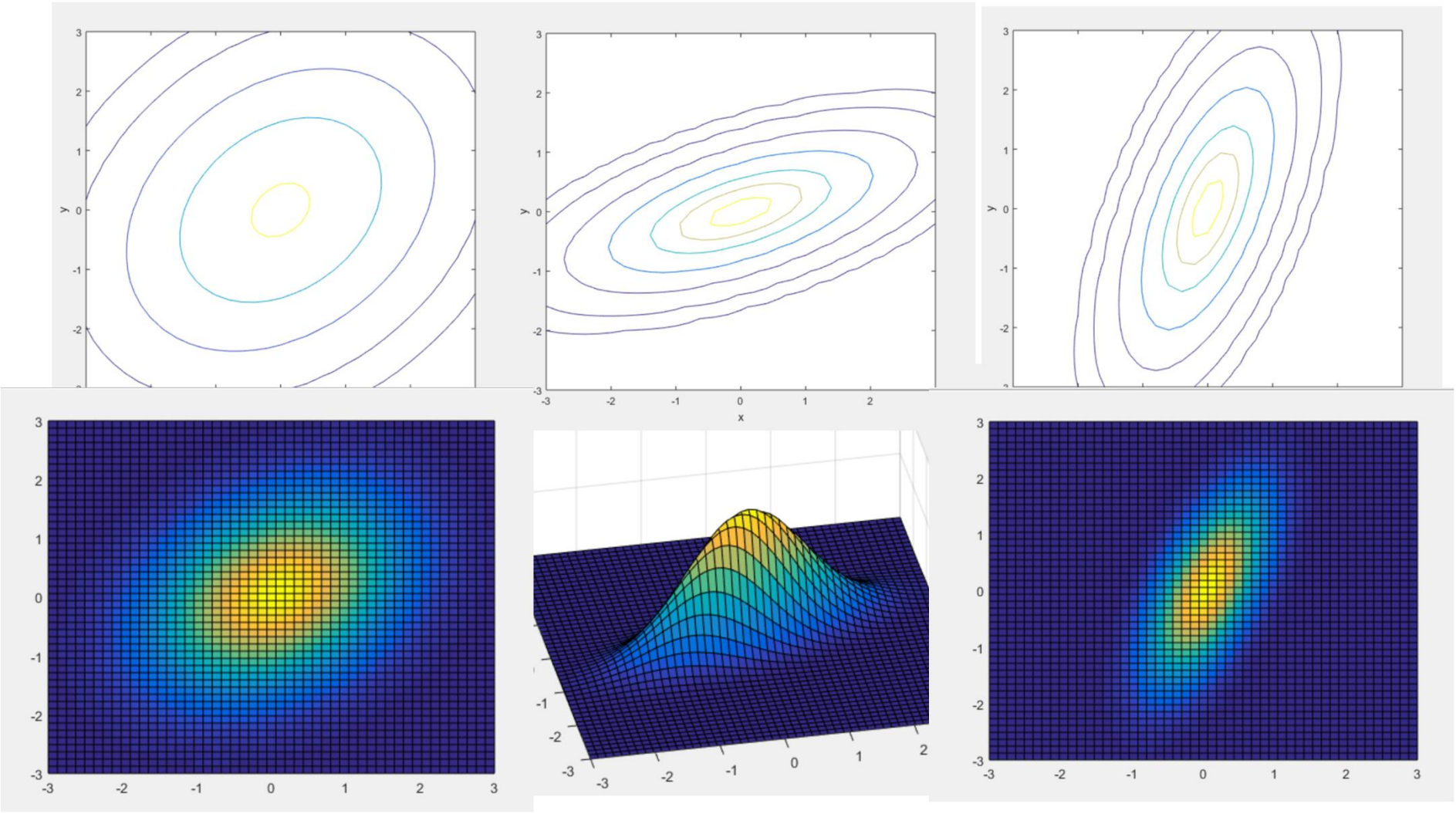
$$p(\mathbf{x}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^T\Sigma^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] = \mathcal{N}(\boldsymbol{\mu}, \Sigma)$$

- Σ is a symmetric and positive definite matrix so that $|\Sigma| > 0$.



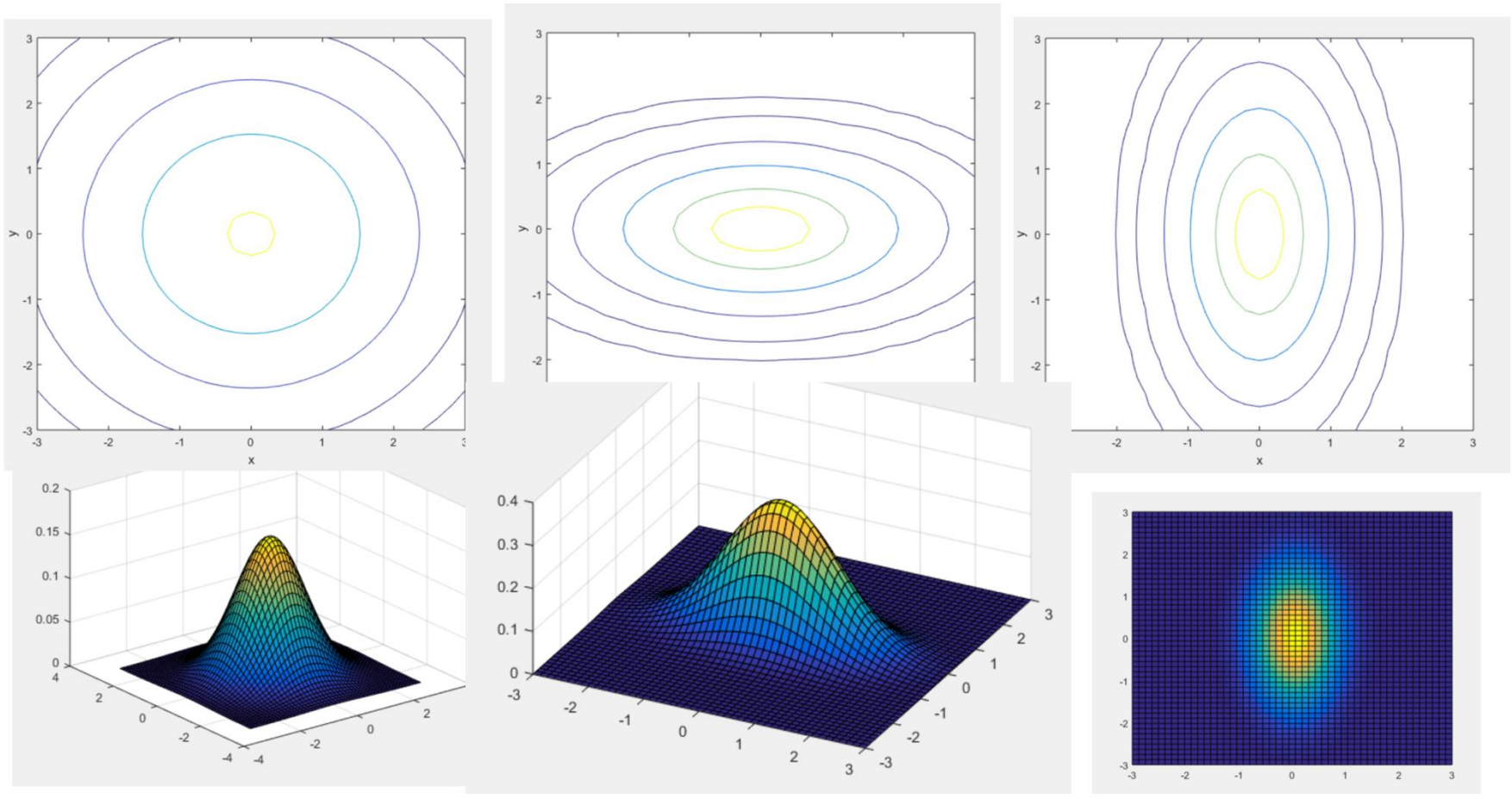


Multivariate Normal Density



Statistical Independence

- If all off diagonal elements of covariance matrix Σ are zero then the associate dimensions are statistically independent.



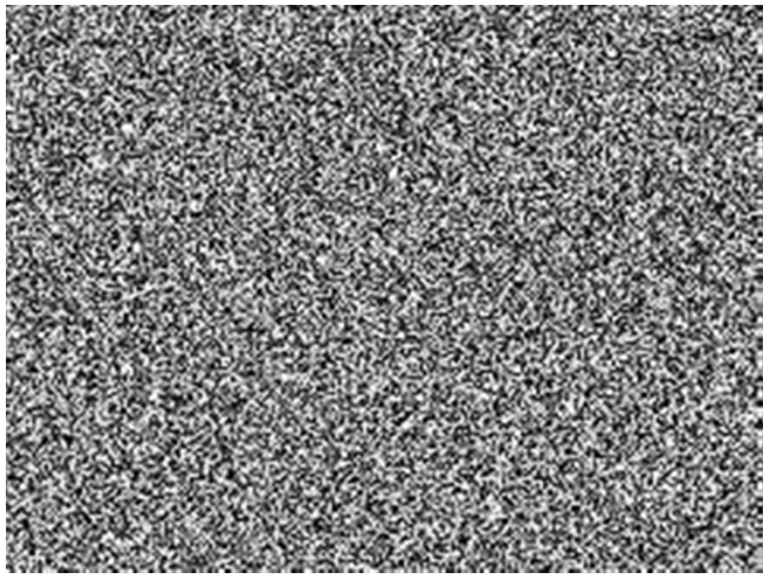
Whitening Transform

- It transforms an arbitrary set of variables having a known covariance matrix into a set of new variables whose covariance is the identity matrix .
- The transformation is called "whitening" because it changes the input vector into a white noise vector.

- Let $\Sigma = \Phi\Lambda\Phi^T$ be the SVD of Σ

  then $A_w = \Phi\Lambda^{-1/2}$

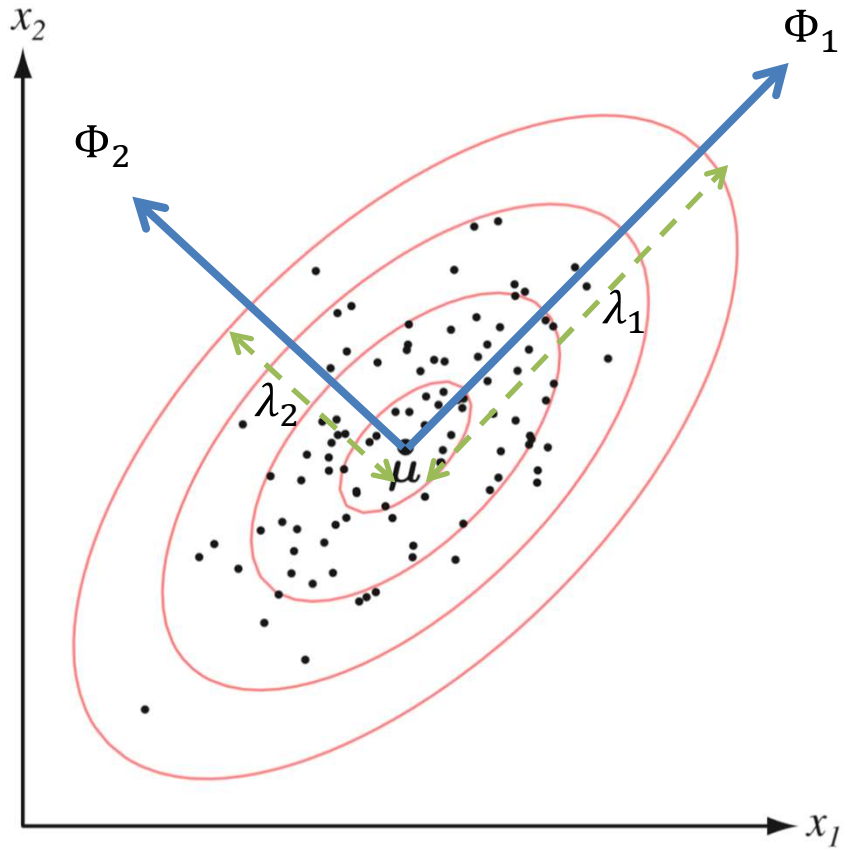
  so that, ${A_w}^T \Sigma A_w = I$



Mahalanobis Distance Metric

- $\Sigma = \Phi\Lambda\Phi^T$ where$\Phi = [\Phi_1, \dots, \Phi_d]$, $\Lambda = diag(\lambda_1, \dots, \lambda_d)$

- $r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$ is also known as Squared Mahalanobis Distance.
- $r^2 = (\mathbf{x} - \mathbf{y})^T \mathrm{S}^{-1} (\mathbf{x} - \mathbf{y})$ is generalized form for computing distance between two RVs.
- Mahalanobis distance becomes Euclidean when $\mathrm{S} = \mathrm{I}$.



Mahalanobis Distance Metric

- $\Sigma^{-1} = U\Lambda^{-1}U^T = \sum_{i=1}^{d} \frac{1}{\lambda_i} \mathrm{u}_i \mathrm{u}_i{}^T$
- $r^2 = (\mathbf{x} - \boldsymbol{\mu})^T \Sigma^{-1} (\mathbf{x} - \boldsymbol{\mu})$

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

- Equation of an ellipse in 2d is: $\frac{z_1{}^2}{\lambda_1} + \frac{z_2{}^2}{\lambda_2} = 1$

Mid Semester Exam # 1 Syllabus

- What all is covered in the class, tutorials till 3rd September and relevant material in the public domain.
- Chapter 1
- Chapter 5 (5.1-5.7, 5.8.1, 5.9)
- Chapter 6 (6.1-6.3,6.5,6.8,6.10*)
- Chapter 4 (4.2, 4.3,4.4)
- Chapter 2 (2.1-2.3)