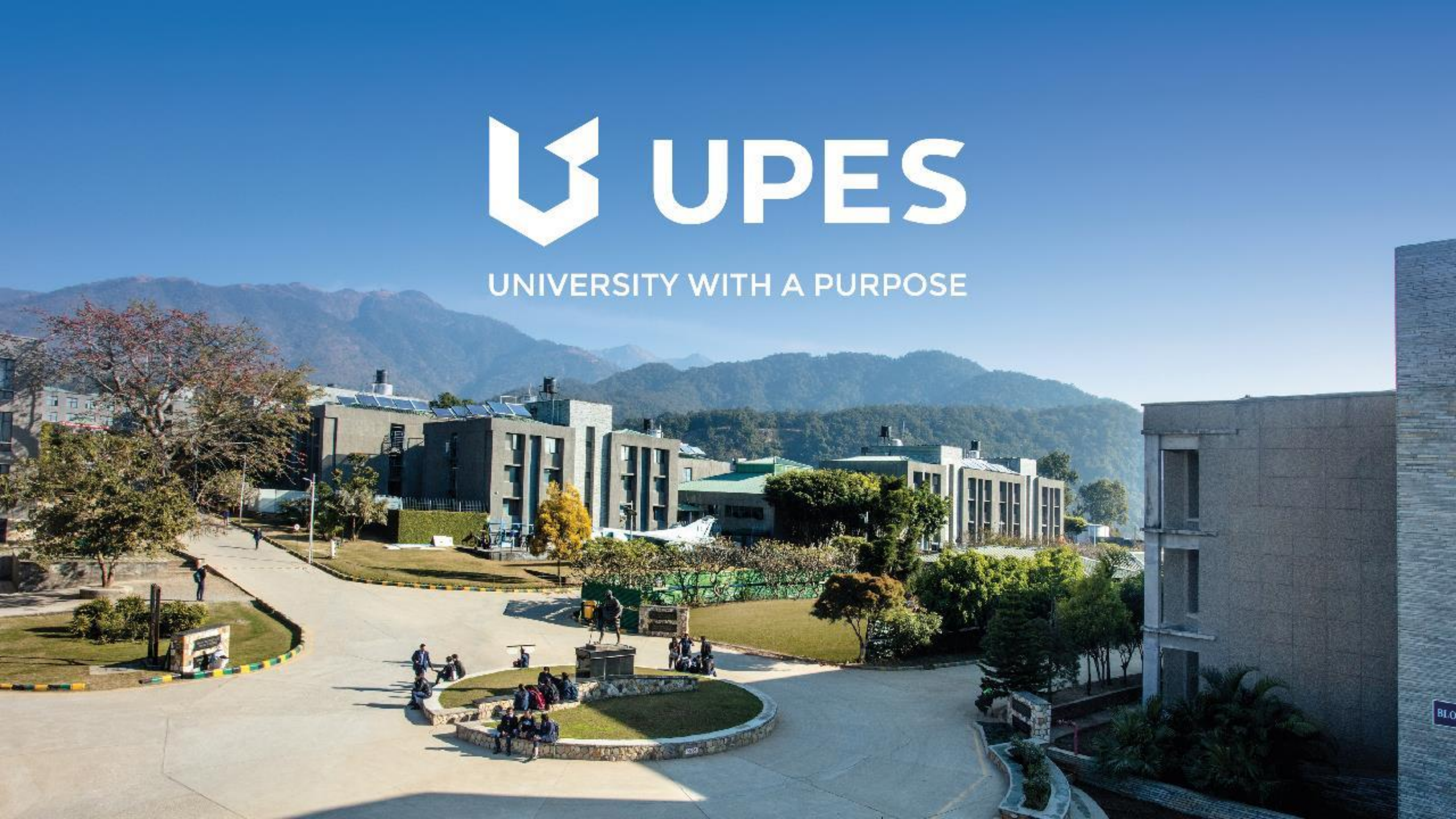
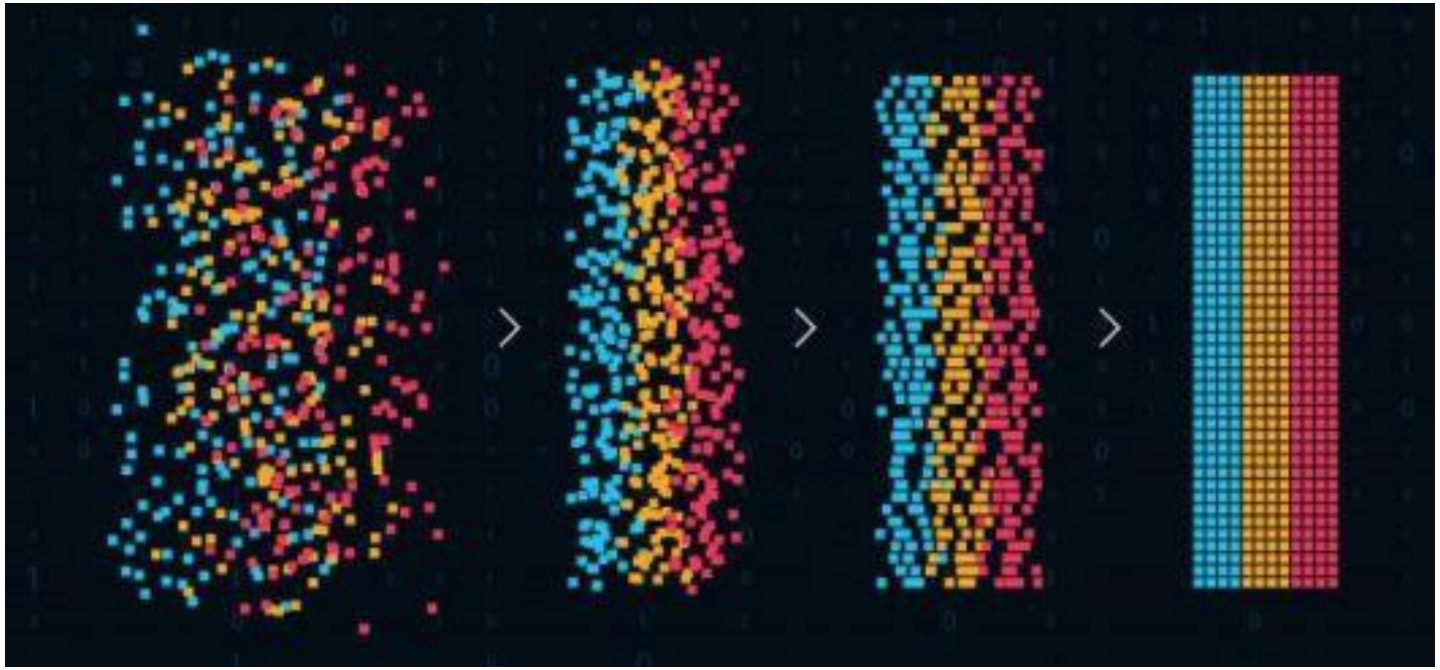

UPES
UNIVERSITY WITH A PURPOSE

Pattern and Anomaly Detection



Source: Edureka

B. Tech., CSE + AI/ML

Dr Gopal Singh Phartiyal

16/09/2021

Recap: Binary Variables

Discrete binary random variable

Two states (True and False)

Popular Distributions

- *Bernoulli*
- *Binomial (Frequenist)*
- *Beta (Bayesian)*

Multinomial Variables

Variable with K states

- 1-of-K coding scheme: $\mathbf{x} = (0, 0, 1, 0, 0, 0)^{\mathrm{T}}$
- Probability distribution $\boxed{p(\mathbf{x}|\boldsymbol{\mu}) = \prod_{k=1}^{K} \mu_k^{x_k}}$ Generalized version of Bernoulli
- Under constraints $\forall k : \mu_k \geqslant 0$ and $\sum_{k=1}^{K} \mu_k = 1$
- Expectation and variance

$$\mathbb{E}[\mathbf{x}|\boldsymbol{\mu}] = \sum_{\mathbf{x}} p(\mathbf{x}|\boldsymbol{\mu})\mathbf{x} = (\mu_1, \ldots, \mu_K)^{\mathrm{T}} = \boldsymbol{\mu}$$

Multinomial Variables: Parameter Estimation

- Given $\mathcal{D} = \{\mathbf{x}_1, \ldots, \mathbf{x}_N\}$

$$p(\mathcal{D}|\boldsymbol{\mu}) = \prod_{n=1}^{N} \prod_{k=1}^{K} \mu_k^{x_{nk}} = \prod_{k=1}^{K} \mu_k^{(\sum_n x_{nk})} = \prod_{k=1}^{K} \mu_k^{m_k}$$

- Ensure $\sum_k \mu_k = 1$, use a Lagrange multiplier, ¸.

$$\sum_{k=1}^{K} m_k \ln \mu_k + \lambda \left(\sum_{k=1}^{K} \mu_k - 1 \right)$$

$$\mu_k = -m_k / \lambda \qquad \mu_k^{\mathrm{ML}} = \frac{m_k}{N}$$

Multinomial Variables: Multinomial Distribution

Joint distribution of $(m_1, m_2, \ldots, m_K$

$$
\begin{aligned}
\operatorname{Mult}(m_1, m_2, \ldots, m_K | \boldsymbol{\mu}, N) &= \binom{N}{m_1 m_2 \ldots m_K} \prod_{k=1}^{K} \mu_k^{m_k} \\
\mathbb{E}[m_k] &= N\mu_k \\
\operatorname{var}[m_k] &= N\mu_k(1-\mu_k) \\
\operatorname{cov}[m_j m_k] &= -N\mu_j\mu_k
\end{aligned}
$$

$$\binom{N}{m_1 m_2 \ldots m_K} = \frac{N!}{m_1! m_2! \ldots m_K!}.$$

$$\sum_{k=1}^{K} m_k = N.$$

Multinomial Variables: Dirichlet Distribution

Bayesian treatment

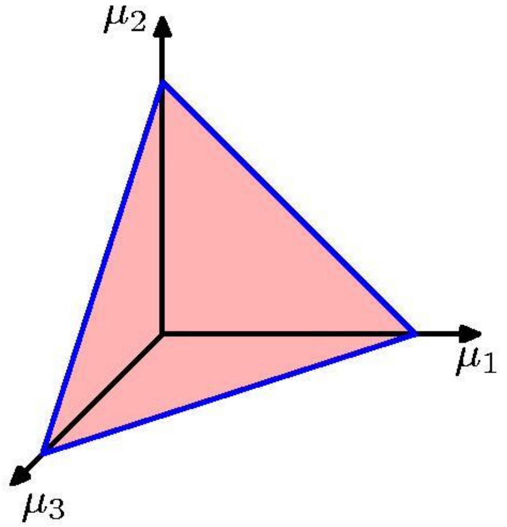
$$\mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}) = \frac{\Gamma(\alpha_0)}{\Gamma(\alpha_1)\cdots\Gamma(\alpha_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k - 1} \qquad \alpha_0 = \sum_{k=1}^{K} \alpha_k$$

A family of priors for $\{\mu_k\}$

Parameters of the distribution

$$\alpha_1, \ldots, \alpha_K$$

Conjugate prior for the multinomial distribution.



Simplex (bounded linear mainfold) of dimensionality K-1.

Multinomial Variables: Dirichlet Distribution

Multiplying likelihood with prior

Posterior =

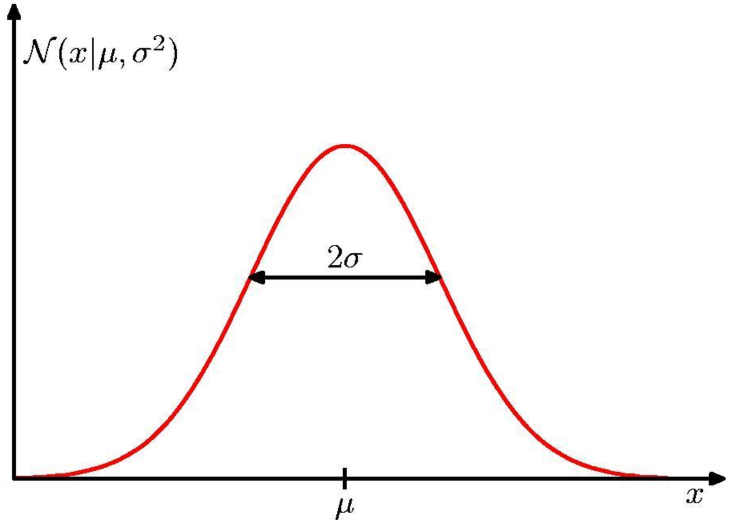
$$p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) \propto p(\mathcal{D}|\boldsymbol{\mu})p(\boldsymbol{\mu}|\boldsymbol{\alpha}) \propto \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1}$$

$$\begin{aligned} p(\boldsymbol{\mu}|\mathcal{D},\boldsymbol{\alpha}) &= \mathrm{Dir}(\boldsymbol{\mu}|\boldsymbol{\alpha}+\mathbf{m}) \\ &= \frac{\Gamma(\alpha_0+N)}{\Gamma(\alpha_1+m_1)\cdots\Gamma(\alpha_K+m_K)} \prod_{k=1}^{K} \mu_k^{\alpha_k+m_k-1} \end{aligned}$$

Two-state (Binary) variables: Either via binomial and Beta or
With Multinomial and Dirichlet (1 of 2 scenario).

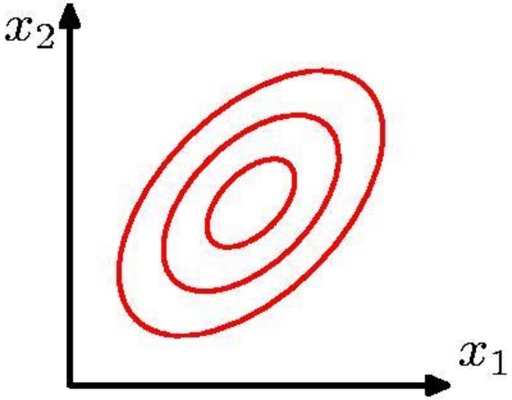
Continuous Variables: Gaussian Distribution

- Widely used model for the distribution of continuous variables
- Also known as Normal distribution



- For single variable

$$\mathcal{N}(x|\mu, \sigma^2) = \frac{1}{(2\pi\sigma^2)^{1/2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$



- Multivariate Gaussian distribution

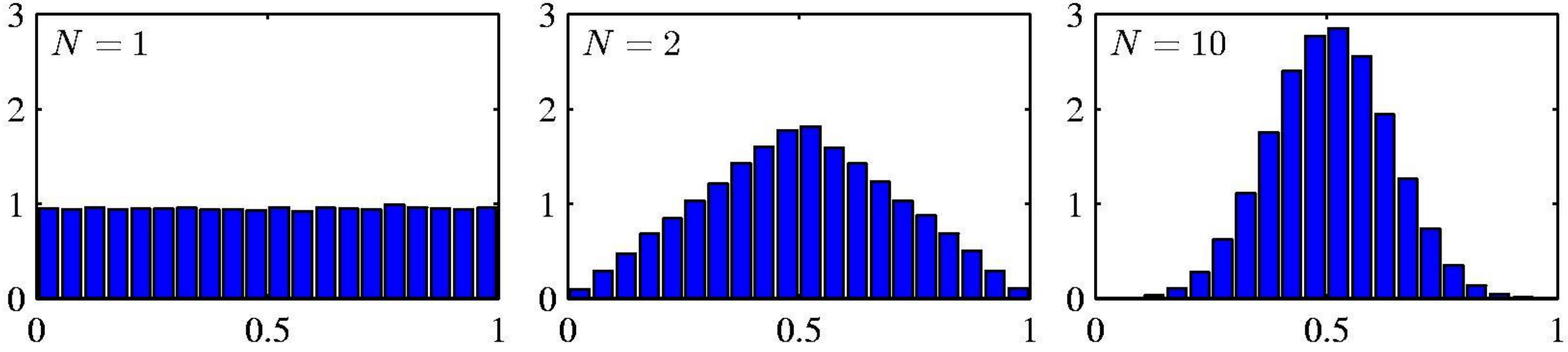
$$\mathcal{N}(\mathbf{x}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \frac{1}{(2\pi)^{D/2}} \frac{1}{|\boldsymbol{\Sigma}|^{1/2}} \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}$$

Continuous Variables: Gaussian Distribution

- For singe or multiple continuous variable, the distribution that maximizes the entropy is the **Gaussian.**

- The distribution of a random variable (which itself is a sum of multiple random variables) tends to be Gaussian as the number of number of variables summing up increases.

Example: N uniform [0,1] random variables



Continuous Variables: Gaussian Distribution

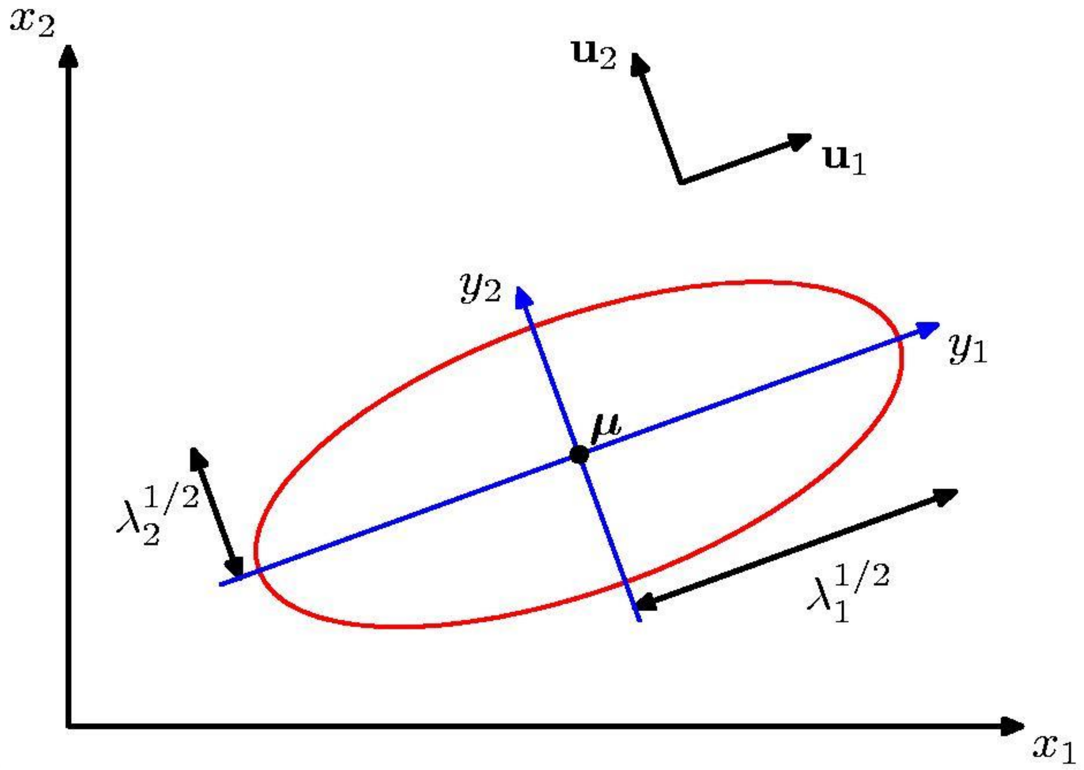
- Gaussian distribution has important analytical properties
  - Geometrical form interpretation: Δ̇ is ***Mahalonobis distance***

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

$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

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

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



Non-Parametric

- Parametric distribution models are restricted to specific forms, which may not always be suitable; for example, consider modelling a multimodal distribution with a single, unimodal model.

- Nonparametric approaches make few assumptions about the overall shape of the distribution being modelled.

- We will focus on frequentist treatment however Bayesian treatment is also interesting.

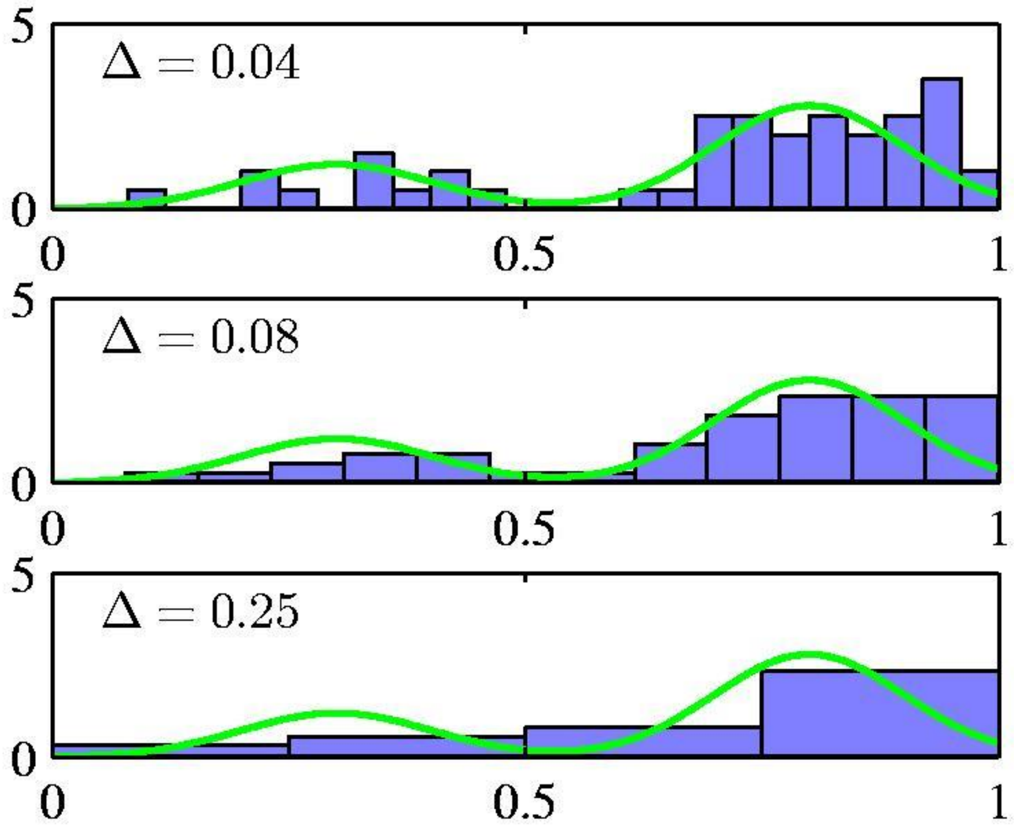
Histogram methods(histogram density models)

Single continuous variable x

Histogram methods partition the data space into distinct bins with widths $\dot{\Delta}_i$ and count the number of observations, n_i, in each bin.

$$p_i = \frac{n_i}{N\Delta_i}$$

- Often, the same width is used for all bins, $\dot{\Delta}_i = \dot{\Delta}$.
- ¢ acts as a smoothing parameter.



•In a D-dimensional space, using M bins in each dimen-sion will require M^D bins!

Next time: Non-parametric methods

Thank You

UPES

UNIVERSITY WITH A PURPOSE