BLG 454E Learning From Data

FALL 2022-2023

Multivariate Methods

(Slides are Prepared by Assoc. Prof. Yusuf Yaslan

& Assist. Prof. Ayşe Tosun)

Lecture Notes from Alpaydın 2010 Introduction to Machine Learning 2e © The MIT Press (V1.0) AND
O. Velksler's Lecture Notes from CS 434s/541a Pattern Recognition, Uni of Western Ontario

Univariate Normal Density

- So far, we have dealt with univariate x (dimension of 1)
- $P(x) = \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2}\left(\frac{x-\mu}{\sigma}\right)^2)$
- $\mu_{MLE} = m = \frac{1}{n}\sum_{i=1}^{N} x_i$
- ${\sigma^2}_{MLE} = s^2 = \frac{1}{n}\sum_{i=1}^{N}(x_i - m)^2$

Multivariate Normal Density

- What if we have several features $x_1, x_2, x_3, \dots, x_d$
  - Each normally distributed
  - Different variances
  - Different means
  - May be dependent or independent of each other
- $P(x) = \frac{1}{2\pi^{d/2}|\Sigma|^{1/2}} \exp(-\frac{1}{2}(x-\mu)^T \Sigma^{-1} (x-\mu))$

X = $\begin{bmatrix} x_1^1 & \cdots & x_d^1 \\ \vdots & \ddots & \vdots \\ x_1^N & \cdots & x_d^N \end{bmatrix}$

Mahalonobis distance between x and mean
(elliptical curve between x's)

Multivariate parameters

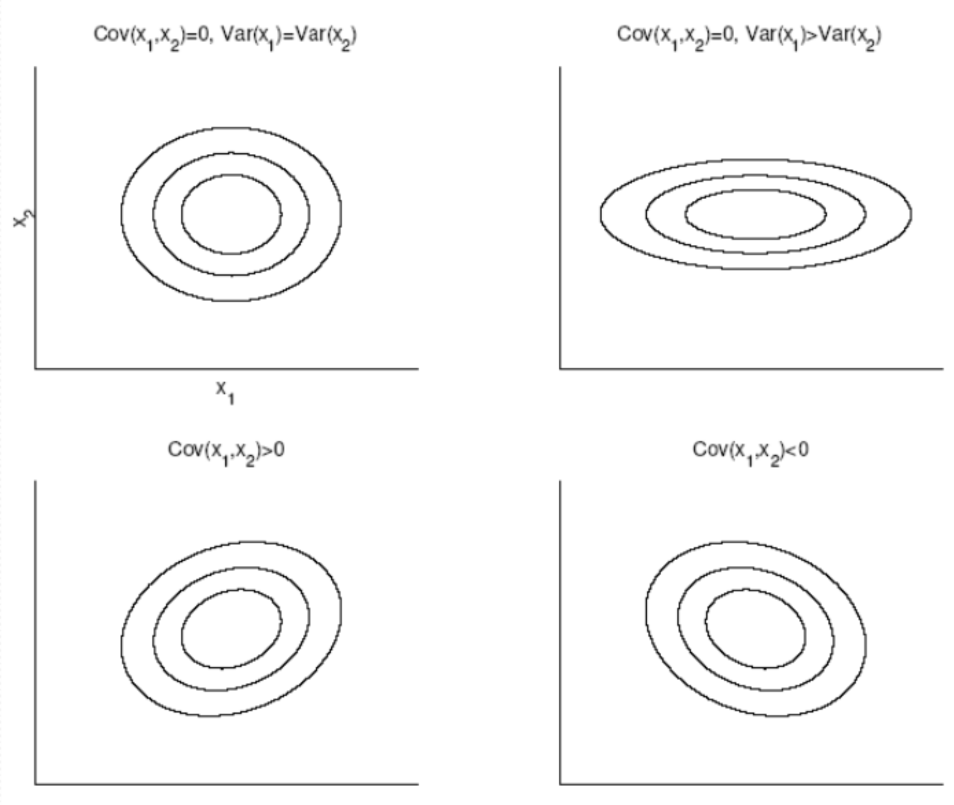
- $E[X] = \mu = [\mu_1, \mu_2, \dots, \mu_d]^T$
- Covariance = $\sigma_{ij} = Cov(X_i, X_j)$
- Correlation =$\mathrm{Corr}(X_i, X_j) = \rho_{ij} = \frac{\sigma_{ij}}{\sigma_i \sigma_j}$
- Covariance matrix $\Sigma = E[(X-\mu)(X-\mu)^T]$

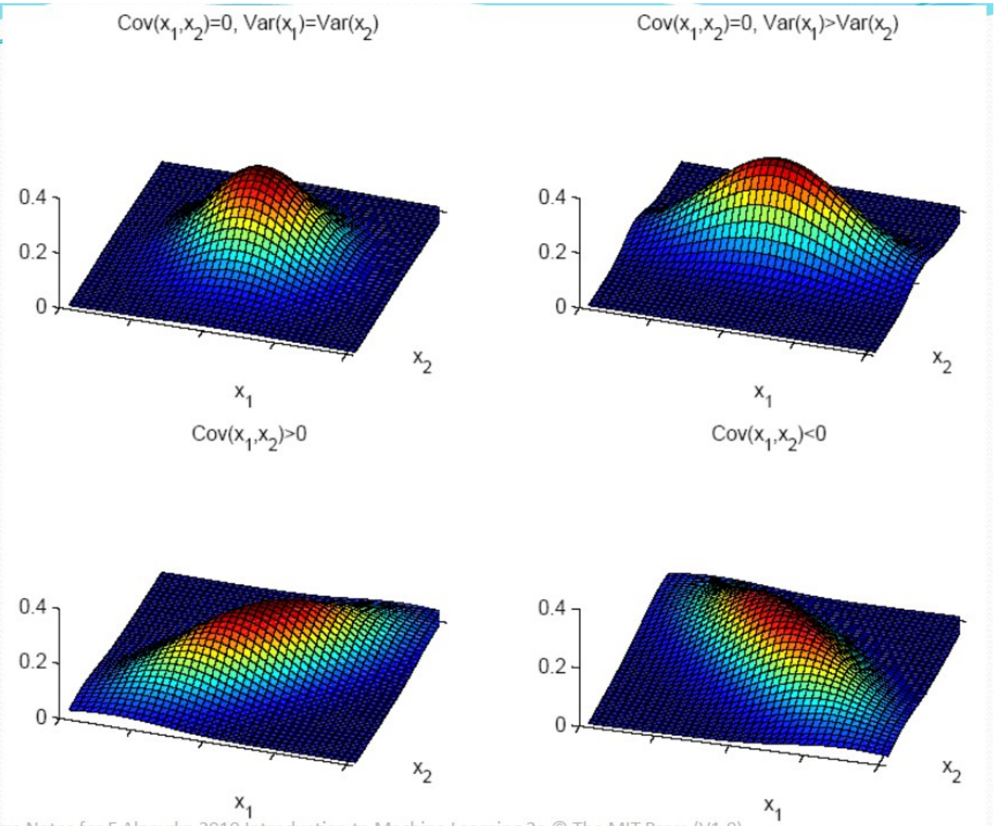
$$= \begin{bmatrix} \sigma_1^2 & \sigma_{12} & \cdots & \sigma_{1d} \\ \vdots & & \ddots & \vdots \\ \sigma_{d1} & \sigma_{d2} & \cdots & \sigma_d^2 \end{bmatrix}$$

Multivariate parameter estimation

- Sample mean m: $m_i = \frac{\sum_{t=1}^{N} x_i^t}{N}, i = 1, \ldots, \mathrm{d}$
- Covariance matrix S: $s_{ij} = \frac{\sum_{t=1}^{N}(x_i^t - m_i)(x_j^t - m_j)}{N}$
- Correlation matrix R: $r_{ij} = \frac{s_{ij}}{s_i s_j}$
- If features x_i, x_j are
  - INDEPENDENT, then σ_{ij}=0 diagonals are non-zero. $\begin{bmatrix} \sigma_1^2 & 0 & \cdots & 0 \\ \vdots & & \ddots & \vdots \\ 0 & 0 & \cdots & \sigma_d^2 \end{bmatrix}$
  - POSITIVE correlation, $\sigma_{ij} > 0$
  - NEGATIVE correlation, $\sigma_{ij} < 0$

Σ in Bivariate Normal

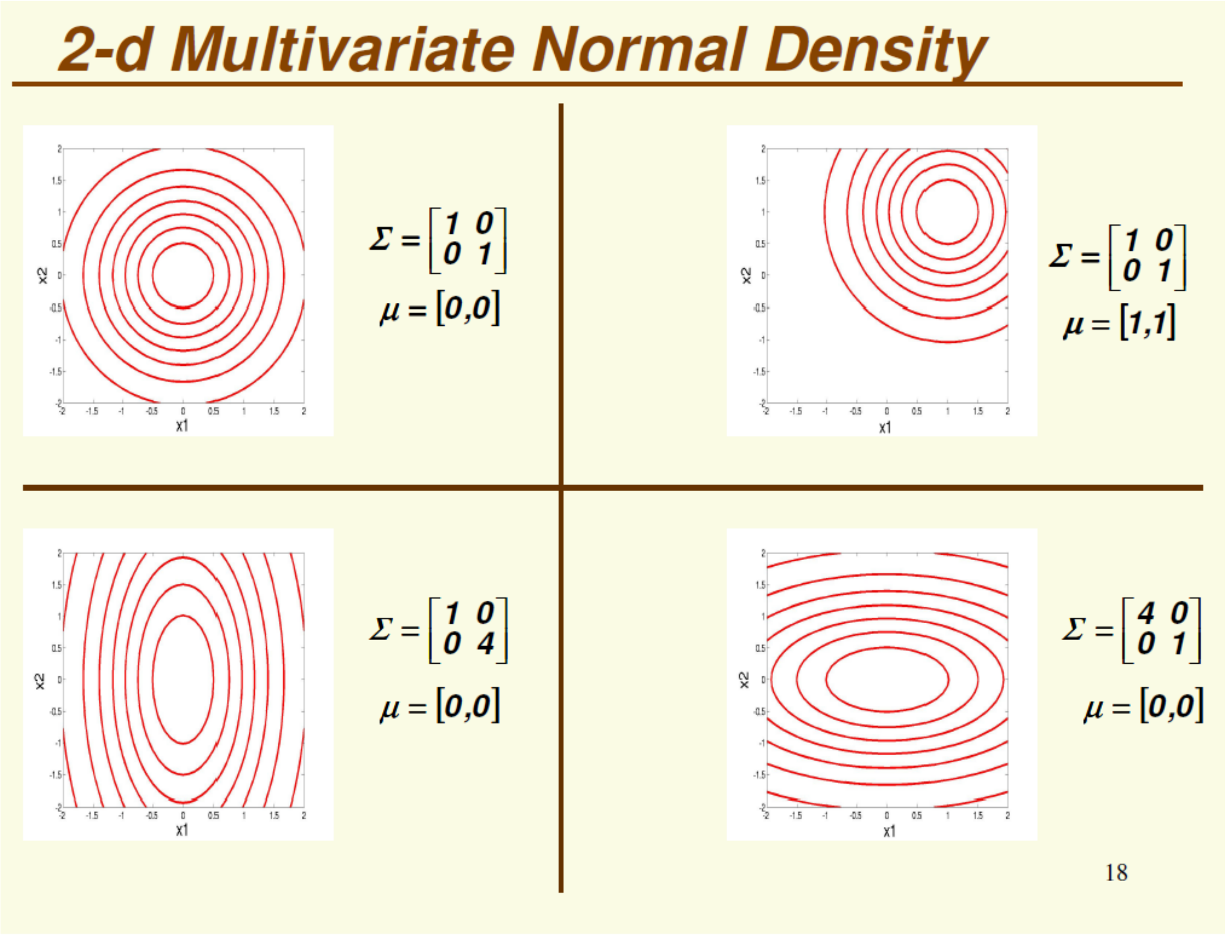


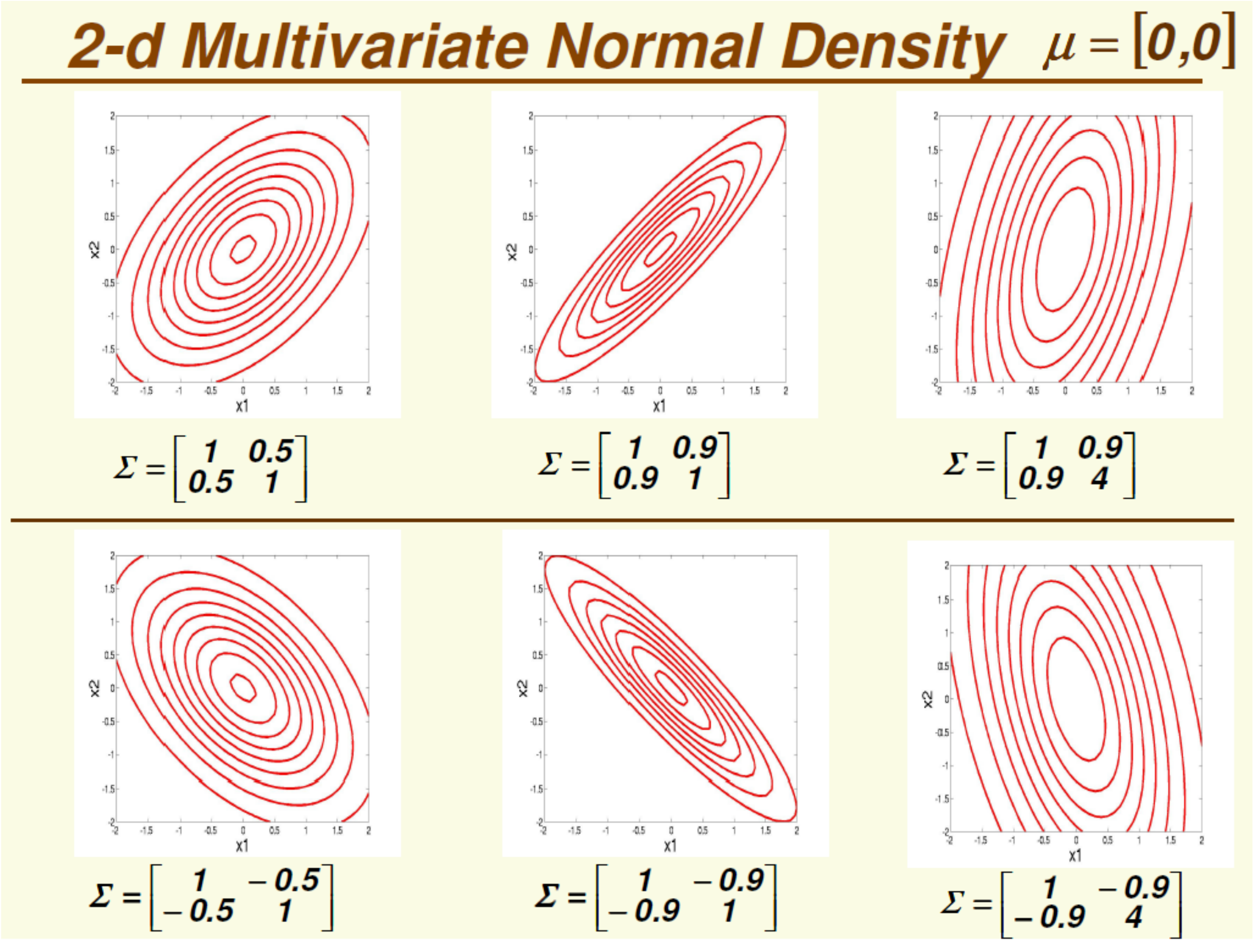


If Σ is diagonal

- Features are independent and
- $P(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma_i} \exp(-\frac{1}{2{\sigma_i}^2}(x_i - \mu_i)^2)$
  - Euclidean distance (circular view between x's)
- If variances are also equal
- $P(x) = \prod_{i=1}^{d} \frac{1}{\sqrt{2\pi}\sigma} \exp(-\frac{1}{2\sigma^2}(x_i - \mu_i)^2)$

Σ and μ relations on topological maps of Gaussian surface





Figures from CS 434s/541a Pattern Recognition, Uni of Western Ontario

Discriminant functions for classification

- Classifier can be viewed as ***m*** *discriminant* functions and the classification is based on selecting the largest discriminant:
  - $g_i(x) = P(c_i|X) = P(X|c_i)\mathrm{P}(c_i)/\mathrm{P(x)}$
- For normal density, it is more convenient to work on logarithms
  - $g_i(x) = logP(X|c_i) + \mathrm{logP}(c_i)$
  - $g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma_i^{-1}(x-\mu_i) - \frac{d}{2}log2\pi - \frac{1}{2}log|\Sigma_i| + \mathrm{logP}(c_i)$

Case $\Sigma_i = \sigma^2 \mathrm{I}$

- Features are independent with different means and equal variances
- $\sigma^2\mathrm{I} = \sigma^2 \begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$
- $g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) - \frac{d}{2}log2\pi - \frac{1}{2}log|\Sigma \quad| + \log\mathrm{P}(c_i)$

$$= -\frac{1}{2}(x-\mu_i)^T(\frac{1}{\sigma^2}\mathrm{I})(x-\mu_i) + \log\mathrm{P}(c_i)$$

$$= -\frac{1}{2\sigma^2}(x-\mu_i)^T(x-\mu_i) + \log\mathrm{P}(c_i)$$

$$= -\frac{1}{2\sigma^2}(x^Tx - x^T\mu_i - {\mu_i}^Tx + \mu_i^T\mu_i)$$

$$= -\frac{1}{2\sigma^2}\left(-2{\mu_i}^Tx + \mu_i^T\mu_i\right) + \log\mathrm{P}(c_i)$$

- Discriminant function is linear wrt x
- $g_i(x) = w_i^Tx + w_{i0}$

Case Σ_i = $\sigma^2\mathrm{I}$

- Decision boundaries $g_i(x) = \ g_j(x)$ are linear
  - when x has a dimension of 2, lines
  - When x has a dimension of 3, plane
  - Larger than 3, hyperplanes

Example

- $\mu_1 = \begin{bmatrix} 1 \\ 2 \end{bmatrix}, \mu_2 = \begin{bmatrix} 4 \\ 6 \end{bmatrix}, \mu_3 = \begin{bmatrix} -2 \\ 4 \end{bmatrix}, \Sigma_1 = \Sigma_2 = \Sigma_3 = \begin{bmatrix} 3 & 0 \\ 0 & 3 \end{bmatrix}$
- $p(c_1) = p(c_2) = \frac{1}{4}, p(c_3) = \frac{1}{2}$
- $g_i(x) = \frac{\mu_i^T}{\sigma^2} x + (-\frac{\mu_i^T \mu_i}{\sigma^2} + \text{logP}(c_i))$
- First form the discriminants for $g_1(x), g_2(x), g_3(x)$
- Then solve $g_i(x)$= $g_j(x)$

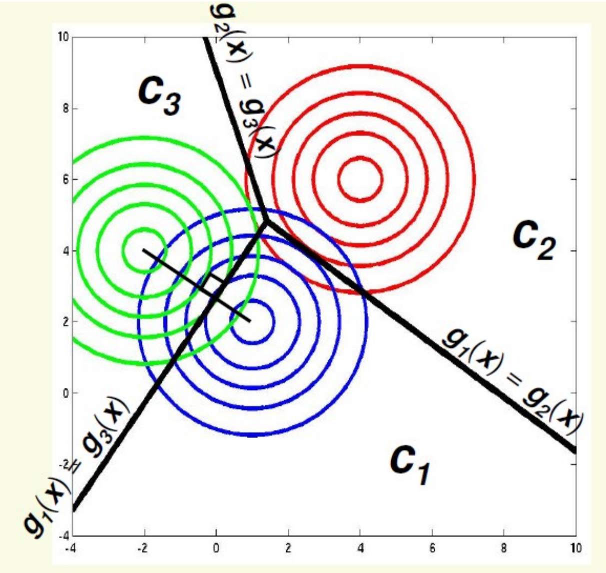


Figure from CS 434s/541a Pattern Recognition, Uni of Western Ontario

Case $\Sigma_i = \Sigma$

- Features are not necessarily independent
- Covariance matrices are equal but arbitrary
- $g_i(x) = -\frac{1}{2}(x-\mu_i)^T\Sigma^{-1}(x-\mu_i) + \log \mathrm{P}(c_i)$

  $= -\frac{1}{2}\left(x^T\Sigma^{-1}x - 2\mu_i^T\Sigma^{-1}x + \mu_i^T\Sigma^{-1}\mu_i\right) + \log \mathrm{P}(c_i)$

  $= \mu_i^T\Sigma^{-1}x + (\log \mathrm{P}(c_i) - \frac{\mu_i^T\Sigma^{-1}\mu_i}{2})$
- This is also linear
- $w_i^T x + w_{i0}$

General case

- $g_i(x) = -\frac{1}{2}(x-\mu_i)^T \Sigma_i^{-1}(x-\mu_i) - \frac{1}{2} log|\Sigma_i| + \log \mathrm{P}(c_i)$

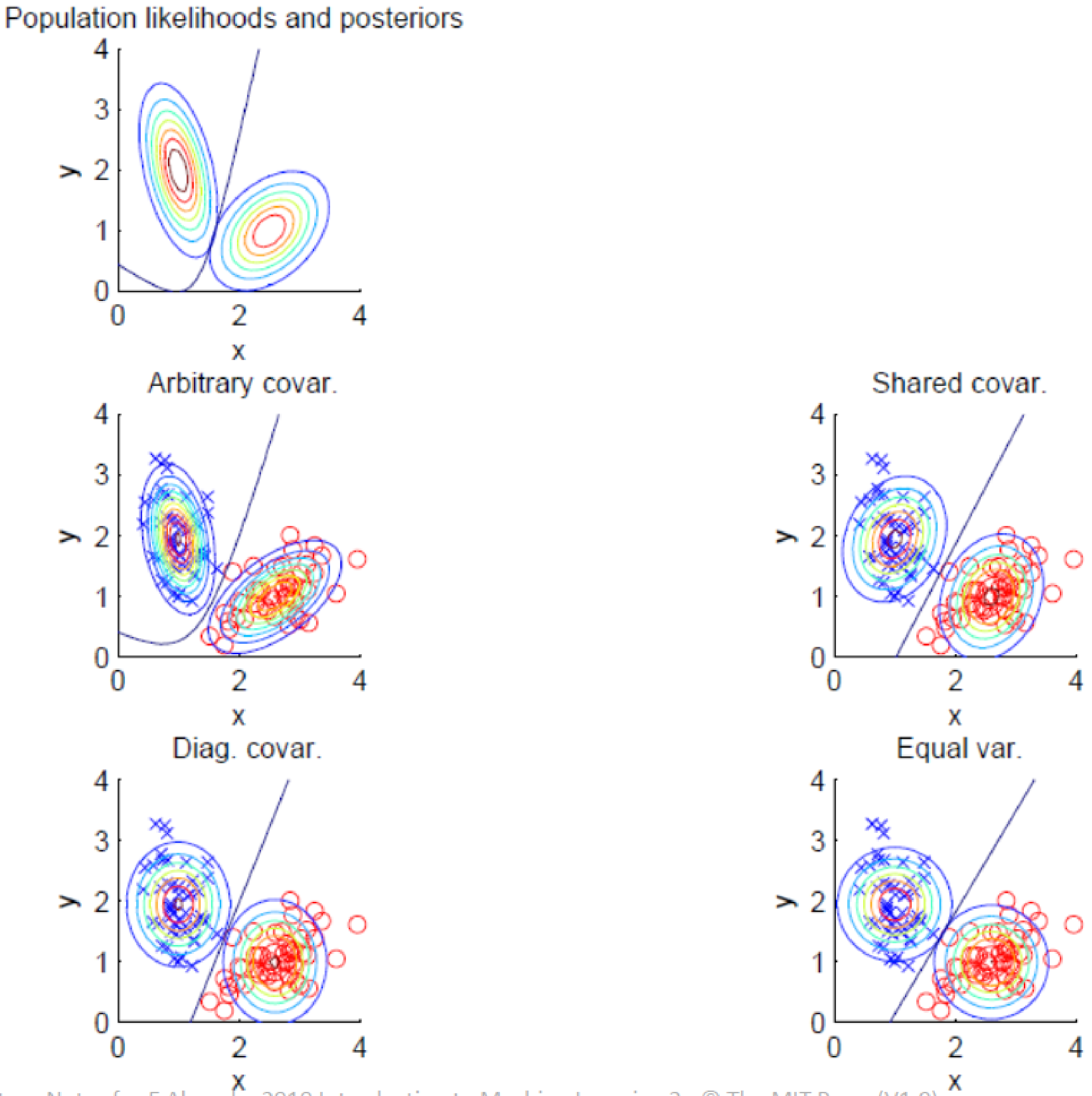
  $= -\frac{1}{2}\ (x^T \Sigma_i^{-1} x - 2\mu_i^T \Sigma_i^{-1} x + \mu_i^T \Sigma_i^{-1} \mu_i) - \frac{1}{2} log|\Sigma_i| + \log \mathrm{P}(c_i)$

  $= \ x^T \mathrm{Wx} + \ w^T x + \ w_{i0}$
- Discriminant is quadratic.
- Decision boundaries are ellipses and parabolloids

Model complexity - Bias - Variance

- As we increase complexity, bias decreases and variance increases
- Assume simple models to control variance (regularization)

Assumption	*Covariance matrix*	*No of parameters*
Shared, Hyperspheric	$\mathbf{S}_i=\mathbf{S}=s^2\mathbf{I}$	1
Shared, Axis-aligned	$\mathbf{S}_i=\mathbf{S}$, with $s_{ij}=0$	d
Shared, Hyperellipsoidal	$\mathbf{S}_i=\mathbf{S}$	$d(d+1)/2$
Different, Hyperellipsoidal	$\mathbf{S}_i$	$K\ d(d+1)/2$


Population likelihoods and posteriors
Arbitrary covar.
Shared covar.
Diag. covar.
Equal var.
x
y
0
2
4
1
3