Last updated: Sept 20, 2012

MULTIVARIATE NORMAL DISTRIBUTION

J. Elder

CSE 4404/5327 Introduction to Machine Learning and Pattern Recognition

Linear Algebra

- Tutorial this Wed 3:00 – 4:30 in Bethune 228
- Linear Algebra Reviews:
  - Kolter, Z., avail at http://cs229.stanford.edu/section/cs229-linalg.pdf
  - Prince, Appendix C (up to and including C.7.1)
  - Bishop, Appendix C
  - Roweis, S., avail at http://www.cs.nyu.edu/~roweis/notes/matrixid.pdf

Credits

- Some of these slides were sourced and/or modified from:
    - Christopher Bishop, Microsoft UK
    - Simon Prince, University College London
    - Sergios Theodoridis, University of Athens & Konstantinos Koutroumbas, National Observatory of Athens

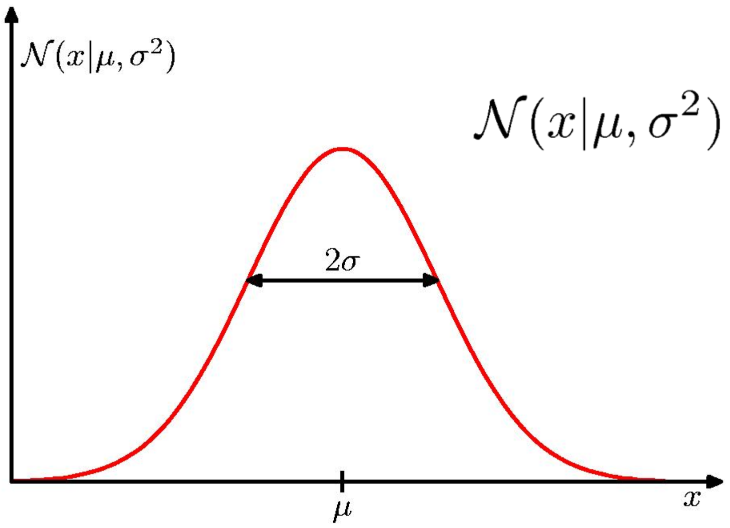
The Multivariate Normal Distribution: Topics

1. The Multivariate Normal Distribution
2. Decision Boundaries in Higher Dimensions
3. Parameter Estimation
    1. Maximum Likelihood Parameter Estimation
    2. Bayesian Parameter Estimation

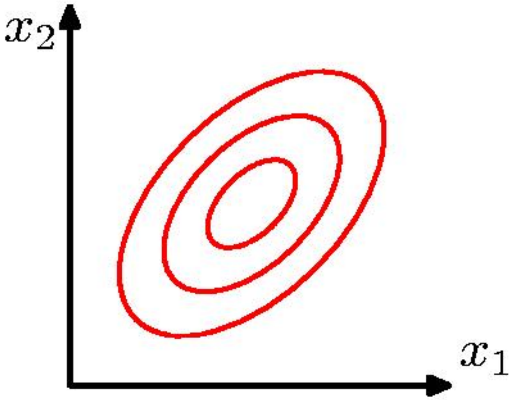
The Multivariate Normal Distribution: Topics

1. **The Multivariate Normal Distribution**
2. Decision Boundaries in Higher Dimensions
3. Parameter Estimation
    1. Maximum Likelihood Parameter Estimation
    2. Bayesian Parameter Estimation

The Multivariate Gaussian



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



MATLAB Statistics Toolbox Function:
mvnpdf(x,mu,sigma)

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

Orthonormal Form

$\Delta^2 = (\mathbf{x} - \boldsymbol{\mu})^{\mathrm{T}} \boldsymbol{\Sigma}^{-1} (\mathbf{x} - \boldsymbol{\mu})$ where $\Delta \equiv$ Mahalanobis distance from μ to x

MATLAB Statistics Toolbox Function:
mahal(x,y)

Let $A \in \mathbb{R}^{D\times D}$. λ is an eigenvalue and u is an eigenvector of A if $Au = \lambda u$.

MATLAB Functions:
[V, D]= eig(A)
[V, D]= eigs(A, k)

Let u_i and λ_i represent the i^{th} eigenvector/eigenvalue pair of Σ: $\Sigma u_i = \lambda_i u_i$

See Linear Algebra Review Resources on Moodle site for a review of eigenvectors.

Orthonormal Form

Since it is used in a quadratic form, we can assume that Σ^{-1} is symmetric.
This means that all of its eigenvalues and eigenvectors are real.

We are also implicitly assuming that Σ, and hence Σ^{-1}, are invertible (of full rank).

Thus Σ can be represented in orthonormal form: $\Sigma = U\Lambda U^t$,
where the columns of U are the eigenvectors u_i of Σ, and
Λ is the diagonal matrix with entries $\Lambda_{ii} = \lambda_i$ equal to the corresponding eigenvalues of Σ.

Thus the Mahalanobis distance Δ^2 can be represented as:

$$\Delta^2 = \left(x-\mu\right)^t \Sigma^{-1}\left(x-\mu\right) = \left(x-\mu\right)^t U\Lambda^{-1}U^t\left(x-\mu\right).$$

Let $y = U^t\left(x-\mu\right)$. Then we have,

$$\Delta^2 = y^t\Lambda^{-1}y = \sum_{ij} y_i\Lambda_{ij}^{-1}y_j = \sum_i \lambda_i^{-1}y_i^2,$$

where $y_i = u_i^t\left(x-\mu\right)$.

Geometry of the Multivariate Gaussian

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

Δ = Mahalanobis distance from μ to x

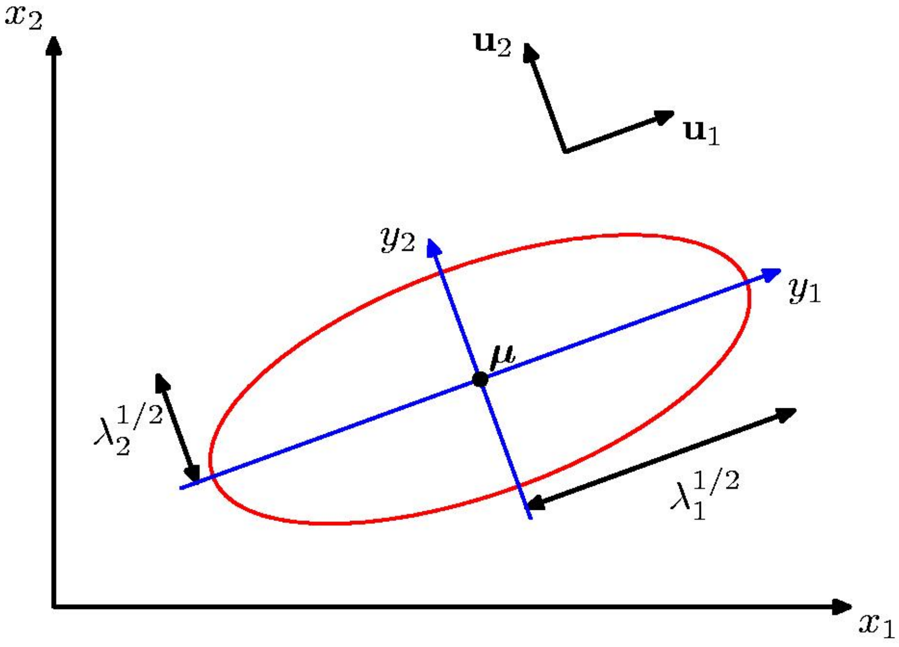
$$\boldsymbol{\Sigma}^{-1} = \sum_{i=1}^{D} \frac{1}{\lambda_i} \mathbf{u}_i \mathbf{u}_i^{\mathrm{T}}$$

where $(\mathbf{u}_i, \lambda_i)$ are the ith eigenvector and eigenvalue of Σ.

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

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

or $\mathbf{y} = \mathbf{U}(\mathbf{x} - \mu)$



Moments of the Multivariate Gaussian

$$\begin{aligned}\mathbb{E}[\mathbf{x}] &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right\}\mathbf{x}\,\mathrm{d}\mathbf{x} \\ &= \frac{1}{(2\pi)^{D/2}}\frac{1}{|\boldsymbol{\Sigma}|^{1/2}}\int \exp\left\{-\frac{1}{2}\mathbf{z}^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}\mathbf{z}\right\}(\mathbf{z}+\boldsymbol{\mu})\,\mathrm{d}\mathbf{z}\end{aligned}$$

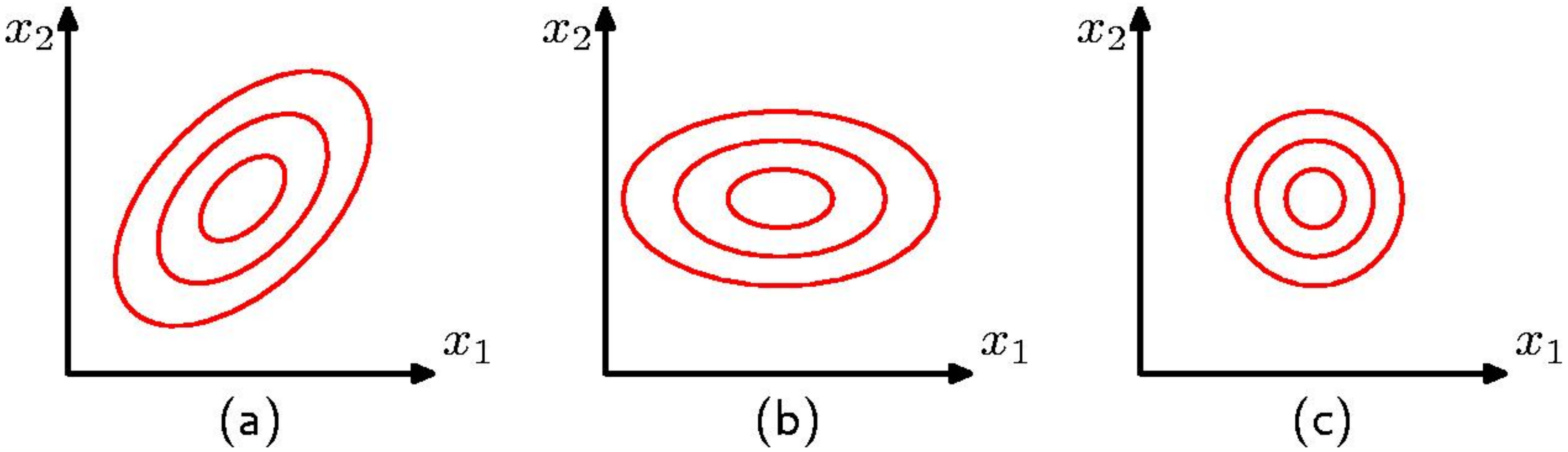
thanks to anti-symmetry of z

$$\mathbb{E}[\mathbf{x}] = \boldsymbol{\mu}$$

Moments of the Multivariate Gaussian

$$\mathbb{E}[\mathbf{x}\mathbf{x}^{\mathrm{T}}] = \boldsymbol{\mu}\boldsymbol{\mu}^{\mathrm{T}} + \boldsymbol{\Sigma}$$

$$\operatorname{cov}[\mathbf{x}] = \mathbb{E}\left[(\mathbf{x} - \mathbb{E}[\mathbf{x}])(\mathbf{x} - \mathbb{E}[\mathbf{x}])^{\mathrm{T}}\right] = \boldsymbol{\Sigma}$$

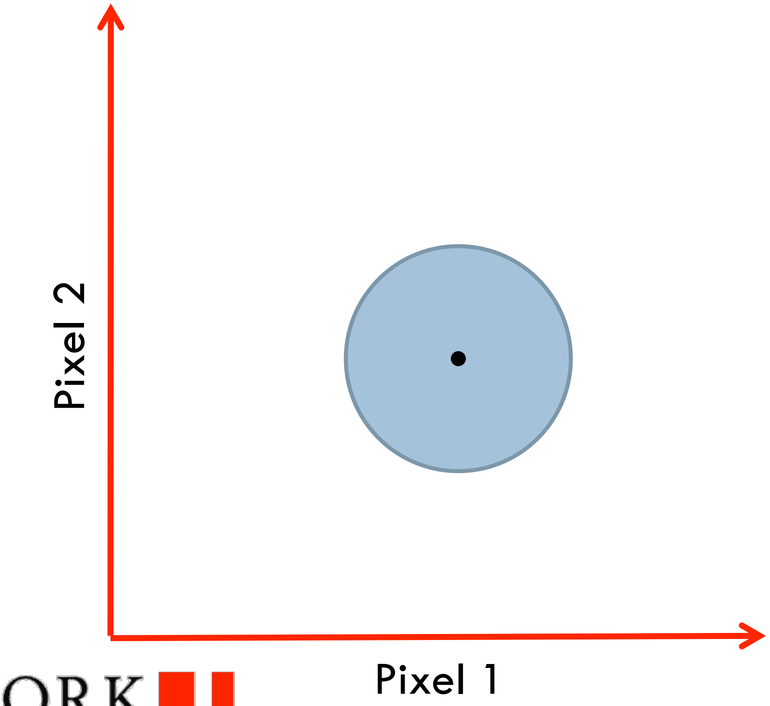


5.1 Application: Face Detection

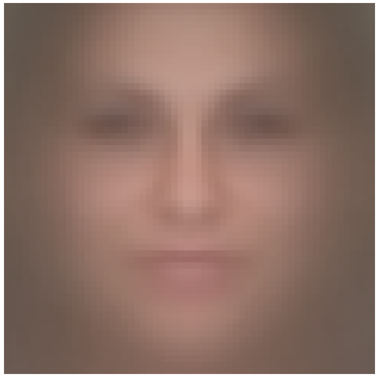
Model # 1: Gaussian, uniform covariance

$$Pr(\mathbf{x}|\text{face}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-0.5(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

Fit model using maximum likelihood criterion



m face



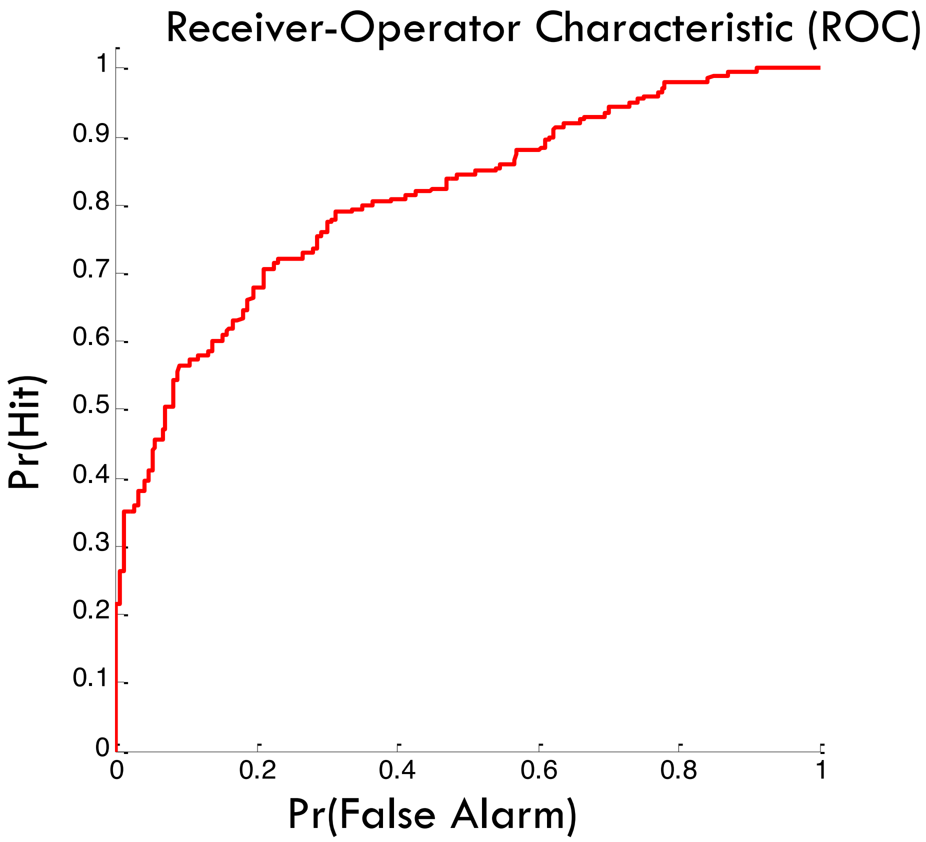
m non-face

Face ‘template’

s Face	s non-face
59.1	69.1

Model 1 Results

Results based on 200 cropped faces and 200 non-faces from the same database.

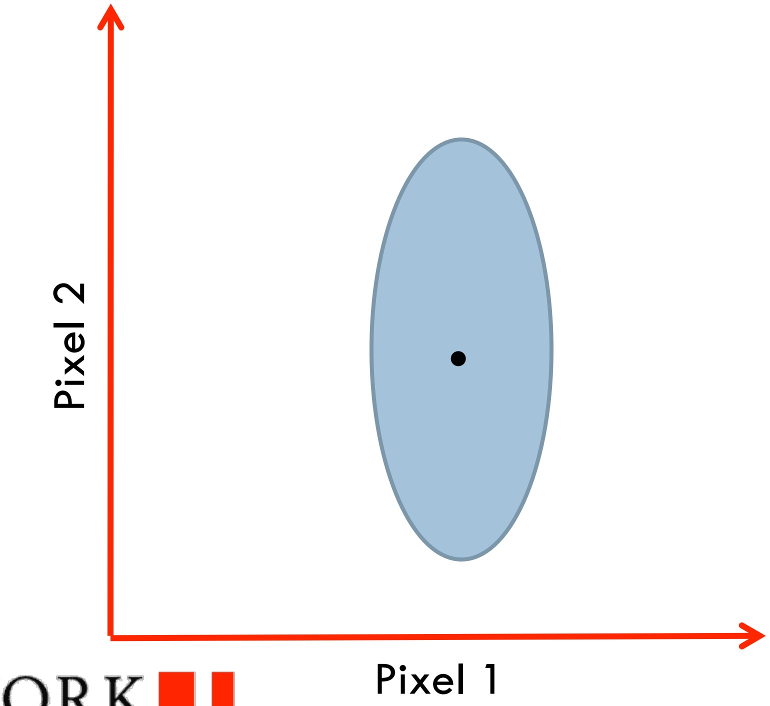


How does this work with a real image?

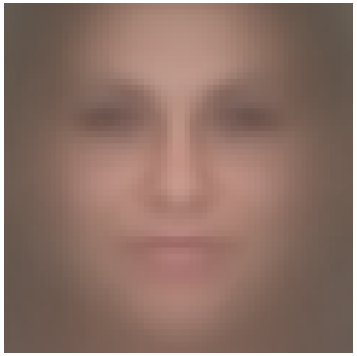
Model # 2: Gaussian, diagonal covariance

$$Pr(\mathbf{x}|\text{face}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}} \exp\left\{-0.5(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

Fit model using maximum likelihood criterion

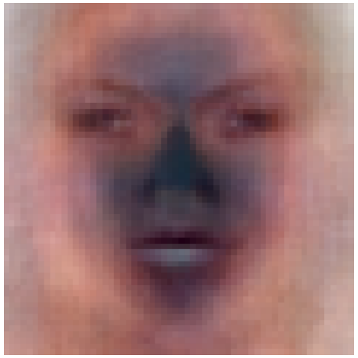


m face



m non-face

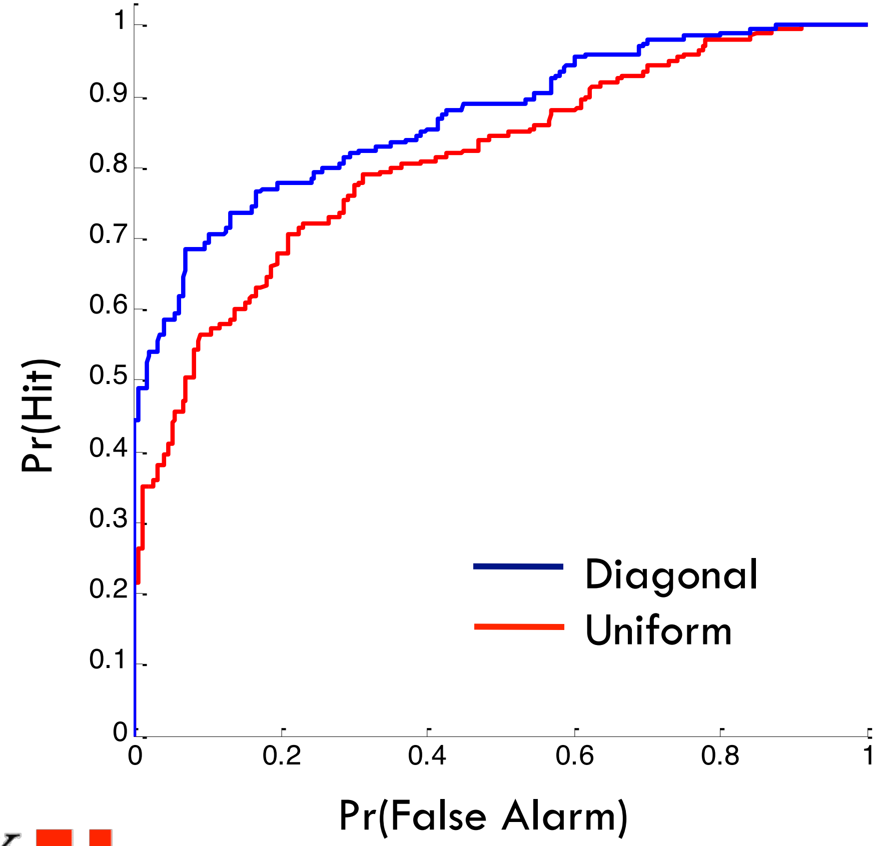
s Face



s non-face

Model 2 Results

Results based on 200 cropped faces and 200 non-faces from the same database.

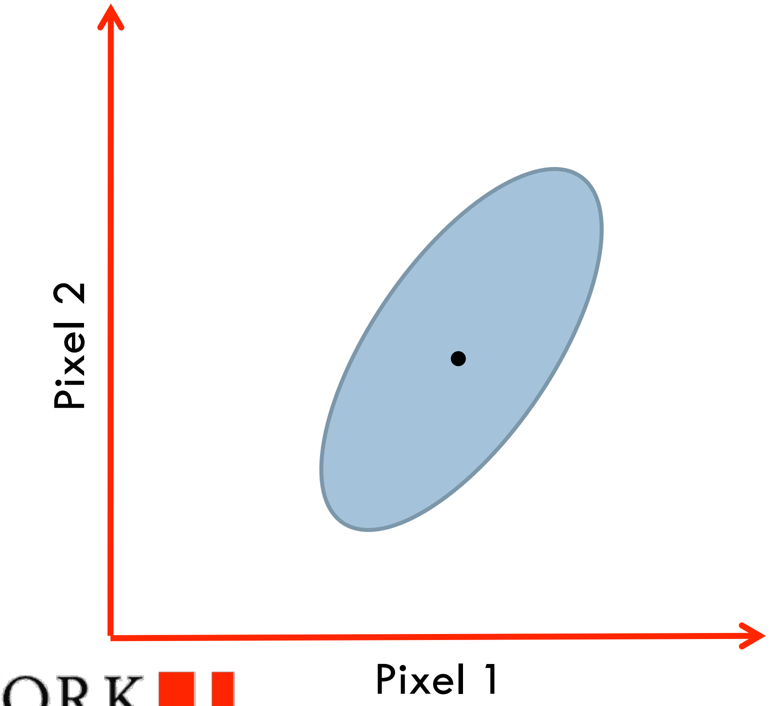


More sophisticated model unsurprisingly classifies new faces and non-faces better.

Model # 3: Gaussian, full covariance

$$Pr(\mathbf{x}|\text{face}) = \frac{1}{(2\pi)^{d/2}|\Sigma|^{1/2}}\exp\left\{-0.5(\mathbf{x}-\mu)^T\Sigma^{-1}(\mathbf{x}-\mu)\right\}$$

Fit model using maximum likelihood criterion



PROBLEM: we cannot fit this model. We don't have enough data to estimate the full covariance matrix.

N=400 training images
D=10800 dimensions

Total number of measured numbers =
ND = 400x10,800 = 4,320,000

Total number of parameters in cov matrix =
(D+1)D/2 = (10,800+1)x10,800/2 =
58,325,400

The Multivariate Normal Distribution: Topics

1. The Multivariate Normal Distribution
2. **Decision Boundaries in Higher Dimensions**
3. Parameter Estimation
    1. Maximum Likelihood Parameter Estimation
    2. Bayesian Parameter Estimation

Decision Surfaces

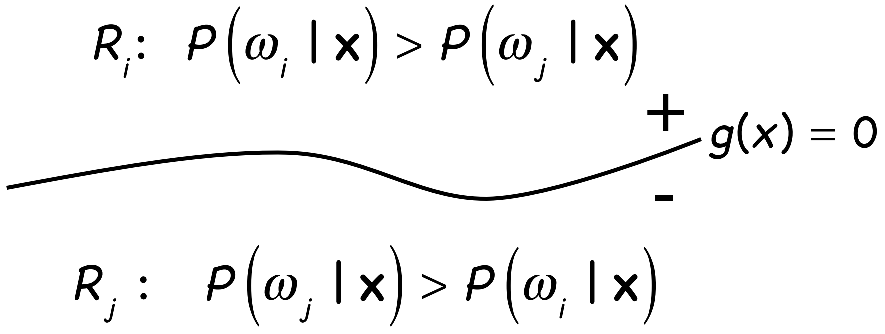
- □ If decision regions R_i and R_j are contiguous, define

$$g(\mathbf{x}) \equiv P(\omega_i \mid \mathbf{x}) - P(\omega_j \mid \mathbf{x})$$

- □ Then the decision surface

$$g(x) = 0$$

separates the two decision regions. *g(x)* is positive on one side and negative on the other.



R_i: $P\left(\omega_i \mid \mathbf{x}\right) > P\left(\omega_j \mid \mathbf{x}\right)$

R_j : $P\left(\omega_j \mid \mathbf{x}\right) > P\left(\omega_i \mid \mathbf{x}\right)$

Discriminant Functions

- If $f(.)$ monotonic, the rule remains the same if we use:

$$\underline{x} \rightarrow \omega_i \text{ if: } f(P(\omega_i|\underline{x})) > f(P(\omega_j|\underline{x})) \quad \forall i \neq j$$

- $g_i(\mathbf{x}) \equiv f(P(\omega_i \mid \mathbf{x}))$ is a **discriminant function**
- In general, discriminant functions can be defined in other ways, independent of Bayes.
- In theory this will lead to a suboptimal solution
- However, non-Bayesian classifiers can have significant advantages:
  - Often a full Bayesian treatment is intractable or computationally prohibitive.
  - Approximations made in a Bayesian treatment may lead to errors avoided by non-Bayesian methods.

Multivariate Normal Likelihoods

- Multivariate Gaussian pdf

$$p(\underline{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)\right)$$

$$\underline{\mu}_i = E\left[\underline{x}|\omega_i\right]$$

$$\Sigma_i = E\left[(\underline{x}-\underline{\mu}_i)(\underline{x}-\underline{\mu}_i)^{\mathrm{T}}|\omega_i\right]$$

Logarithmic Discriminant Function

$$p(\underline{x}|\omega_i) = \frac{1}{(2\pi)^{\frac{D}{2}}\left|\Sigma_i\right|^{\frac{1}{2}}}\exp\left(-\frac{1}{2}(\underline{x}-\underline{\mu}_i)^{\mathrm{T}}\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i)\right)$$

- □ ln(·) is monotonic. Define:

$$g_i(\underline{x}) = \ln\left(p\left(\underline{x}\mid\omega_i\right)P\left(\omega_i\right)\right) = \ln p\left(\underline{x}\mid\omega_i\right) + \ln P(\omega_i)$$

$$= -\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T\Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) + \ln P(\omega_i) + C_i$$

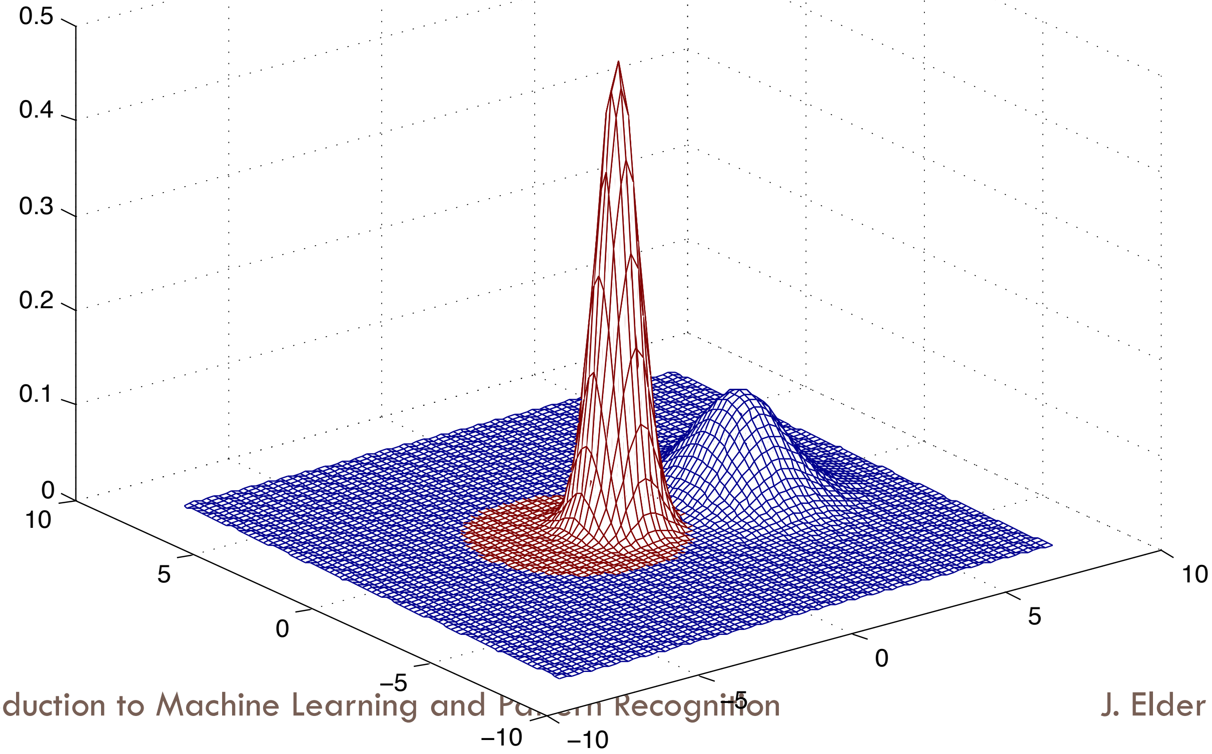
where

$$C_i = -\frac{D}{2}\ln 2\pi - \frac{1}{2}\ln\left|\Sigma_i\right|$$

Quadratic Classifiers

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x} - \underline{\mu}_i) + \ln P(\omega_i) + C_i$$

- Thus the decision surface has a quadratic form.
- For a 2D input space, the decision curves are quadrics (ellipses, parabolas, hyperbolas or, in degenerate cases, lines).



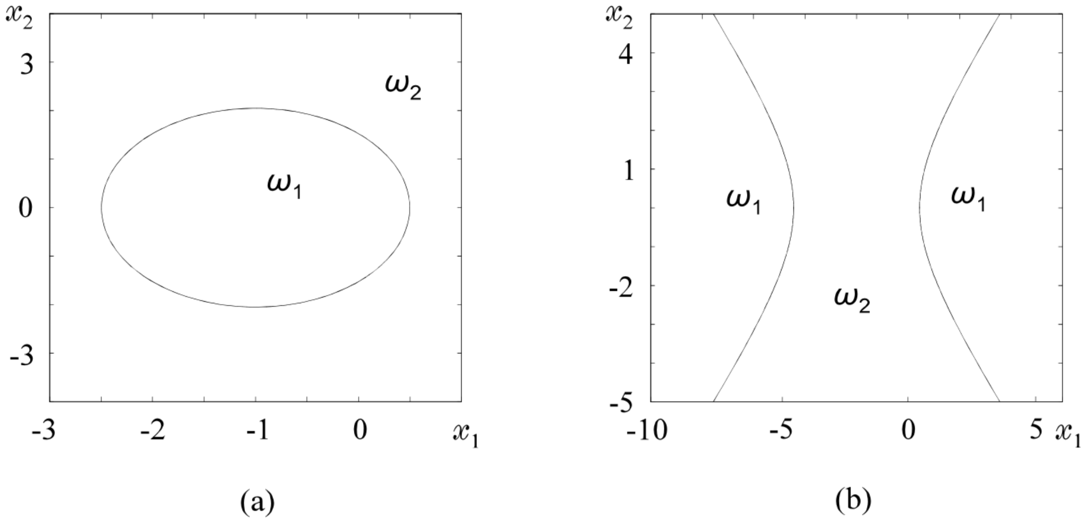
Example: Isotropic Likelihoods

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x} - \underline{\mu}_i) + \ln P(\omega_i) + C_i$$

- Suppose that the two likelihoods are both isotropic, but with different means and variances. Then

$$g_i(x) = -\frac{1}{2\sigma_i^2}(x_1^2 + x_2^2) + \frac{1}{\sigma_i^2}(\mu_{i1}x_1 + \mu_{i2}x_2) - \frac{1}{2\sigma_i^2}(\mu_{i1}^2 + \mu_{i2}^2) + \ln\left(P\left(\omega_i\right)\right) + C_i$$

- And $g_i(\underline{x}) - g_j(\underline{x}) = 0$ will be a quadratic equation in 2 variables.



Equal Covariances

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x} - \underline{\mu}_i) + \ln P(\omega_i) + C_i$$

- The quadratic term of the decision boundary is given by

$$\frac{1}{2}\mathbf{x}^T\left(\Sigma_j^{-1} - \Sigma_i^{-1}\right)\mathbf{x}$$

- Thus if the covariance matrices of the two likelihoods are identical, the decision boundary is linear.

Linear Classifier

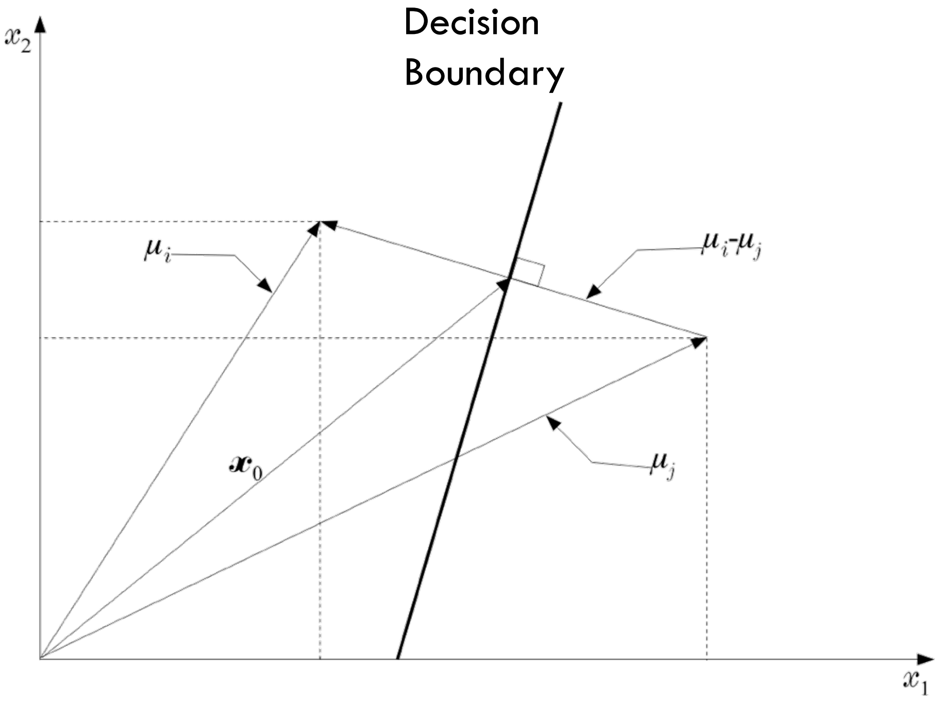
$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma^{-1}(\underline{x} - \underline{\mu}_i) + \ln P(\omega_i) + C_i$$

- □ In this case, we can drop the quadratic terms and express the discriminant function in linear form:

$$\boxed{\begin{aligned} g_i(\underline{x}) &= \underline{w}_i^T \underline{x} + w_{io} \\ \underline{w}_i &= \Sigma^{-1}\underline{\mu}_i \\ w_{i0} &= \ln P(\omega_i) - \frac{1}{2}\underline{\mu}_i^T \Sigma^{-1} \underline{\mu}_i \end{aligned}}$$

Example 1: Isotropic, Identical Variance

$$\boxed{\begin{aligned} g_i(\underline{x}) &= \underline{w}_i^T \underline{x} + w_{io} \\ \underline{w}_i &= \Sigma^{-1}\underline{\mu}_i \\ w_{i0} &= \ln P(\omega_i) - \frac{1}{2}\underline{\mu}_i^T \Sigma^{-1}\underline{\mu}_i \end{aligned}}$$



$\Sigma = \sigma^2 I$. Then the decision surface has the form

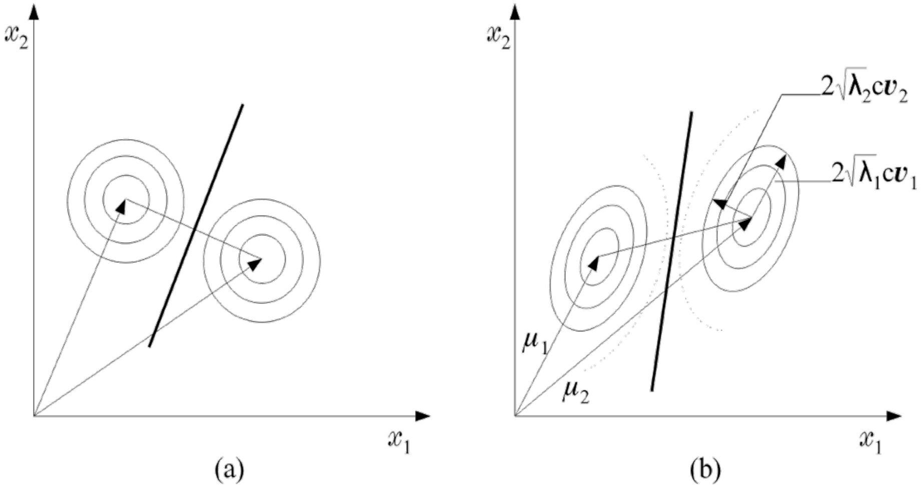
$\underline{w}^T(\underline{x} - \underline{x}_o) = 0$, where

$\underline{w} = \underline{\mu}_i - \underline{\mu}_j$, and

$$\underline{x}_o = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \sigma^2 \ln \frac{P(\omega_i)}{P(\omega_j)} \frac{\underline{\mu}_i - \underline{\mu}_j}{\left\|\underline{\mu}_i - \underline{\mu}_j\right\|^2}$$

Example 2: Equal Covariance

$$\boxed{\begin{aligned} g_i(\underline{x}) &= \underline{w}_i^T \underline{x} + w_{io} \\ \underline{w}_i &= \Sigma^{-1}\underline{\mu}_i \\ w_{i0} &= \ln P(\omega_i) - \frac{1}{2}\underline{\mu}_i^T \Sigma^{-1} \underline{\mu}_i \end{aligned}}$$



$g_{ij}(\underline{x}) = \underline{w}^T(\underline{x} - \underline{x}_0) = 0$ where

$$\underline{w} = \Sigma^{-1}(\underline{\mu}_i - \underline{\mu}_j),$$

$$\underline{x}_0 = \frac{1}{2}(\underline{\mu}_i + \underline{\mu}_j) - \ln\left(\frac{P(\omega_i)}{P(\omega_j)}\right)\frac{\underline{\mu}_i - \underline{\mu}_j}{\left\|\underline{\mu}_i - \underline{\mu}_j\right\|_{\Sigma^{-1}}^2},$$

and

$$\left\|\underline{x}\right\|_{\Sigma^{-1}} \equiv (\underline{x}^T \Sigma^{-1} \underline{x})^{\frac{1}{2}}$$

Minimum Distance Classifiers

- If the two likelihoods have identical covariance AND the two classes are equiprobable, the discrimination function simplifies:

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma_i^{-1}(\underline{x}-\underline{\mu}_i) + \ln P(\omega_i) + C_i$$

$$\Big\Downarrow$$

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x}-\underline{\mu}_i)^T \Sigma^{-1}(\underline{x}-\underline{\mu}_i)$$

Isotropic Case

- In the isotropic case,

$$g_i(\underline{x}) = -\frac{1}{2}(\underline{x} - \underline{\mu}_i)^T \Sigma^{-1}(\underline{x} - \underline{\mu}_i) = -\frac{1}{2\sigma^2}\left\|\underline{x} - \underline{\mu}_i\right\|^2$$

- Thus the Bayesian classifier simply assigns the class that minimizes the Euclidean distance d_e between the observed feature vector and the class mean.

$$d_e = \left\|\underline{x} - \underline{\mu}_i\right\|$$

General Case: Mahalanobis Distance

- To deal with anisotropic distributions, we simply classify according to the Mahalanobis distance, defined as

$$\Delta = g_i(\underline{x}) = \left((\underline{x} - \underline{\mu}_i)^T \Sigma^{-1} (\underline{x} - \underline{\mu}_i)\right)^{1/2}$$

Let $y = U^t\left(x - \mu\right)$. Then we have,

$$\Delta^2 = y^t \Lambda^{-1} y = \sum_{ij} y_i \Lambda_{ij}^{-1} y_j = \sum_i \lambda_i^{-1} y_i^2,$$

where $y_i = u_i^t\left(x - \mu\right)$.

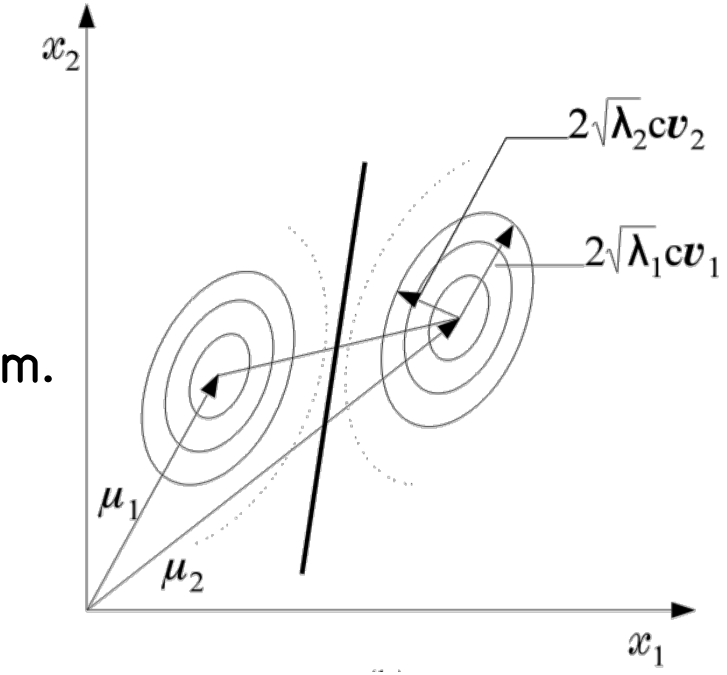
General Case: Mahalanobis Distance

Let $y = U^t(x - \mu)$. Then we have,

$$\Delta^2 = y^t \Lambda^{-1} y = \sum_{ij} y_i \Lambda_{ij}^{-1} y_j = \sum_i \lambda_i^{-1} y_i^2,$$

where $y_i = u_i^t(x - \mu)$.

Thus the curves of constant Mahalanobis distance c have ellipsoidal form.


x_2
$2\sqrt{\lambda_2}c\boldsymbol{v}_2$
$2\sqrt{\lambda_1}c\boldsymbol{v}_1$
μ_1
μ_2
x_1

Example:

Given ω_1, ω_2 : $P(\omega_1) = P(\omega_2)$ and $p(\underline{x}|\omega_1) = N(\underline{\mu}_1, \Sigma)$, $p(\underline{x}|\omega_2) = N(\underline{\mu}_2, \Sigma)$,

$$\underline{\mu}_1 = \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \quad \underline{\mu}_2 = \begin{bmatrix} 3 \\ 3 \end{bmatrix}, \quad \Sigma = \begin{bmatrix} 1.1 & 0.3 \\ 0.3 & 1.9 \end{bmatrix}$$

classify the vector $\underline{x} = \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix}$ using Bayesian classification:

- $\Sigma^{-1} = \begin{bmatrix} 0.95 & -0.15 \\ -0.15 & 0.55 \end{bmatrix}$

- Compute Mahalanobis d_m from μ_1, μ_2 :

$$d^2_{m,1} = \begin{bmatrix} 1.0, & 2.2 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} 1.0 \\ 2.2 \end{bmatrix} = 2.952, \; d^2_{m,2} = \begin{bmatrix} -2.0, & -0.8 \end{bmatrix} \Sigma^{-1} \begin{bmatrix} -2.0 \\ -0.8 \end{bmatrix} = 3.672$$

- Classify $\underline{x} \to \omega_1$. Observe that $d_{E,2} < d_{E,1}$

The Multivariate Normal Distribution: Topics

1. The Multivariate Normal Distribution
2. Decision Boundaries in Higher Dimensions
3. **Parameter Estimation**
    1. **Maximum Likelihood Parameter Estimation**
    2. Bayesian Parameter Estimation

Maximum Likelihood Parameter Estimation

Suppose we believe input vectors $\underline{x}$ are distributed as

$p(\underline{x}) \equiv p(\underline{x};\underline{\theta})$, where $\underline{\theta}$ is an unknown parameter.

Given independent training input vectors $X = \left\{\underline{x}_1, \underline{x}_2, \ldots \underline{x}_N\right\}$

we want to compute the maximum likelihood estimate $\underline{\theta}_{ML}$ for $\underline{\theta}$.

Since the input vectors are independent, we have

$$p(X;\underline{\theta}) \equiv p(\underline{x}_1, \underline{x}_2, \ldots \underline{x}_N;\underline{\theta}) = \prod_{k=1}^{N} p(\underline{x}_k;\underline{\theta})$$

Maximum Likelihood Parameter Estimation

$$p(X;\underline{\theta}) = \prod_{k=1}^{N} p(\underline{x}_k;\underline{\theta})$$

Let $L(\underline{\theta}) \equiv \ln p(X;\underline{\theta}) = \sum_{k=1}^{N} \ln p(\underline{x}_k;\underline{\theta})$

The general method is to take the derivative of L with respect to $\underline{\theta}$, set it to 0 and solve for $\underline{\theta}$:

$$\underline{\hat{\theta}}_{ML} : \quad \frac{\partial L(\underline{\theta})}{\partial(\underline{\theta})} = \sum_{k=1}^{N} \frac{\partial \ln p(\underline{x}_k;\underline{\theta})}{\partial(\underline{\theta})} = \underline{0}$$

Properties of the Maximum Likelihood Estimator

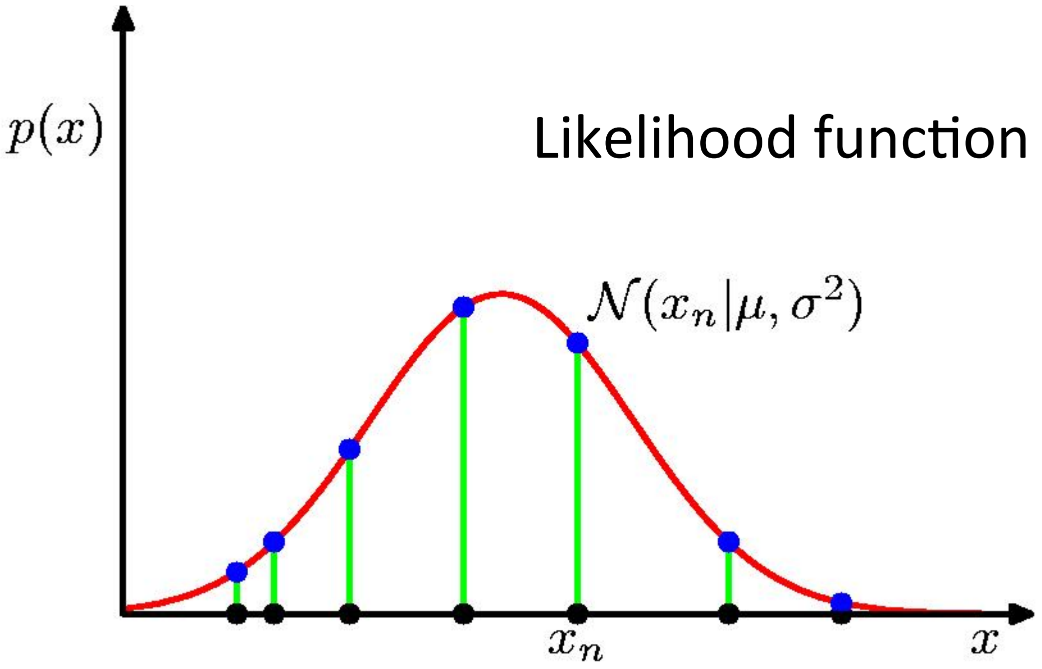
Let $\underline{\theta}_0$ be the true value of the unknown parameter vector.

Then

$\underline{\theta}_{ML}$ is asymptotically unbiased: $\lim\limits_{N \to \propto} E[\underline{\theta}_{ML}] = \underline{\theta}_0$

$\underline{\theta}_{ML}$ is asymptotically consistent: $\lim\limits_{N \to \infty} E \left\| \underline{\hat{\theta}}_{ML} - \underline{\theta}_0 \right\|^2 = 0$

Example: Univariate Normal



$$p(\mathbf{x}|\mu,\sigma^2) = \prod_{n=1}^{N} \mathcal{N}\left(x_n|\mu,\sigma^2\right)$$

Example: Univariate Normal

$$\ln p\left(\mathbf{x}|\mu,\sigma^2\right) = -\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n-\mu)^2 - \frac{N}{2}\ln\sigma^2 - \frac{N}{2}\ln(2\pi)$$

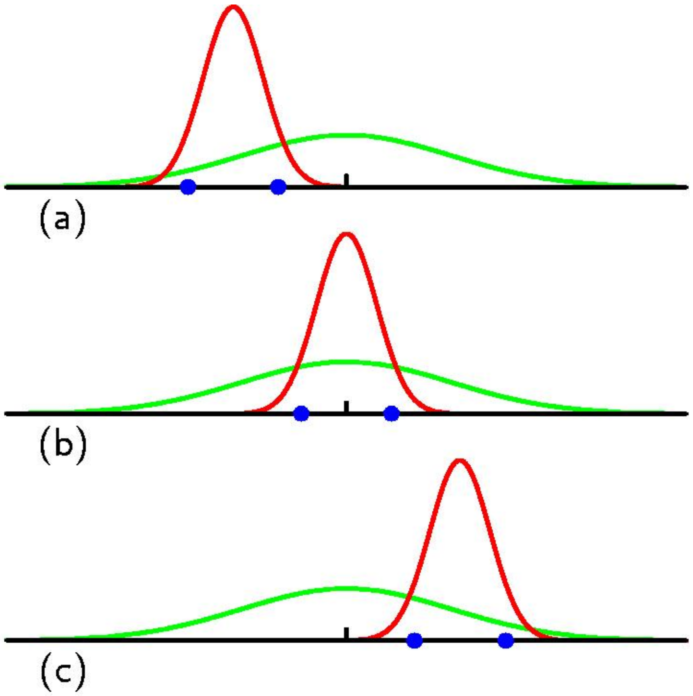
$$\mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n \qquad \sigma^2_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N}(x_n-\mu_{\mathrm{ML}})^2$$

Example: Univariate Normal

$$\mathbb{E}[\mu_{\mathrm{ML}}] = \mu$$

$$\mathbb{E}[\sigma^2_{\mathrm{ML}}] = \left(\frac{N-1}{N}\right)\sigma^2$$

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



Thus σ_{ML} is biased (although asymptotically unbiased).

Example: Multivariate Normal

- Given i.i.d. data $\mathbf{X} = (\mathbf{x}_1, \dots, \mathbf{x}_N)^{\mathrm{T}}$, the log likelihood function is given by

$$\ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = -\frac{ND}{2}\ln(2\pi) - \frac{N}{2}\ln|\boldsymbol{\Sigma}| - \frac{1}{2}\sum_{n=1}^{N}(\mathbf{x}_n - \boldsymbol{\mu})^{\mathrm{T}}\boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu})$$

Maximum Likelihood for the Gaussian

□ Set the derivative of the log likelihood function to zero,

$$\frac{\partial}{\partial \boldsymbol{\mu}} \ln p(\mathbf{X}|\boldsymbol{\mu}, \boldsymbol{\Sigma}) = \sum_{n=1}^{N} \boldsymbol{\Sigma}^{-1}(\mathbf{x}_n - \boldsymbol{\mu}) = 0$$

□ and solve to obtain

$$\boldsymbol{\mu}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} \mathbf{x}_n.$$

□ One can also show that

$$\boldsymbol{\Sigma}_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} (\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})(\mathbf{x}_n - \boldsymbol{\mu}_{\mathrm{ML}})^{\mathrm{T}}.$$

$$\left(\text{Recall: If } \mathbf{x} \text{ and } \mathbf{a} \text{ are vectors, then } \frac{\partial}{\partial \mathbf{x}}\left(\mathbf{x}^t\mathbf{a}\right) = \frac{\partial}{\partial \mathbf{x}}\left(\mathbf{a}^t\mathbf{x}\right) = \mathbf{a}\right)$$

The Multivariate Normal Distribution: Topics

1. The Multivariate Normal Distribution
2. Decision Boundaries in Higher Dimensions
3. **Parameter Estimation**
    1. Maximum Likelihood Parameter Estimation
    2. **Bayesian Parameter Estimation**

Bayesian Inference for the Gaussian (Univariate Case)

□ Assume σ^2 is known. Given i.i.d. data $\mathbf{x} = \{x_1, \ldots, x_N\}$, the likelihood function for μ is given by

$$p(\mathbf{x}|\mu) = \prod_{n=1}^{N} p(x_n|\mu) = \frac{1}{(2\pi\sigma^2)^{N/2}} \exp\left\{-\frac{1}{2\sigma^2}\sum_{n=1}^{N}(x_n - \mu)^2\right\}.$$

□ This has a Gaussian shape as a function of μ (but it is *not* a distribution over μ).

Bayesian Inference for the Gaussian (Univariate Case)

- Combined with a Gaussian prior over μ ,

$$p(\mu) = \mathcal{N}\left(\mu|\mu_0, \sigma_0^2\right).$$

- this gives the posterior

$$p(\mu|\mathbf{x}) \propto p(\mathbf{x}|\mu)p(\mu).$$

- Completing the square over μ, we see that

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

Bayesian Inference for the Gaussian

- … where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

Shortcut: $p\left(\mu \mid X\right)$ has the form $C\exp\left(-\Delta^2\right)$.

Get Δ^2 in form $a\mu^2 - 2b\mu + c = a(\mu - b/a)^2 + \text{const}$ and identify

$$\mu_N = b/a$$

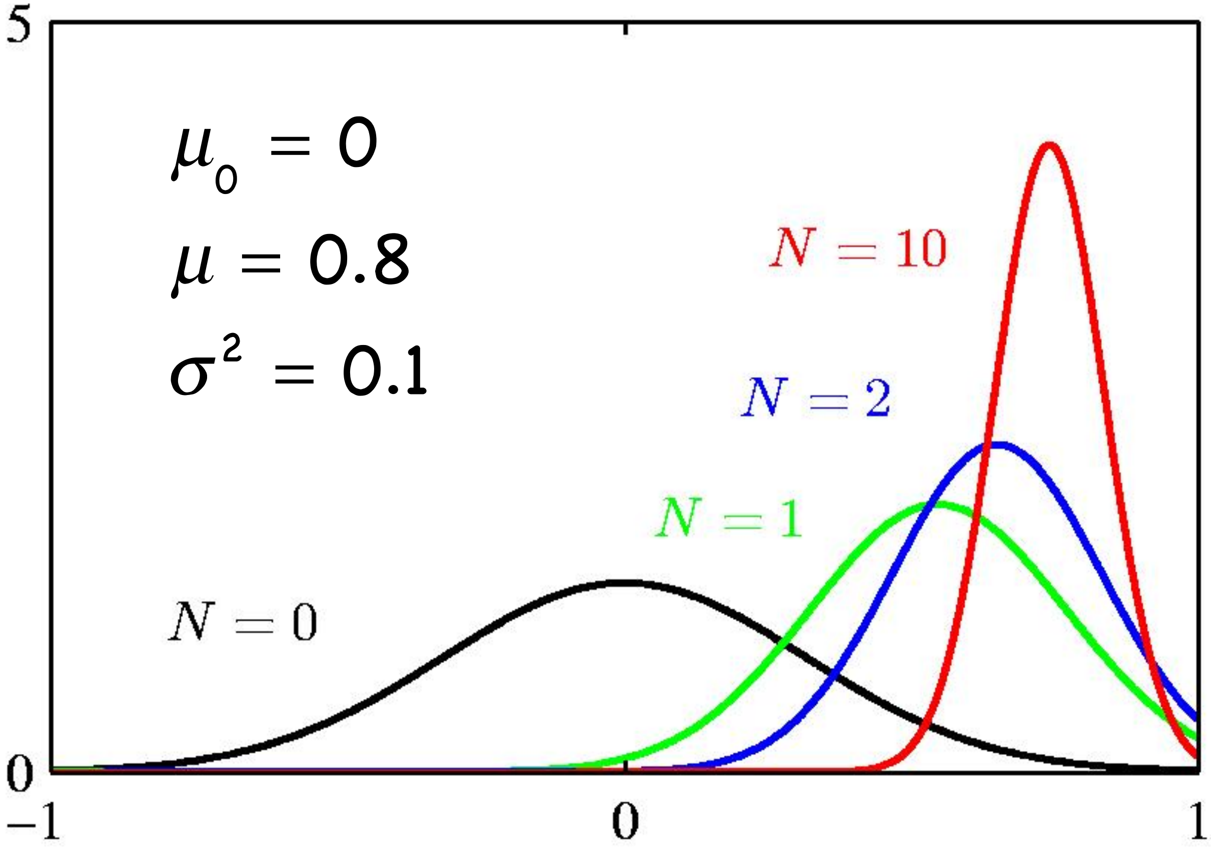
$$\frac{1}{\sigma_N^2} = a$$

- Note:

	$N = 0$	$N \rightarrow \infty$
μ_N	μ_0	μ_{ML}
σ_N^2	σ_0^2	0

Bayesian Inference for the Gaussian

- Example: $p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$



$\mu_0 = 0$

$\mu = 0.8$

$\sigma^2 = 0.1$

Maximum a Posteriori (MAP) Estimation

$$p(\mu|\mathbf{x}) = \mathcal{N}\left(\mu|\mu_N, \sigma_N^2\right)$$

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

In MAP estimation, we use the value of μ that maximizes the posterior $p\left(\mu \mid X\right)$:

$$\mu_{MAP} = \mu_N.$$

Full Bayesian Parameter Estimation

- In both ML and MAP, we use the training data ***X*** to estimate a specific value for the unknown parameter vector $\underline{\theta}$, and then use that value for subsequent inference on new observations **x:** $p\left(\mathbf{x} \mid \underline{\theta}\right)$
- These methods are suboptimal, because in fact we are always uncertain about the exact value of $\underline{\theta}$, and to be optimal we should take into account the possibility that $\underline{\theta}$ assumes other values.

Full Bayesian Parameter Estimation

- In full Bayesian parameter estimation, we do not estimate a specific value for $\underline{\theta}$.
- Instead, we compute the posterior over $\underline{\theta}$, and then integrate it out when computing $p\left(\mathrm{x} \mid \mathrm{X}\right)$:

$$p(\underline{x}|X) = \int p(\underline{x}|\underline{\theta})p(\underline{\theta}|X)d\underline{\theta}$$

$$p(\underline{\theta}|X) = \frac{p(X|\underline{\theta})p(\underline{\theta})}{p(X)} = \frac{p(X|\underline{\theta})p(\underline{\theta})}{\int p(X|\underline{\theta})p(\underline{\theta})d\underline{\theta}}$$

$$p(X|\underline{\theta}) = \prod_{k=1}^{N} p(\underline{x}_k|\underline{\theta})$$

Example: Univariate Normal with Unknown Mean

Consider again the case $p(\underline{x}|\mu) \sim N(\mu,\sigma)$ where σ is known and $\mu \sim N(\mu_0,\sigma_0)$

We showed that $p(\mu|X) \sim N(\mu_N,\sigma_N^2)$, where

$$\mu_N = \frac{\sigma^2}{N\sigma_0^2+\sigma^2}\mu_0 + \frac{N\sigma_0^2}{N\sigma_0^2+\sigma^2}\mu_{\mathrm{ML}}, \qquad \mu_{\mathrm{ML}} = \frac{1}{N}\sum_{n=1}^{N} x_n$$

$$\frac{1}{\sigma_N^2} = \frac{1}{\sigma_0^2} + \frac{N}{\sigma^2}.$$

In the MAP approach, we approximate $p(\underline{x}|\underline{X}) \sim N(\mu_N,\sigma^2)$

In the full Bayesian approach, we calculate $p(\underline{x}|X) = \int p(\underline{x} \mid \mu)p(\mu|X)\,d\mu$

which can be shown to yield $p(\underline{x}|X) \sim N(\mu_N,\sigma^2+\sigma_N^2)$

Comparison: MAP vs Full Bayesian Estimation

- MAP: $p(\underline{x}|\underline{X}) \sim N(\mu_N, \sigma^2)$

- Full Bayesian: $p(\underline{x}|X) \sim N(\mu_N, \sigma^2 + \sigma_N^2)$

- The higher (and more realistic) uncertainty in the full Bayesian approach reflects our posterior uncertainty about the exact value of the mean μ.