

CS 215

Data Analysis and Interpretation

Multivariate Statistics: Multivariate Gaussian

Suyash P. Awate

Multivariate Gaussian – Definition

- Consider a vector random variable $X := [X_1; X_2; \dots; X_D]$
 - Column vector of length D

Definition: The RV X has a multivariate (jointly) Gaussian PDF if \exists a finite set of i.i.d. univariate standard-normal RVs W_1, \dots, W_N (with $D \leq N$) such that each X_d can be expressed as $X_d = \mu_d + \sum_n A_{dn} W_n$ (i.e., $X = AW + \mu$).

Multivariate Gaussian – Identity A

- Consider a vector random variable $X := [X_1; X_2; \dots; X_D]$
 - Column vector of length D

Definition: The RV X has a multivariate (jointly) Gaussian PDF if \exists a finite set of i.i.d. univariate standard-normal RVs W_1, \dots, W_N (with $D \leq N$) such that each X_d can be expressed as $X_d = \mu_d + \sum_n A_{dn} W_n$ (i.e., $X = AW + \mu$).

- Example 1 (Zero-Mean + Isotropic / Spherical Gaussian): The case of independent standard-normal RVs W_1, \dots, W_D with $A := I_{D \times D}$ and $\mu := 0$, i.e. $X = W$

Then, the Gaussian PDF is $p(w) = \prod_d p(w_d) = \frac{1}{(2\pi)^{D/2}} \exp(-0.5w^\top w)$

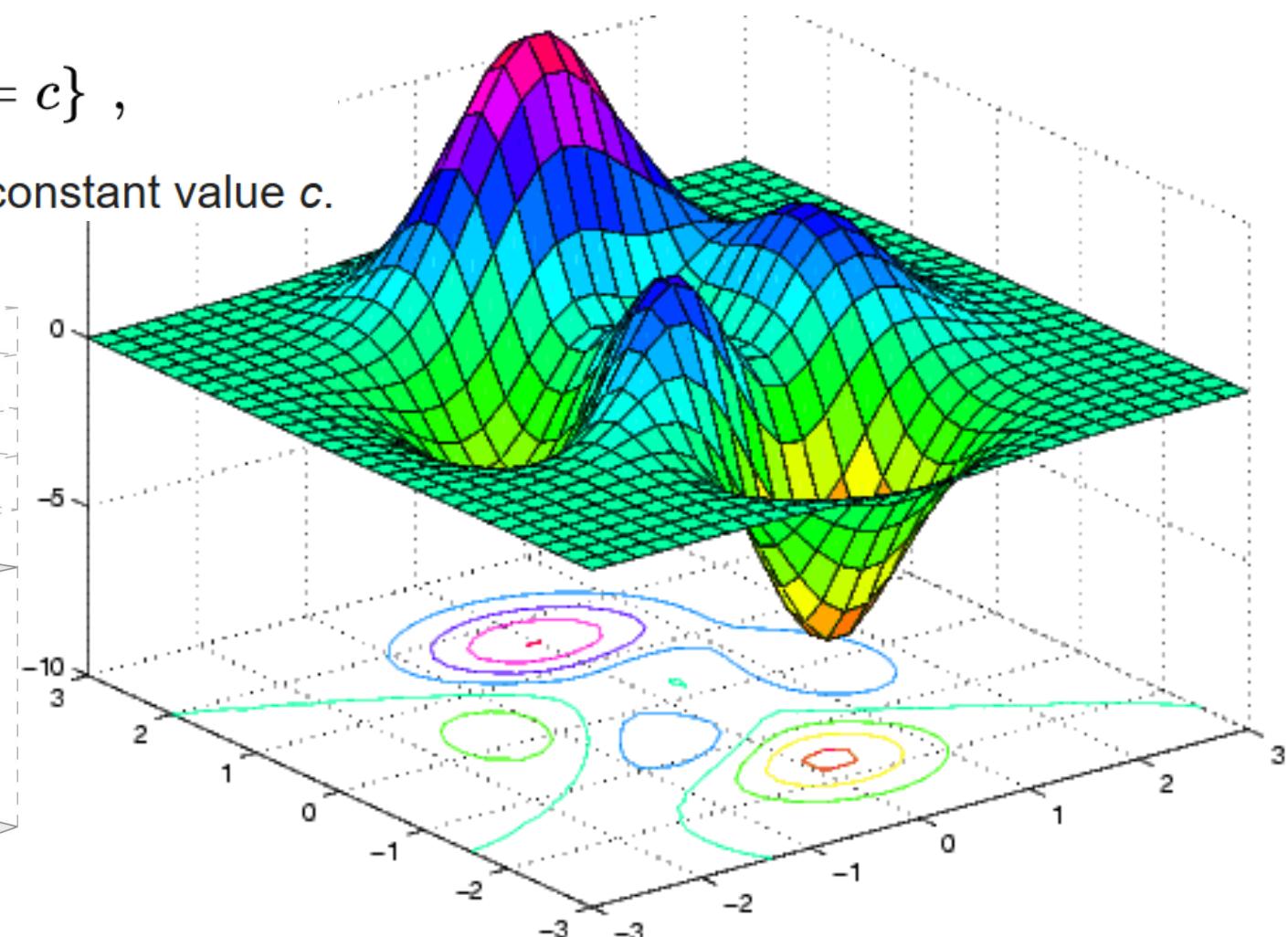
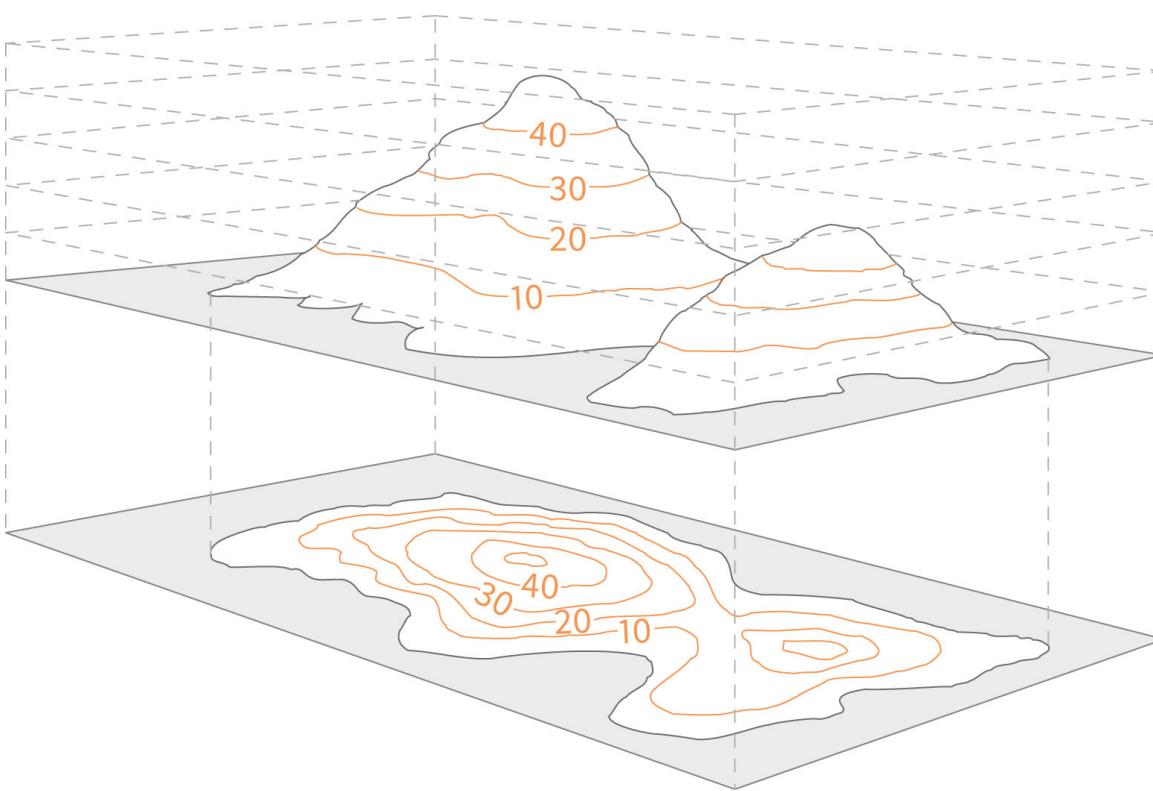
Multivariate Gaussian – Identity A

- What are the level sets of the PDF ?

In mathematics, a **level set** of a real-valued function f of n real variables is a set of the form

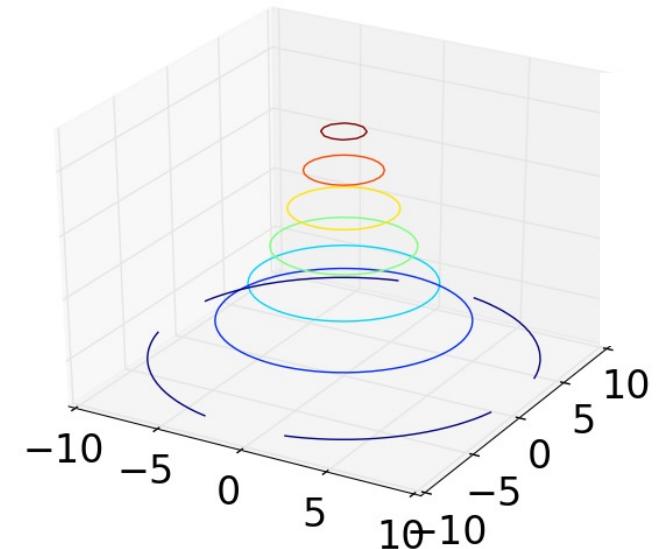
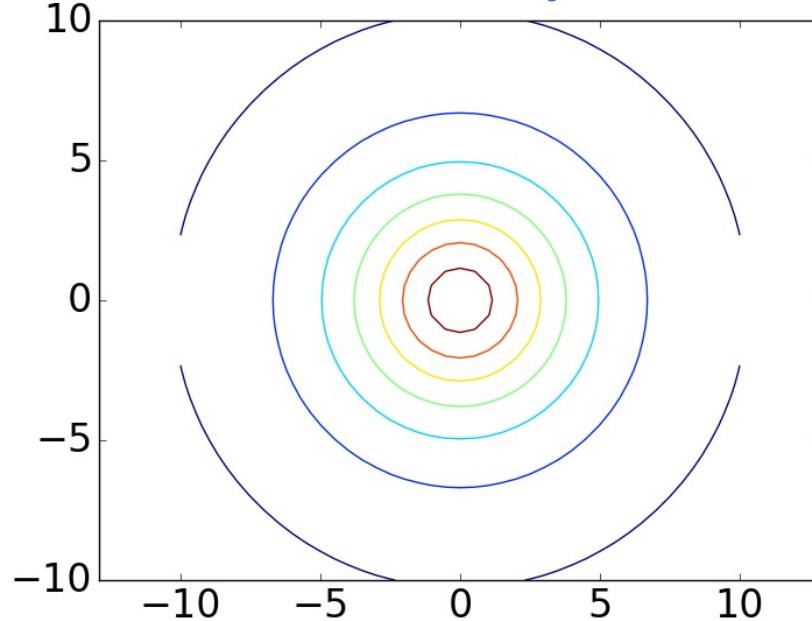
$$L_c(f) = \{(x_1, \dots, x_n) \mid f(x_1, \dots, x_n) = c\},$$

that is, a set where the function takes on a given constant value c .

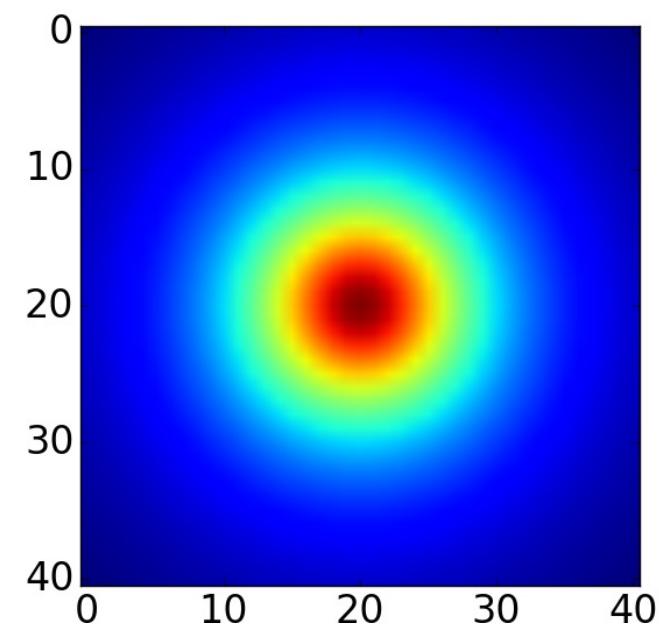
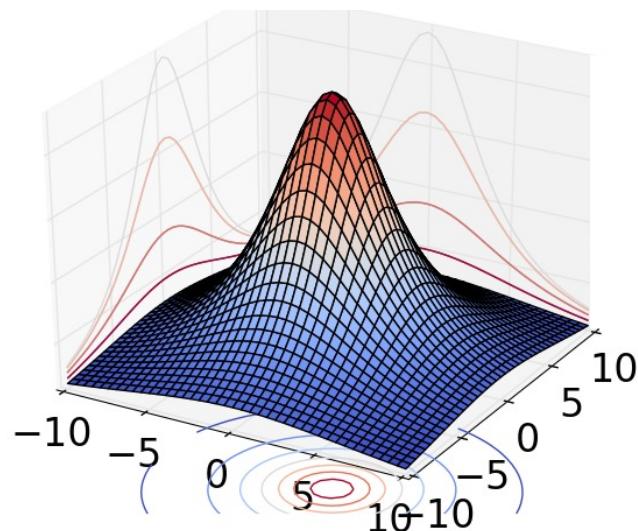


Multivariate Gaussian – Identity A

- Isotropic / spherical multivariate Gaussian
 - Level sets



$$p(w) = \prod_d p(w_d) = \frac{1}{(2\pi)^{D/2}} \exp(-0.5w^\top w)$$

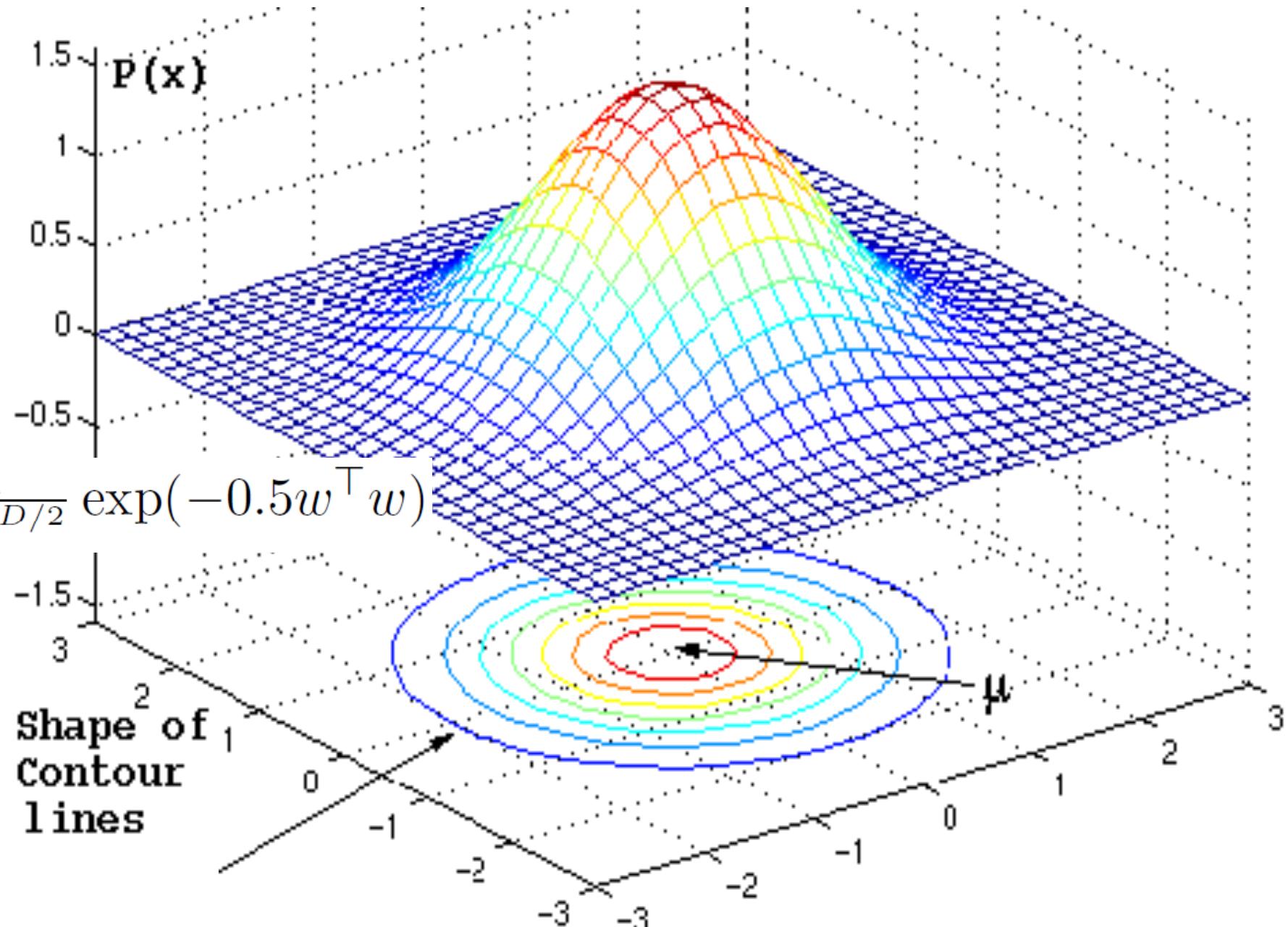


Multivariate Gaussian – Identity A

- Isotropic / spherical multivariate Gaussian

- Level sets

$$p(w) = \prod_d p(w_d) = \frac{1}{(2\pi)^{D/2}} \exp(-0.5w^\top w)$$

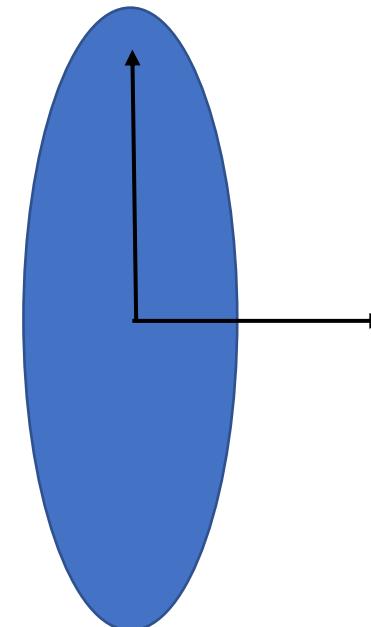
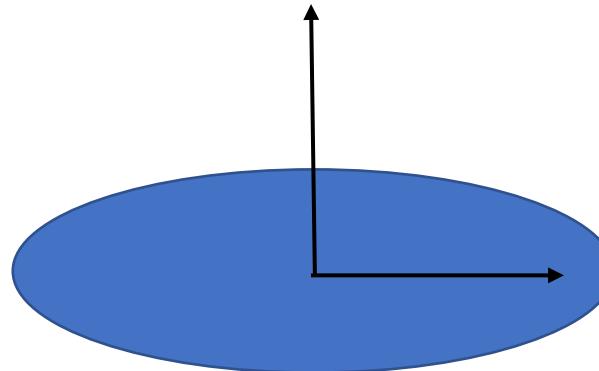
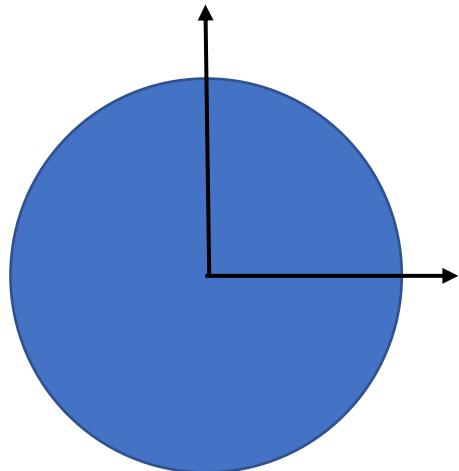


Multivariate Gaussian – Diagonal A

- $X = AW + \mu$
- What is PDF $q(X)$ for **non-singular square diagonal** matrix A , and some μ ?
 - $X_1 = A_{11} W_1 + \mu_1$: Gaussian with mean μ_1 , standard deviation $\sigma_1 = |A_{11}|$
 - $X_2 = A_{22} W_2 + \mu_2$: Gaussian with mean μ_2 , standard deviation $\sigma_2 = |A_{22}|$
 - ...
 - $X_D = A_{DD} W_D + \mu_D$: Gaussian with mean μ_D , standard deviation $\sigma_D = |A_{DD}|$
 - $P(X) = P(X_1, X_2, \dots, X_D) = G(X_1; \mu_1, \sigma_1^2) G(X_2; \mu_2, \sigma_2^2) \dots G(X_D; \mu_D, \sigma_D^2)$
 - Any level set of PDF $q(X)$ is a hyper-ellipsoid with:
 - Center at μ
 - Axes aligned with cardinal axes

Multivariate Gaussian – Diagonal A

- $X = AW + \mu$
- What is PDF $q(X)$ for **non-singular square diagonal** matrix A, some μ ?
 - $P(X) = P(X_1, X_2, \dots, X_D) = G(X_1; \mu_1, \sigma_1^2) G(X_2; \mu_2, \sigma_2^2) \dots G(X_D; \mu_D, \sigma_D^2)$
 - Example 1-3 (left to right):
means are $(\mu_1, \mu_2) : (0,0), (0,0), (0,0)$
variances are $(\sigma_1^2, \sigma_2^2) : (4,4), (6,2), (2,6)$

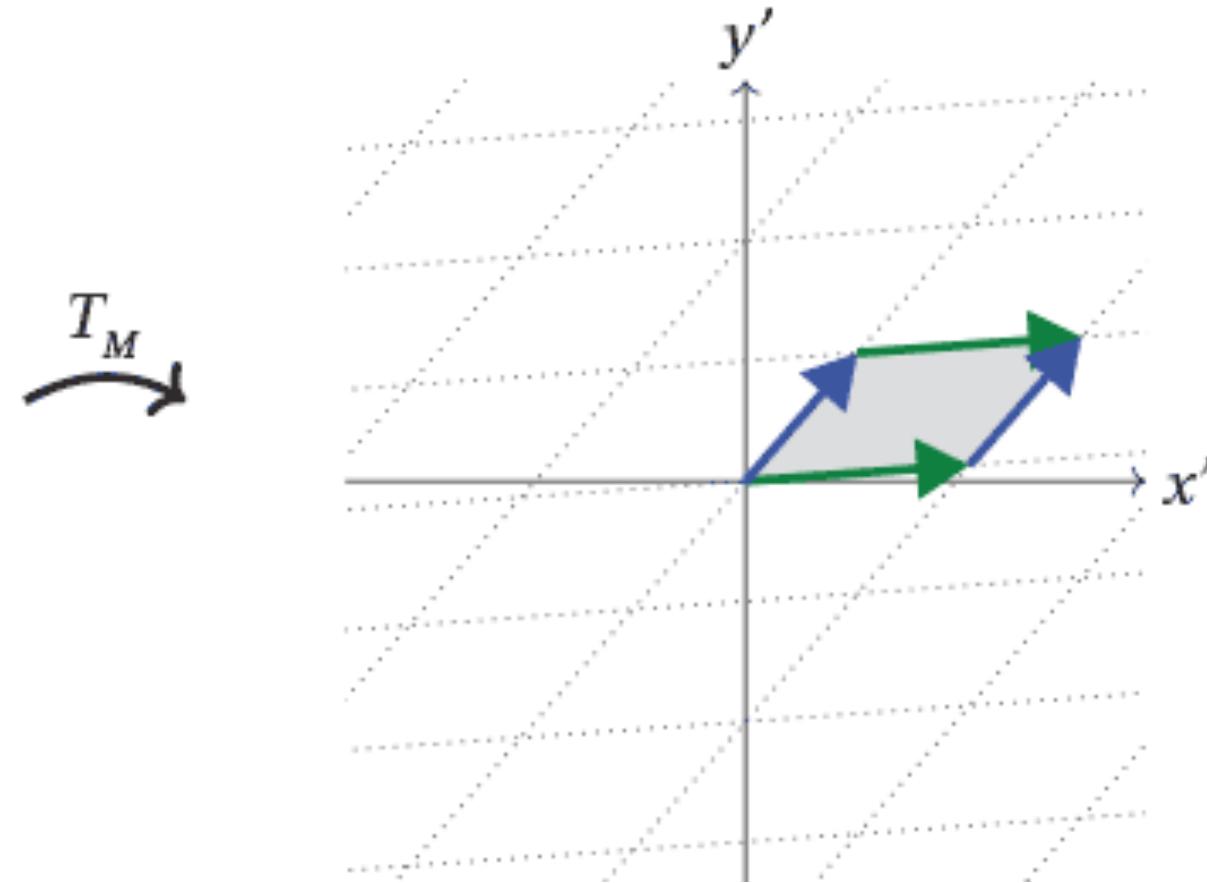
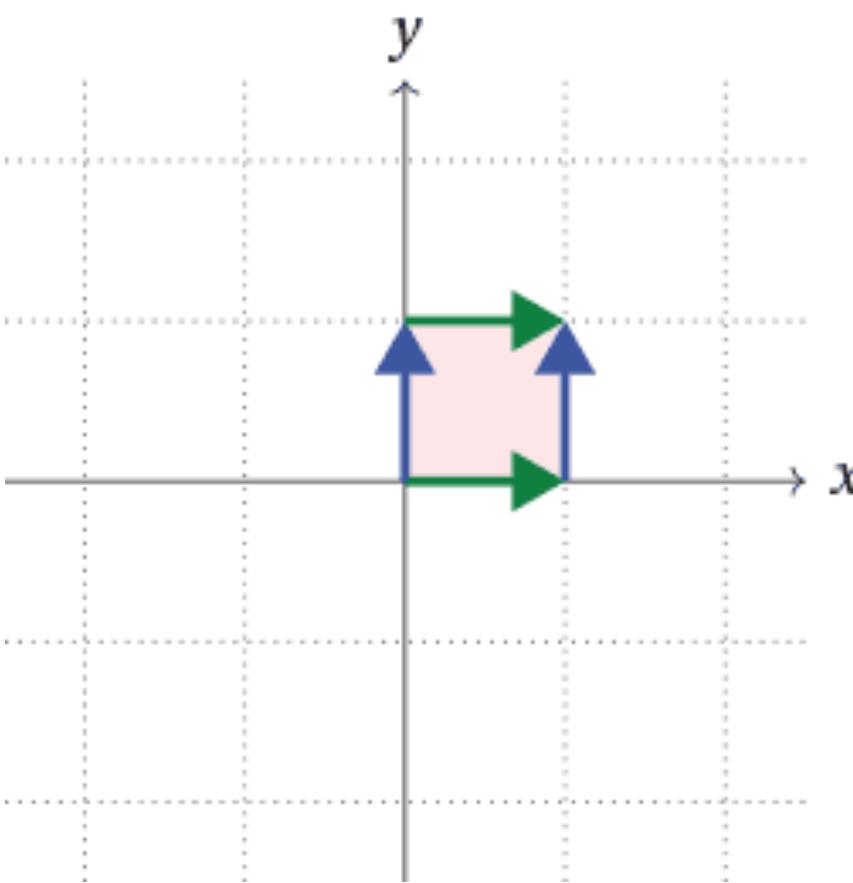
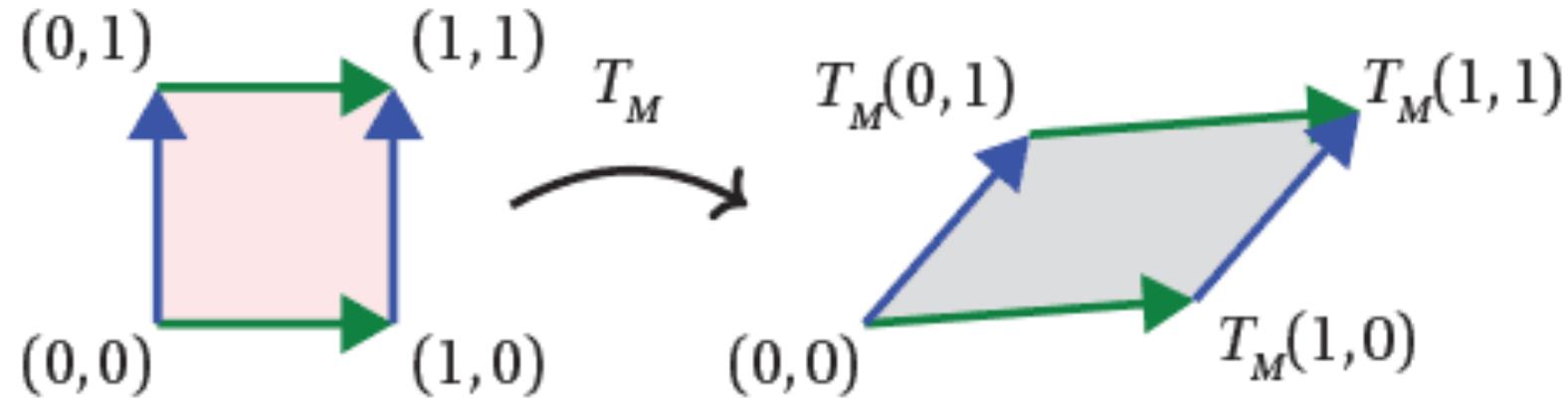


Multivariate Gaussian – Non-Singular A

- $X = A W + \mu$
- What is PDF $q(X)$ for **non-singular square** matrix A and $\mu = 0$?
- Transformation of random variables (multivariate case)
 - Transformation is $X := g(W) := A W$
 - Inverse transformation is $W = g^{-1}(X) = A^{-1}X$
 - Univariate case
 - We wanted magnitude of **derivative** of $g^{-1}(.)$
 - Measured local scaling in **lengths** caused by $g^{-1}(.)$
 - Multivariate case
 - Measure local scaling in **volumes** caused by $g^{-1}(.)$
 - We want the magnitude of the volume-scaling given by Jacobian of $g^{-1}(.)$
 - Magnitude of determinant of **Jacobian** of $g^{-1}(.)$, where the Jacobian is A^{-1}

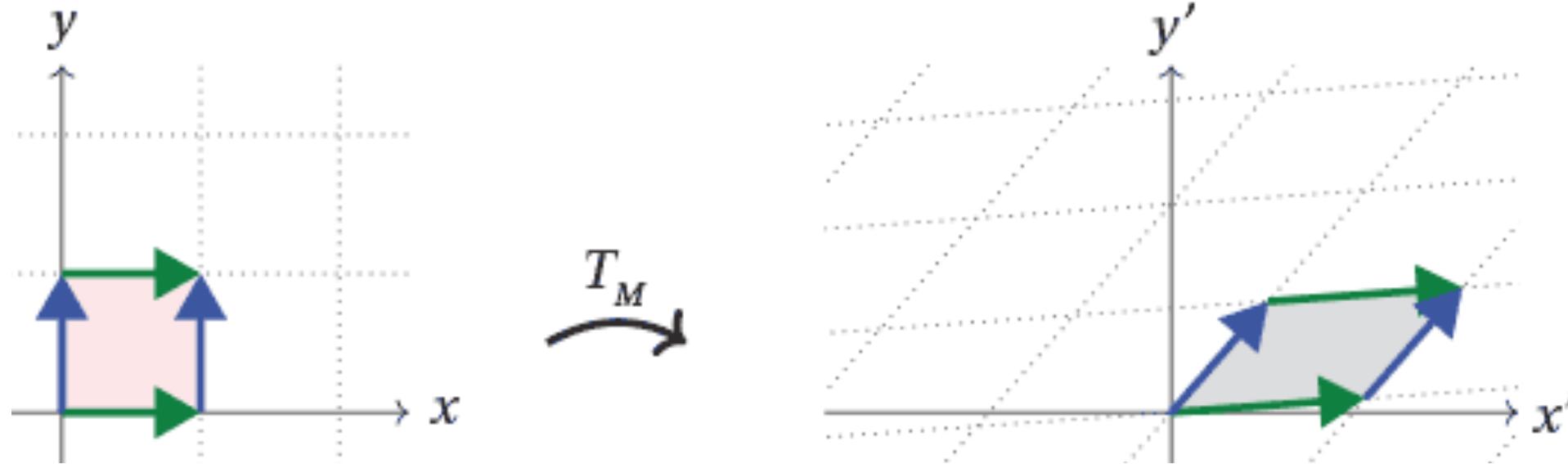
Multivariate Gaussian – Non-Singular A

- Linear transformation
 $W := A^{-1} X$



Multivariate Gaussian – Non-Singular A

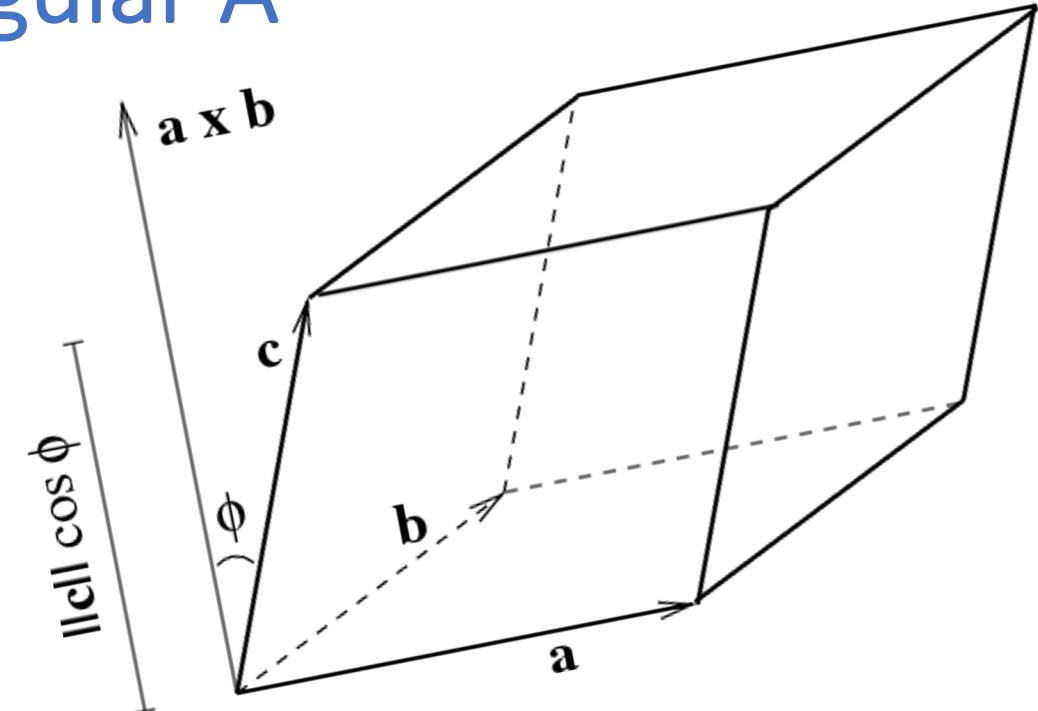
- Linear transformation $W := A^{-1} X$
 - Transformation A^{-1} maps an infinitesimal hyper-cube (dX) $\delta \times \delta \times \dots \times \delta$ (D times) → an infinitesimal hyper-parallelepiped (dW)
 - If axes of hyper-cube were unit vectors along cardinal axes, then axes of hyper-parallelepiped are columns of A^{-1}
 - If volume of the hyper-cube (dX) is δ^D , then volume of hyper-parallelepiped (dW) is $\delta^D \det(A^{-1}) = \delta^D / \det(A)$



Multivariate Gaussian – Non-Singular A

- Volume of a parallelepiped (in 3D)
 - Scalar triple product

$$\begin{aligned}\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c}) &= \begin{vmatrix} a_x & a_y & a_z \\ b_x & b_y & b_z \\ c_x & c_y & c_z \end{vmatrix} = \begin{vmatrix} a_x & b_x & c_x \\ a_y & b_y & c_y \\ a_z & b_z & c_z \end{vmatrix} \\ &= \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b}) \\ &= -\mathbf{a} \cdot (\mathbf{c} \times \mathbf{b}) = -\mathbf{c} \cdot (\mathbf{b} \times \mathbf{a}) = -\mathbf{b} \cdot (\mathbf{a} \times \mathbf{c})\end{aligned}$$



$$\begin{aligned}\text{Volume} &= \text{area of base} \cdot \text{height} \\ &= \|\mathbf{a} \times \mathbf{b}\| \|\mathbf{c}\| |\cos \phi| = |(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}|\end{aligned}$$

The notation $[\mathbf{a}, \mathbf{b}, \mathbf{c}]$ is also used for $\mathbf{a} \cdot (\mathbf{b} \times \mathbf{c})$.

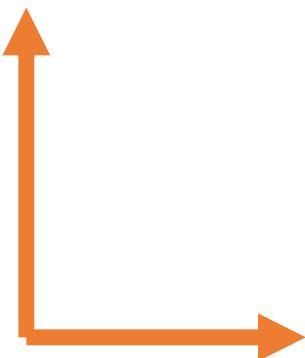
Multivariate Gaussian – Non-Singular A

- Why is volume of hyper-parallelepiped given by determinant of matrix with columns as sides of hyper-parallelepiped ?
 - An inductive proof exists. We consider the following non-inductive reasoning.
 - Two important properties from linear algebra and geometry:
 - a) **Adding multiples of one column** to another doesn't change determinant, because determinant function is multi-linear
 - b) **Adding multiples of one side vector** to another doesn't change volume, because it causes a skew translation of hyper-parallelepiped
- 1. Using **Gram-Schmidt orthogonalization**, transform matrix A^{-1} to a matrix, say, B_{orthog} with orthogonal columns (NOT orthonormal columns; that would have determinant 1)
 - This doesn't change determinant or volume

Multivariate Gaussian – Non-Singular A

1. Gram–Schmidt orthogonalization

- $\{\mathbf{v}_1, \mathbf{v}_2, \dots\}$ to $\{\mathbf{u}_1, \mathbf{u}_2, \dots\}$



$$\mathbf{u}_1 = \mathbf{v}_1,$$

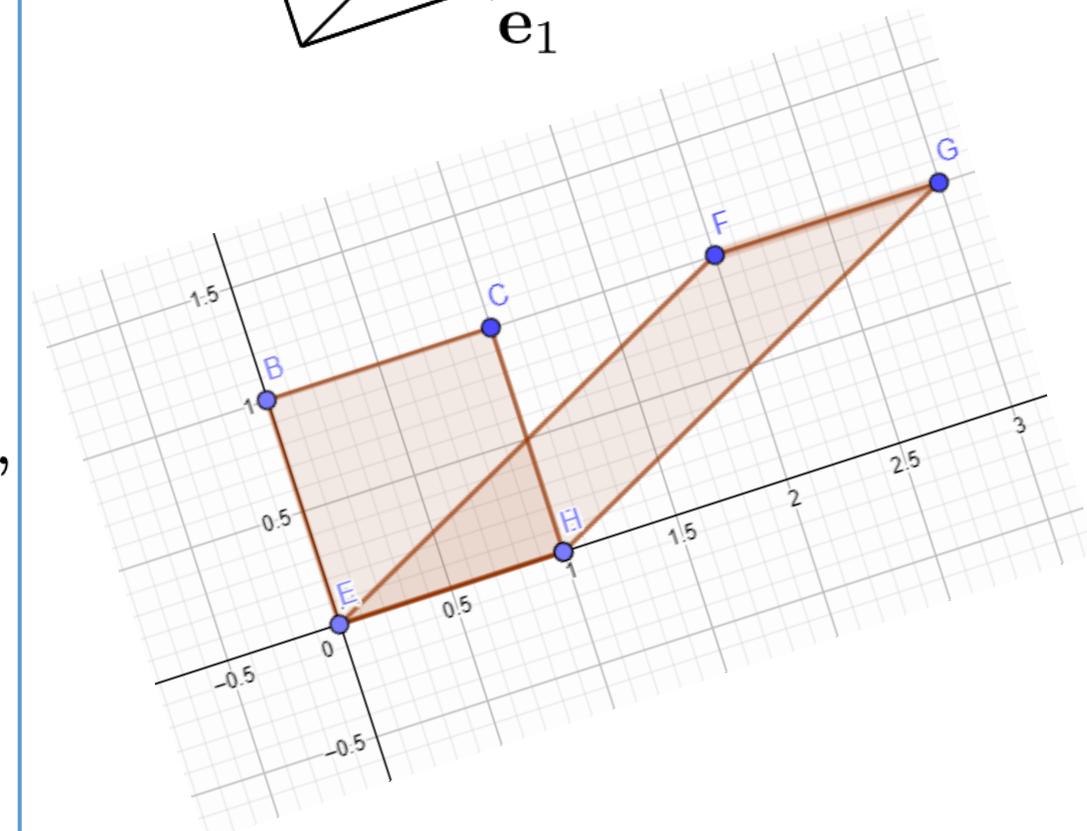
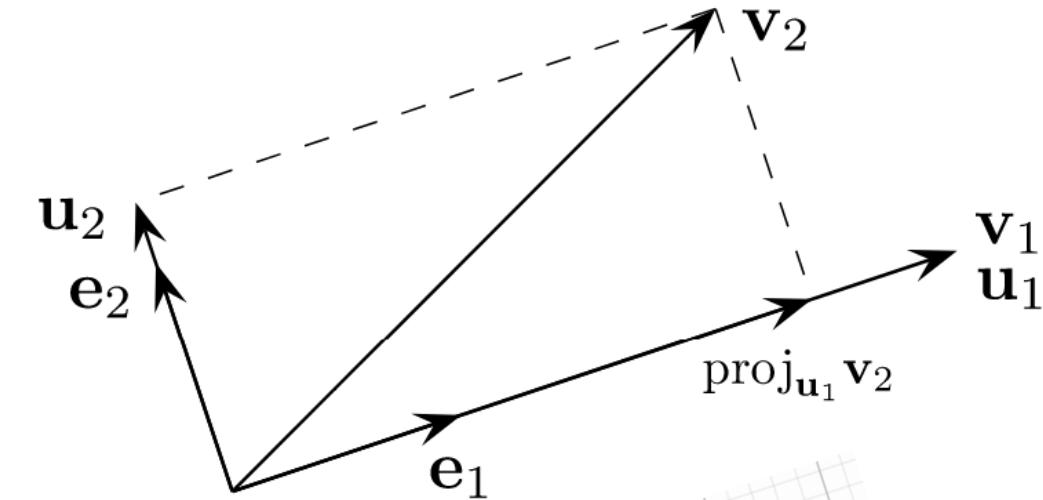
$$\mathbf{u}_2 = \mathbf{v}_2 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_2),$$

$$\mathbf{u}_3 = \mathbf{v}_3 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_3) - \text{proj}_{\mathbf{u}_2}(\mathbf{v}_3),$$

$$\mathbf{u}_4 = \mathbf{v}_4 - \text{proj}_{\mathbf{u}_1}(\mathbf{v}_4) - \text{proj}_{\mathbf{u}_2}(\mathbf{v}_4) - \text{proj}_{\mathbf{u}_3}(\mathbf{v}_4),$$

⋮

$$\mathbf{u}_k = \mathbf{v}_k - \sum_{j=1}^{k-1} \text{proj}_{\mathbf{u}_j}(\mathbf{v}_k),$$

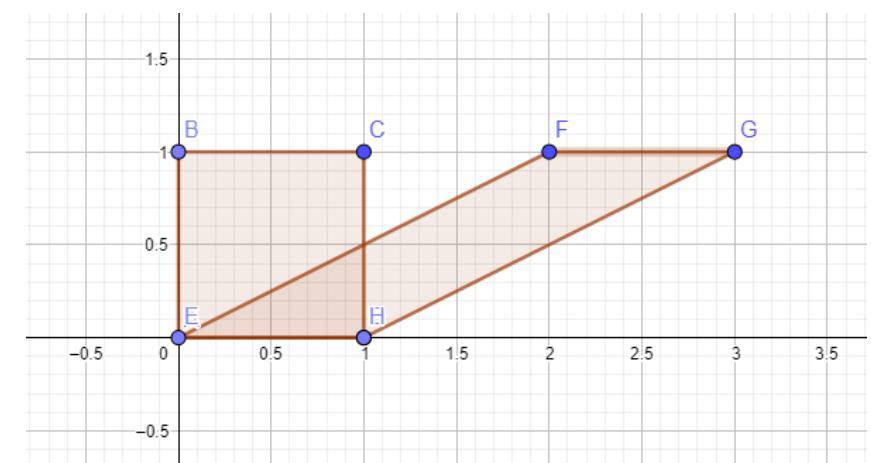
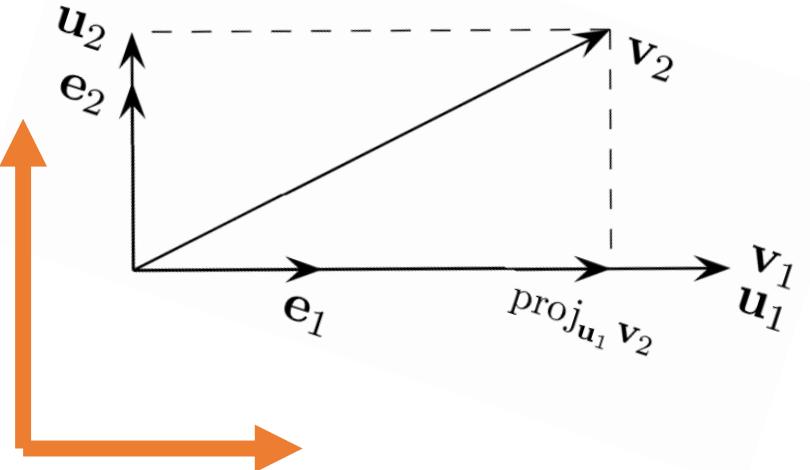
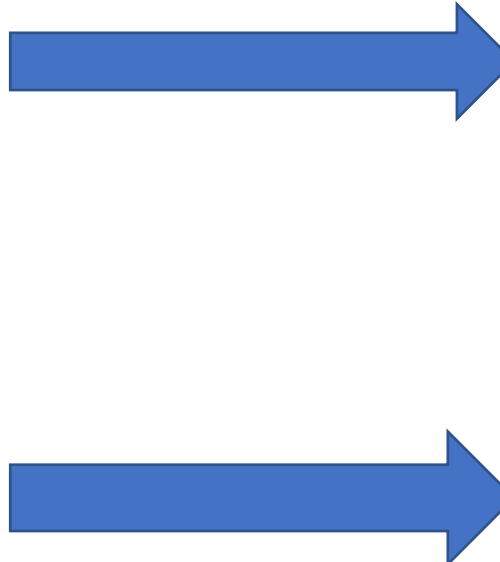
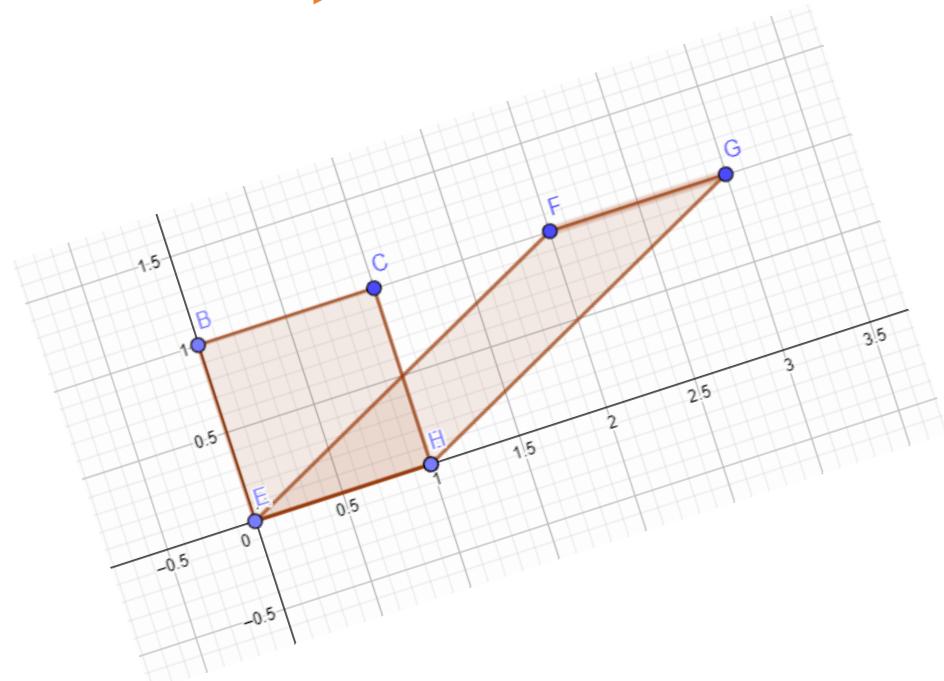
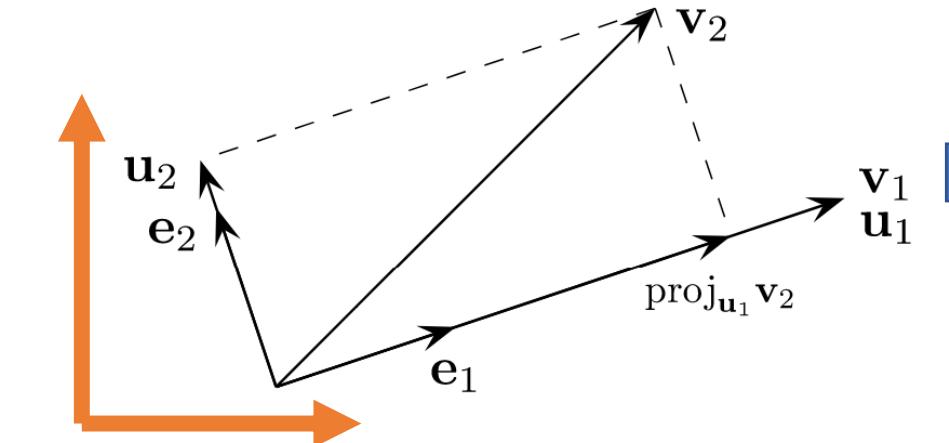


Multivariate Gaussian – Non-Singular A

- Why is volume of hyper-parallelepiped given by determinant of matrix with columns as sides of hyper-parallelepiped ?
 - 2 important properties from linear algebra and geometry:
Adding multiples of one column, or side vector, to another:
 - a) doesn't change determinant, because determinant function is multi-linear
 - b) doesn't change volume, because it causes a skew translation of hyper-parallelepiped
- 1. Using Gram-Schmidt orthogonalization,
transform matrix A^{-1} to a matrix, say, B_{orthog} with orthogonal columns
(NOT orthonormal columns; that would have determinant 1)
- 2. Rotate B_{orthog} to make it to diagonal form (align columns to cardinal axes)
 - This doesn't change determinant or volume

Multivariate Gaussian – Non-Singular A

2. Rotate $\{u_1, u_2, \dots\}$ to align it to cardinal axes



Multivariate Gaussian – Non-Singular A

- Why is volume of hyper-parallelepiped given by determinant of matrix with columns as sides of hyper-parallelepiped ?
 - 2 important properties from linear algebra and geometry:
Adding multiples of one column, or side vector, to another:
 - a) doesn't change determinant, because determinant function is multi-linear
 - b) doesn't change volume, because it causes a skew translation of hyper-parallelepiped
- 1. Using Gram-Schmidt orthogonalization,
transform matrix A^{-1} to a matrix, say, B_{orthog} with orthogonal columns
(NOT orthonormal columns; that would have determinant 1)
- 2. Rotate B_{orthog} to make it to diagonal form (align columns to cardinal axes)
- 3. For this diagonal matrix (aligned hyper-rectangle),
determinant magnitude (= product of diagonal-entries' magnitudes) =
volume of a hyper-rectangle (= product of side lengths)
- 4. Now trace back all operations

Multivariate Gaussian – Non-Singular A

- $X = A W + \mu$
- What is the PDF $q(X)$ for non-singular square matrix A and $\mu = 0$?
- Transformation of random variables (multivariate case)
 - Transformation is $X := g(W) := A W$
 - Inverse transformation is $W = g^{-1}(X) = A^{-1}X$
 - Multivariate case
 - Measure local scaling in volumes caused by $g^{-1}(\cdot)$
 - We want the magnitude determinant of Jacobian of $g^{-1}(\cdot)$

$$q(X) = p(A^{-1}X) \frac{1}{|\det(A)|} = \frac{1}{(2\pi)^{D/2} |\det(A)|} \exp(-0.5X^\top (A^{-1})^\top A^{-1} X)$$

Let $C := AA^\top$. Then, $C^{-1} = (A^{-1})^\top A^{-1}$ and $\det(C) = \det(A)\det(A^\top) = (\det(A))^2$

$$q(X) = \frac{1}{(2\pi)^{D/2} |C|^{0.5}} \exp(-0.5X^\top C^{-1} X)$$

Multivariate Gaussian – Non-Singular A, Non-Zero μ

- If $X := A W$ is a multivariate Gaussian,
then $Y := X + \mu$ is a multivariate Gaussian with

$$p(y) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$$

- Proof:
 - Follows from the transformation $X := Y - \mu := g^{-1}(Y)$

Multivariate Gaussian – Composite Transformations

- If Y is multivariate Gaussian,
then $Z := BY + c$ is multivariate Gaussian,
where matrix B is square invertible
- Proof:
 - Because Y is multivariate Gaussian, we have $Y = AW + \mu$, where A is invertible
 - Thus,
$$\begin{aligned} Z \\ &= B(AW + \mu) + c \\ &= (BA)W + (B\mu + c), \text{ where matrix } BA \text{ is invertible} \end{aligned}$$

Multivariate Statistics – Mean and Covariance

Multivariate Statistics – Mean

- For an general random (column) vector X ,
the mean vector is

$$E_{P(X)}[X]$$

= a (column) vector with the i -th component as $E_{P(X)}[X_i] = E_{P(X_i)}[X_i]$

Multivariate Statistics – Covariance

- Covariance matrix for a general random (column) vector Y is defined as:

$$C := E_{P(Y)} [(Y - E[Y]) (Y - E[Y])^T]$$

- So,

$$C_{ij} = ?$$

$$= E_{P(Y)} [(Y_i - E[Y_i]) (Y_j - E[Y_j])]$$

$$= E_{P(Y_i, Y_j)} [(Y_i - E[Y_i]) (Y_j - E[Y_j])]$$

$$= \text{Cov}(Y_i, Y_j)$$

Multivariate Statistics – Covariance

- More properties of covariance matrix C (for a general random vector X)

$$(1) C = E[XX^\top] - E[X](E[X])^\top$$

Proof: Expand the terms in the definition

$$(2) C \text{ is symmetric}$$

Proof: $C_{ij} = \text{Cov}(X_i, X_j) = \text{Cov}(X_j, X_i) = C_{ji}$

$$(3) C \text{ is positive semi-definite (PSD)}$$

Proof: For any $D \times 1$ non-zero vector a , we get $a^\top Ca = E[a^\top(X - E[X])(X - E[X])^\top a] = E[(f(X))^\top f(X)] \geq 0$ that is the variance of a scalar RV $f(X) = (X - E[X])^\top a$

- If A is invertible, then C is invertible, and then C is positive definite (PD) where $a^\top Ca > 0$ for any non-zero vector a

Multivariate Gaussian – Mean and Covariance

Multivariate Gaussian – Mean

- The **mean** vector of $X := AW + \mu$ is μ
- Proof:
 - When $X = AW + \mu$,
 $E_{P(X)}[X] = E_{P(W)}[AW + \mu] = \mu + E_{P(W)}[AW] = \mu + A E_{P(W)}[W] = \mu$
- Notes:
 - Take the expectation of first component of AW , i.e.,
$$\begin{aligned}E_{P(W)} [A_{11}W_1 + A_{12}W_2 + \dots A_{1D}W_D] \\= A_{11} E_{P(W)} [W_1] + A_{12} E_{P(W)} [W_2] + \dots + A_{1D} E_{P(W)} [W_D]\end{aligned}$$
 - So, for the whole vector: $E_{P(W)} [AW] = A E_{P(W)} [W]$

Multivariate Gaussian – Covariance

- The **covariance** matrix of $X := AW + \mu$ is AA^\top

$\text{Cov}(W) = E[WW^\top] = I$ because:

- (i) $\text{Cov}(W_i, W_i) = 1$ and
- (ii) $\text{Cov}(W_i, W_{j \neq i}) = 0$ because of independence of W_i and W_j

$$\begin{aligned}\text{Cov}(X) &= E[(X - E[X])(X - E[X])^\top] = E[(AW)(AW)^\top] = E[AWW^\top A^\top] = AE[WW^\top]A^\top = \\ &= AA^\top\end{aligned}$$

Thus, the RV $X = AW + \mu$ has covariance $C = AA^\top$, where $C_{ij} = \text{Cov}(X_i, X_j)$.

Multivariate Gaussian – Different Cases

Multivariate Gaussian – Special Cases

- Diagonal matrix
- Orthogonal matrix
 - Definition: **Real square matrix Q** whose columns and rows are **orthogonal unit vectors** (i.e., orthonormal vectors) $Q Q^T = Q^T Q = \text{Identity matrix}$
 - Determinant $\det(Q)$ is either +1 or -1
 - “orthogonal” is an over-used term in mathematics
 - Rotation matrix
 - When $\det(Q) = +1$, then Q is a **rotation** matrix
 - When $\det(Q) = -1$, then Q models either reflection (called as an improper rotation) or a combination of rotation and reflection
 - “Rotation” is an over-used term (sometimes includes improper rotations) in mathematics
 - Reflection matrix
 - An orthogonal matrix that is also symmetric

Multivariate Gaussian – Special Cases

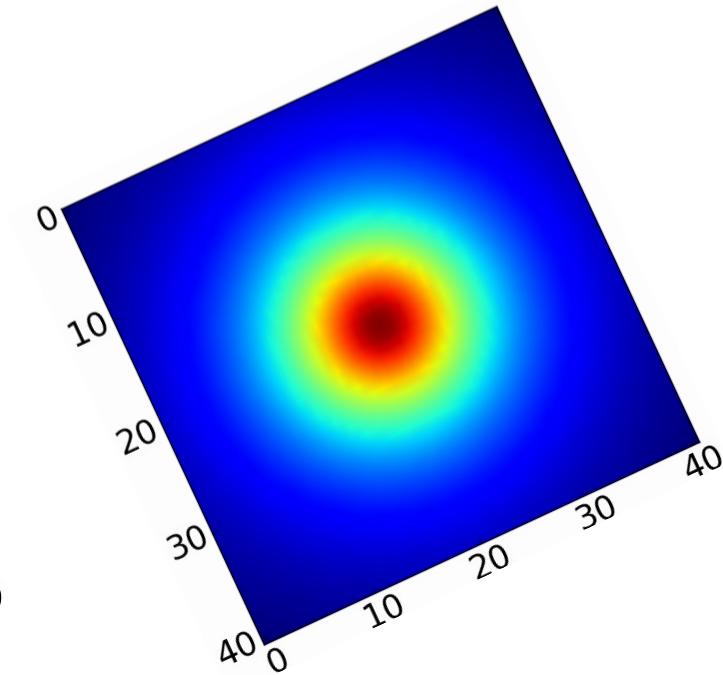
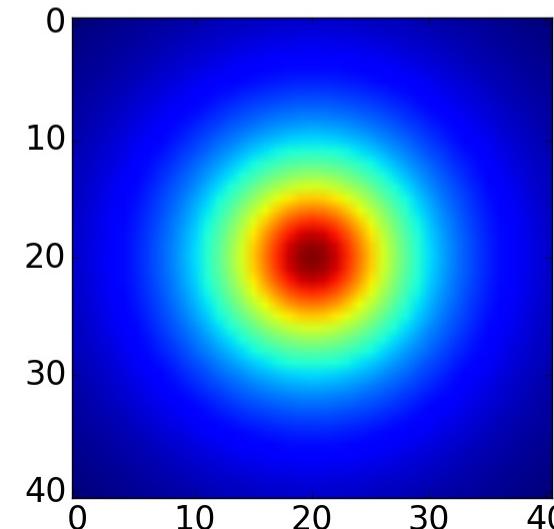
- Property (Rotation and/or Reflection):

If $\mu = 0$; and $A = R$ where R is orthogonal;
then $Y := RW$ has PDF:

$$P(y) = 1/(2\pi)^{D/2} \exp(-0.5y^\top y)$$

- Proof:

- Transformation of random vectors
 - $|\det(R)| = 1$
 - Inverse transformation is
 $W = R^\top Y$



$$\text{So, } Q(y) = (1/(2\pi)^{D/2}) \exp(-0.5(R^\top y)^\top R^\top y) = (1/(2\pi)^{D/2}) \exp(-0.5y^\top y)$$

Thus, $Y := RW$ is also a zero-mean isotropic multivariate Gaussian, just like W

Multivariate Gaussian – Special Cases

- Property (Scaling):
If $\mu = 0$; and $A = S$ square diagonal with positive entries on diagonal;
then $Y := SW$ has PDF:
- $$P(y) = (1/(2\pi)^{D/2})(1/\det(S)) \exp(-0.5y^\top (S^2)^{-1}y)$$

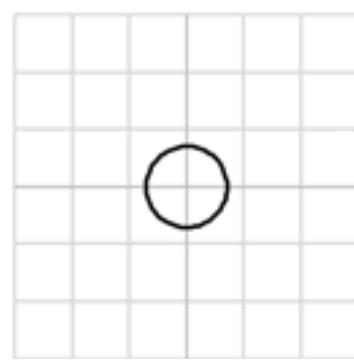
Proof: Transformation of random vectors. $|\det(S)| = \prod_d S_{dd}$

Inverse transformation is $W = S^{-1}Y$

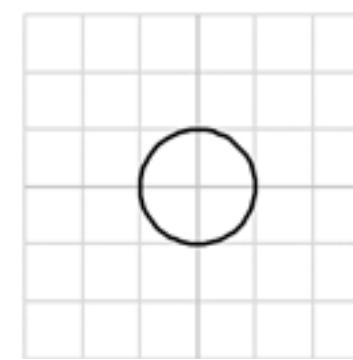
$$\text{So, } Q(y) = (1/(2\pi)^{D/2})(1/\det(S)) \exp(-0.5(S^{-1}y)^\top S^{-1}y) = (1/(2\pi)^{D/2}(1/\det(S)) \exp(-0.5y^\top (S^2)^{-1}y)$$

Thus, $Y := SW$ is a zero-mean anisotropic (axis-aligned) multivariate Gaussian.

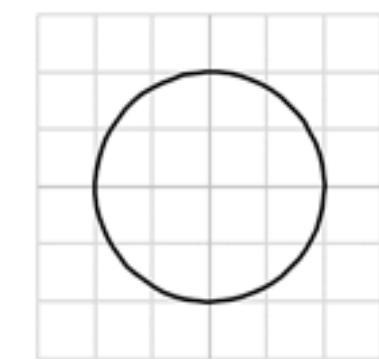
$$\begin{bmatrix} 0.5 & 0 \\ 0 & 0.5 \end{bmatrix}$$



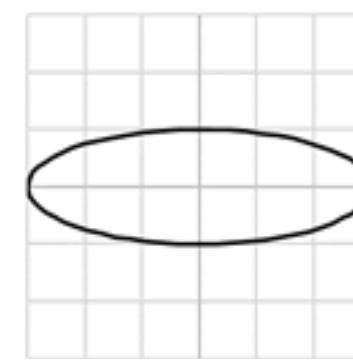
$$\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}$$



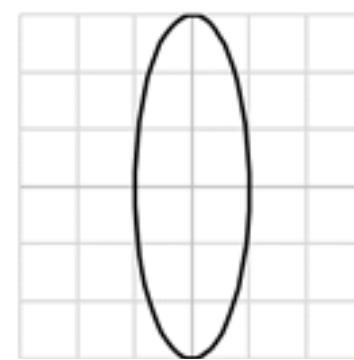
$$\begin{bmatrix} 4 & 0 \\ 0 & 4 \end{bmatrix}$$



$$\begin{bmatrix} 9 & 0 \\ 0 & 1 \end{bmatrix}$$



$$\begin{bmatrix} 1 & 0 \\ 0 & 9 \end{bmatrix}$$



Multivariate Gaussian – Special Cases

- Property (first Scaling, and then Rotation and/or Reflection):
If $\mu = 0$; $A = RS$,
then $Y := RSW$ has the PDF:

$$P(y) = (1/(2\pi)^{D/2})(1/\det(S)) \exp(-0.5y^\top (RS^2R^\top)^{-1}y)$$

Proof: Transformation of random vectors. $|\det(RS)| = \prod_d S_{dd}$.
Inverse transformation is $W = (RS)^{-1}Y = S^{-1}R^\top y$

$$\begin{aligned} \text{So, } Q(y) &= (1/(2\pi)^{D/2})(1/\det(S)) \exp(-0.5(S^{-1}R^\top y)^\top S^{-1}R^\top y) \\ &= (1/(2\pi)^{D/2}(1/\det(S)) \exp(-0.5y^\top (RS^2R^\top)^{-1}y) \end{aligned}$$

Thus, $Y := RSW$ is zero-mean rotated+reflected anisotropic multivariate Gaussian with covariance $C = RS^2R^\top$.

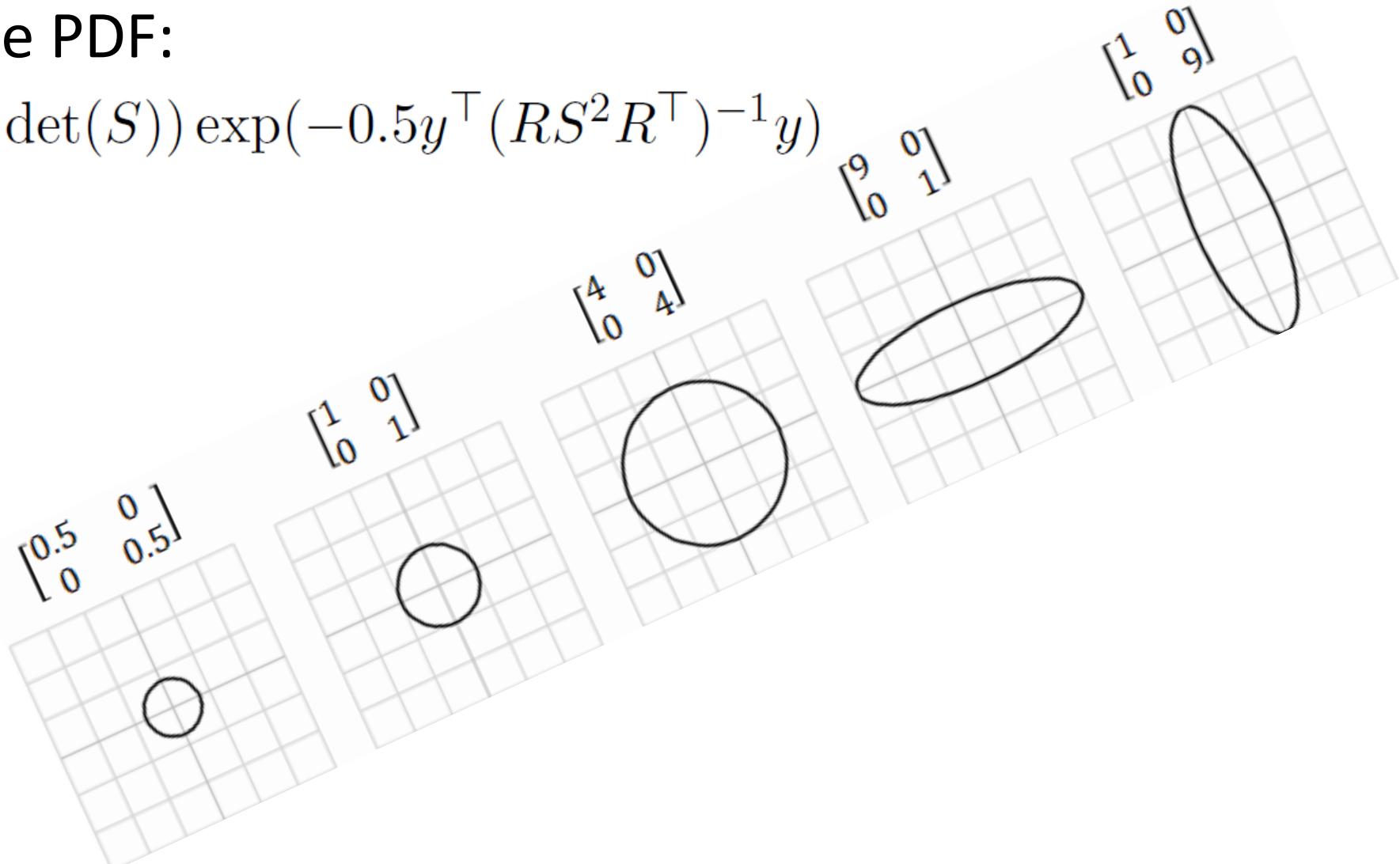
Multivariate Gaussian – Special Cases

- Property (first Scaling, and then Rotation and/or Reflection):

If $\mu = 0$; $A = RS$,

then $Y := RSW$ has the PDF:

$$P(y) = (1/(2\pi)^{D/2})(1/\det(S)) \exp(-0.5y^\top (RS^2R^\top)^{-1}y)$$



Multivariate Gaussian – General Case

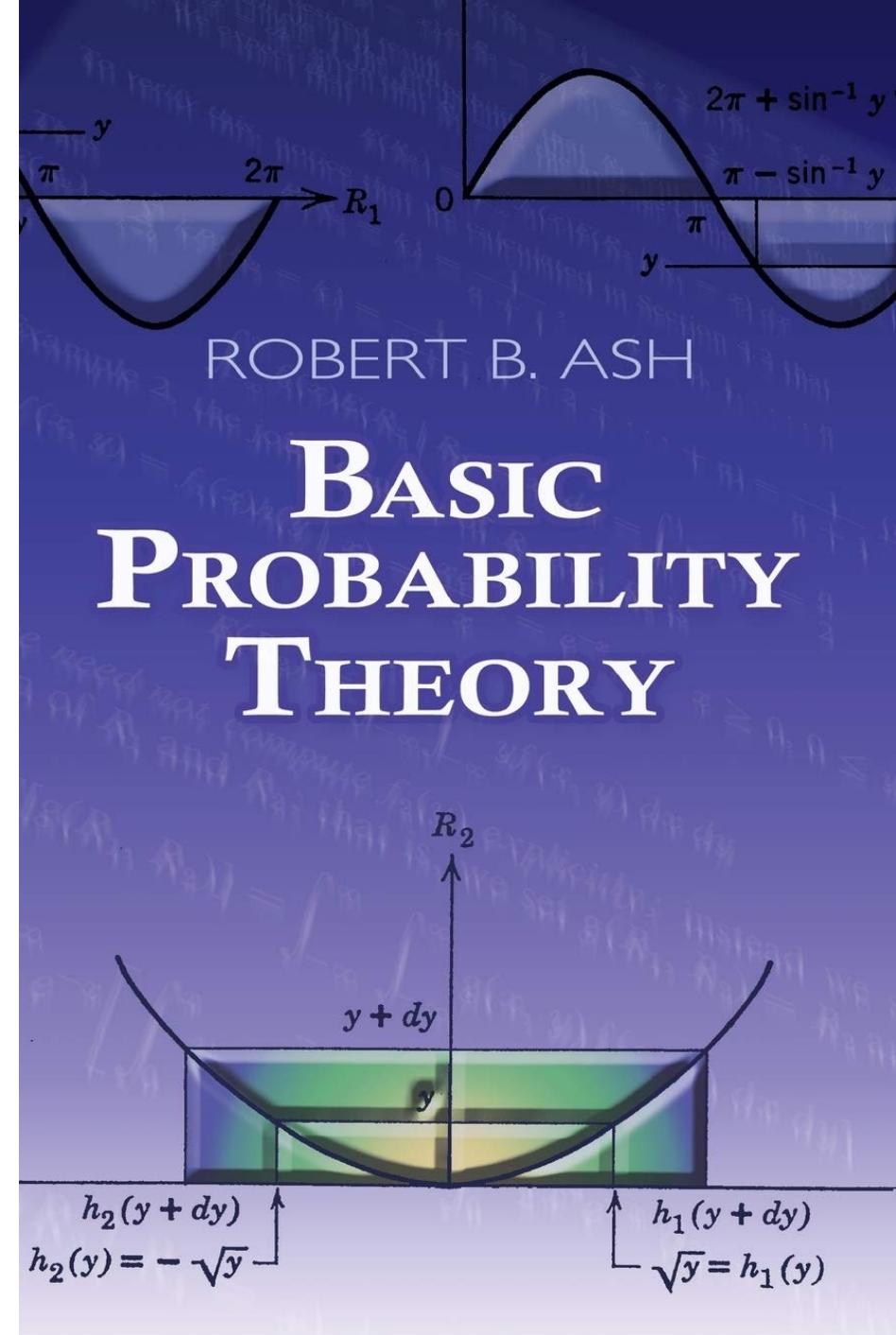
- If $X := A W$ is a multivariate Gaussian,
then $Y := X + \mu$ is a multivariate Gaussian with

$$p(y) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$$

- What are the level sets of this PDF ?
- We need some linear algebra
 - Analyze properties of covariance matrix C that is:
 - In general: real symmetric positive semi-definite
 - When $C = AA^\top$, and A is invertible, then C is positive definite

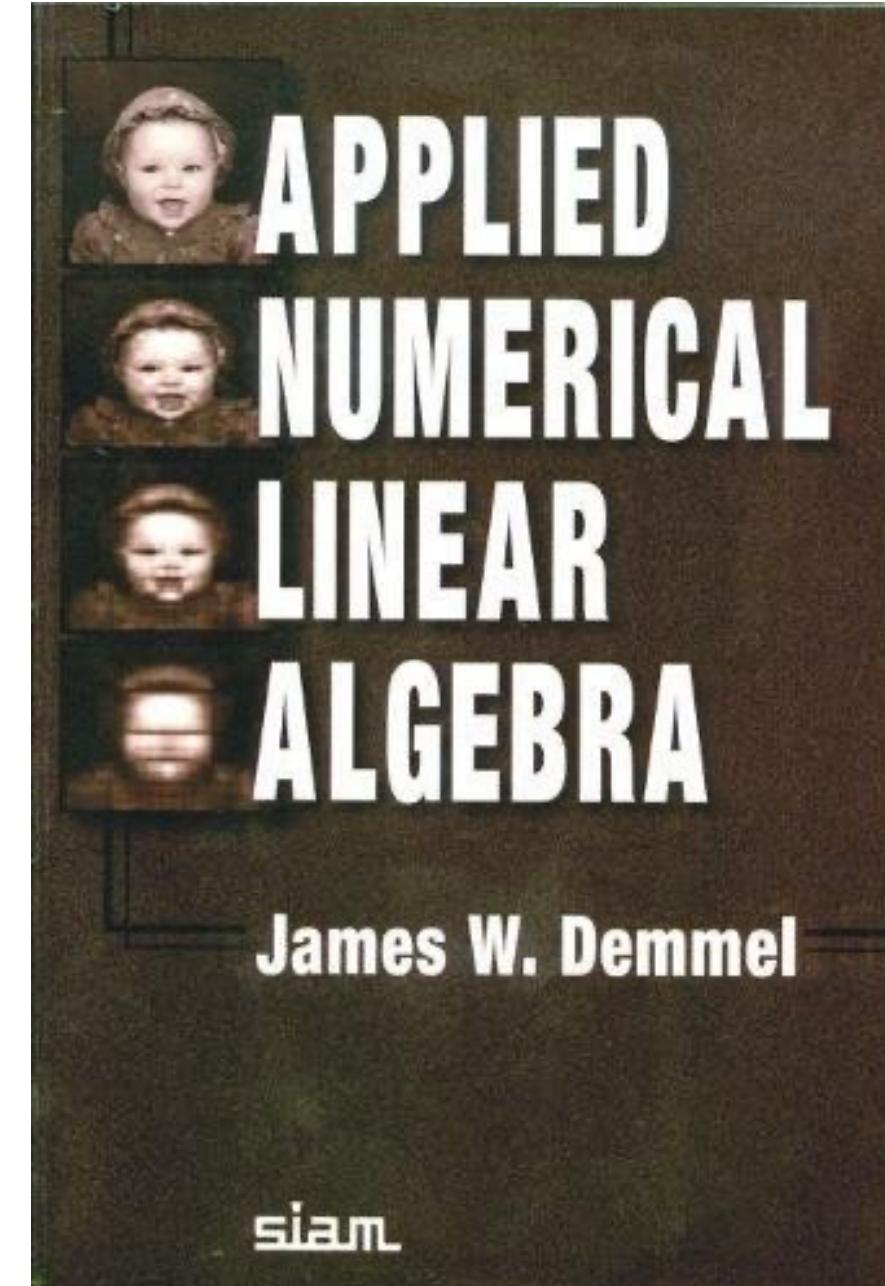
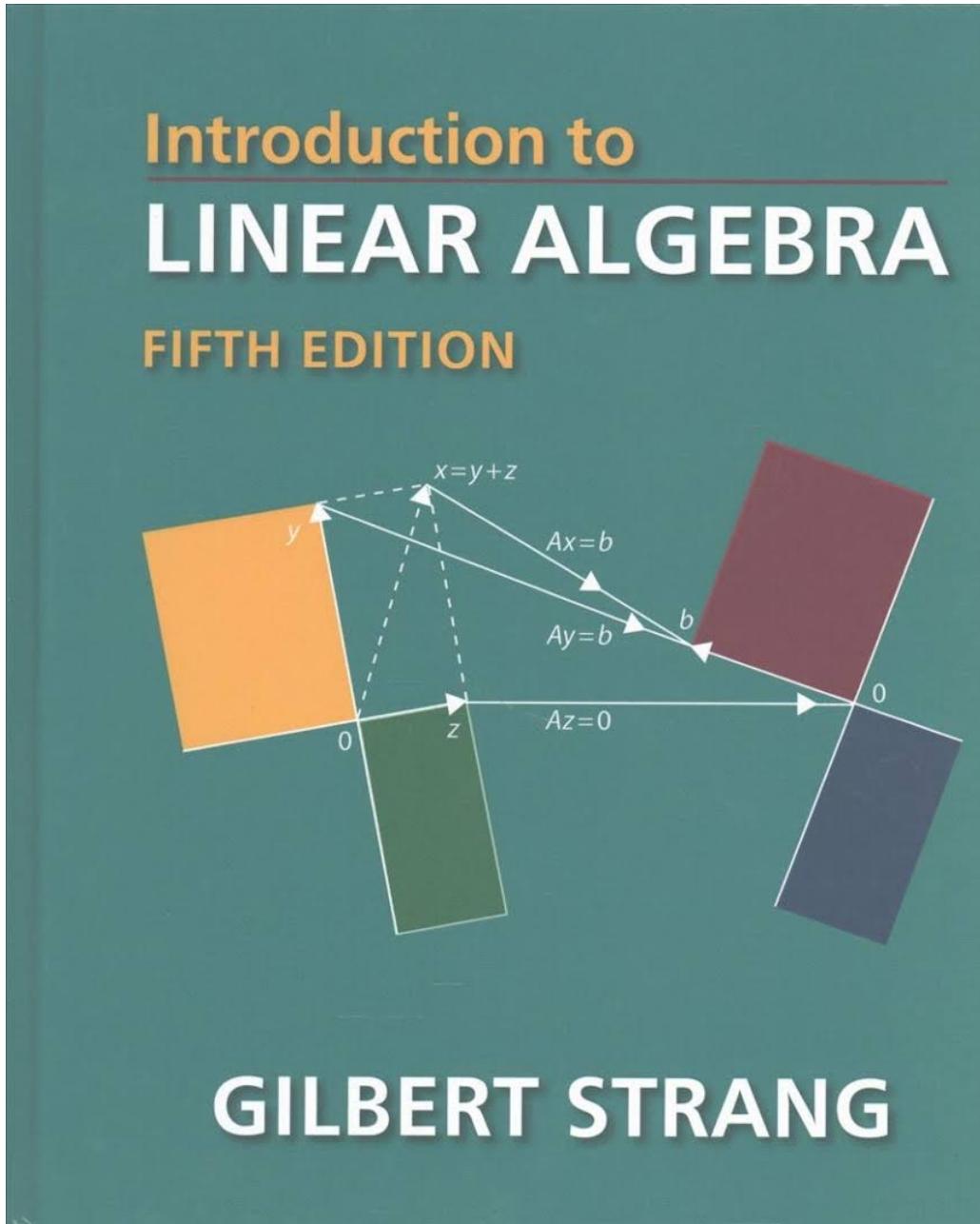
Probability and Statistics

- Reference books specifically for multivariate Gaussian
- Basic Probability Theory, by Robert Ash
 - faculty.math.illinois.edu/~r-ash/BPT.html
 - [Link 2](#)



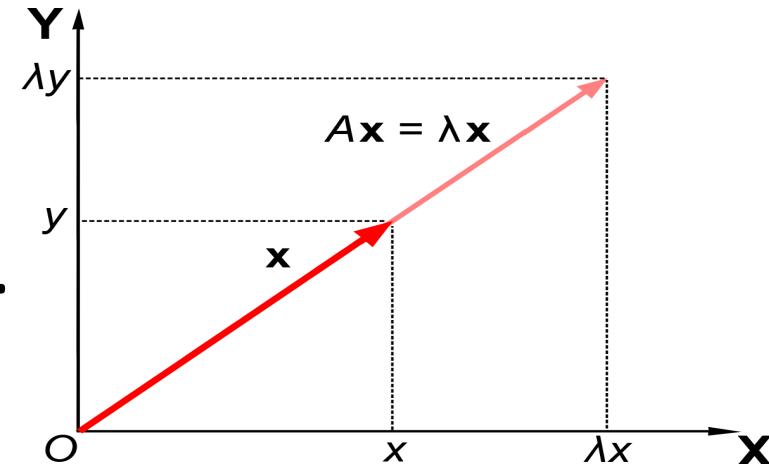
Linear Algebra

- Reference books



Linear Algebra – Eigen Decomposition

- Eigenvalue and Eigenvector
- For any square NxN matrix A,
an **eigenvector** is a non-zero vector ‘v’ s.t. $Av = \lambda v$.
Then, λ is the associated **eigenvalue**
- Square matrix A is **diagonalizable** if it is “similar” to a diagonal matrix,
i.e., if there exists an invertible matrix P and a diagonal matrix D
such that $P^{-1}AP = D$



Linear Algebra – Eigen Decomposition

- If A is diagonalizable,
then it has
 N linearly-independent eigenvectors
 - The eigenvectors needn't be orthogonal to each other

If a matrix A can be diagonalized, that is,

$$P^{-1}AP = \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix},$$

then:

$$AP = P \begin{pmatrix} \lambda_1 & 0 & \dots & 0 \\ 0 & \lambda_2 & \dots & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & \dots & \lambda_n \end{pmatrix}.$$

Writing P as a **block matrix** of its column vectors $\vec{\alpha}_i$

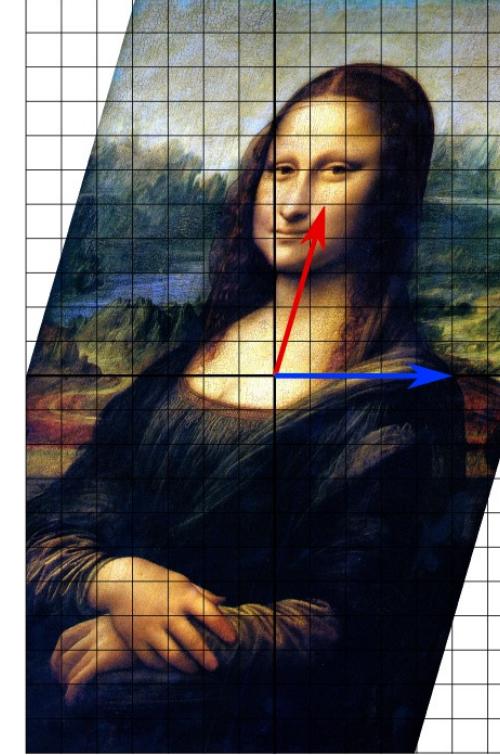
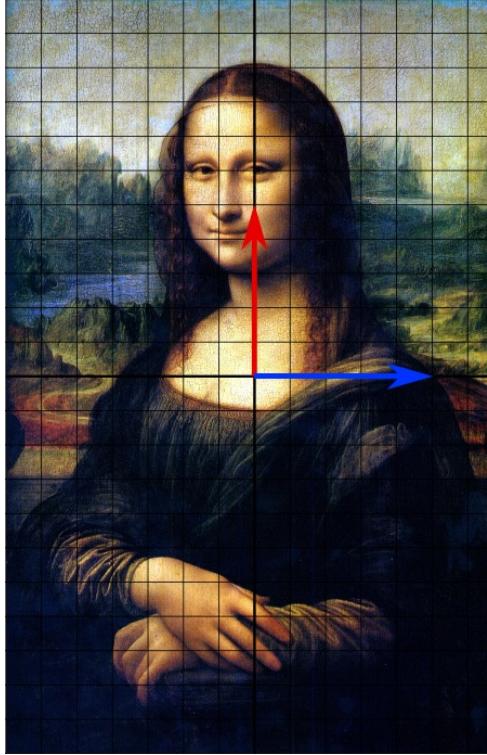
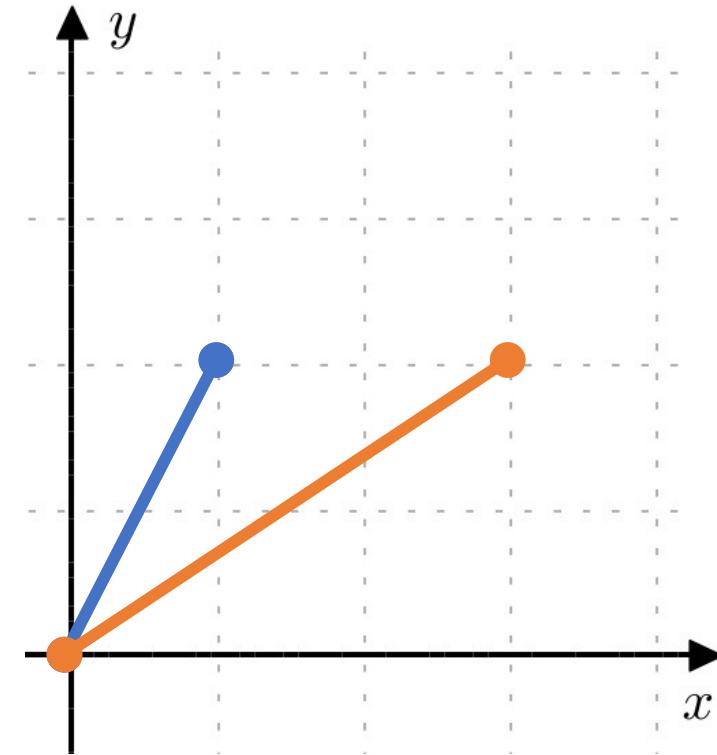
$$P = (\vec{\alpha}_1 \quad \vec{\alpha}_2 \quad \dots \quad \vec{\alpha}_n),$$

the above equation can be rewritten as

$$A\vec{\alpha}_i = \lambda_i \vec{\alpha}_i \quad (i = 1, 2, \dots, n).$$

Linear Algebra – Eigen Decomposition

- Invertible doesn't imply diagonalizable
 - A non-diagonalizable matrix is called a defective matrix
 - e.g., 2x2 matrix A as shown. $B = \text{inv}(A)$. $[V D] = \text{eig}(A)$
 - Doesn't have a complete basis of eigenvectors
 - Intuition: Action of matrix is to map vector (x,y) to $(x+y,y)$
So, any eigenvalue must be 1, any eigenvector must have $y=0$



```
A =
1 1
0 1
B =
1 -1
0 1
V =
1.0000 -1.0000
0 0.0000
D =
1 0
0 1
```

Linear Algebra – Eigen Decomposition

- Diagonalizable doesn't imply invertible
 - e.g., some eigenvalues can be zero

Linear Algebra – Eigen Decomposition

- Eigenvalue and Eigenvector
- For any square NxN matrix A,
an **eigenvector** is a non-zero vector ‘v’ s.t. $Av = \lambda v$.
Then, λ is the associated **eigenvalue**
- Theorem:
Every real symmetric matrix (e.g., covariance C) is diagonalizable
 - There exists an invertible matrix Q such that $Q^{-1} C Q$ is diagonal
 - This implies C has N linearly-independent eigenvectors
- Theorem:
Every real symmetric matrix (e.g., covariance C) is diagonalizable by an orthogonal matrix
 - There exists an orthogonal matrix Q such that $Q^T C Q$ is diagonal

Linear Algebra – Eigen Decomposition

- **Spectral Theorem:** If A is a **real symmetric** $N \times N$ matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors

Linear Algebra – Eigen Decomposition

- In general:
 - If $\{u,v\}$ are distinct/non-parallel eigenvectors associated with same eigenvalue λ , then any vector within span of $\{u,v\}$ is also an eigenvector with eigenvalue λ .
 - Because $A(au + bv) = aAu + bAv = a\lambda u + b\lambda v = \lambda(au + bv)$
 - The above equation holds for real/complex-valued a,b, u,v, λ, A
 - If $\{u,v\}$ are distinct eigenvectors with distinct eigenvalues $\{\lambda_1,\lambda_2\}$, then action of A on any vector within span of $\{u,v\}$ outputs a vector within the span of $\{u,v\}$.
 - Because $A(au + bv) = aAu + bAv = a\lambda_1 u + b\lambda_2 v = (a\lambda_1)u + (b\lambda_2)v$
 - Notation: $(.)^*$ = conjugate and $(.)^T$ = transpose (e.g., converts col to row)
 - For a real-symmetric matrix A
 - Inner product of Ax and y = inner product of x and Ay because $(Ax)^T y^* = x^T A^T y^* = x^T A y^* = x^T A^* y^* = x^T (Ay)^*$ (A is "self-adjoint")
 - If v is a eigenvector and if u is orthogonal to v (i.e., if $u^T v^* = 0$), then action of A on u produces a vector also orthogonal to v (A maintains orthogonality)
 - Because $(Au)^T v^* = u^T (A^T v^*) = u^T (Av)^* = u^T \lambda^* v^* = 0$

Linear Algebra – Eigen Decomposition

- **Spectral Theorem:** If A is a **real symmetric** $N \times N$ matrix, then A has N **real eigenvalues** with N **real-valued orthogonal eigenvectors**
 - First, we show that A has **all real eigenvalues** (i.e., A cannot have a complex-valued eigenvalue)

Proof:

Let $v \in \mathbb{C}^N$ be a (unit-norm) eigenvector with eigenvalue $\lambda \in \mathbb{C}$.

Then, $\lambda v^\top v^* = (\lambda v)^\top v^* = (Av)^\top v^* = v^\top (Av^*)$ (because A is symmetric)

$= v^\top (Av)^*$ (because A is real)

$= v^\top (\lambda v)^* = \lambda^* v^\top v^*$

Because $v^\top v^* = 1 \neq 0$, we have $\lambda = \lambda^*$ that implies that λ is real

Linear Algebra – Eigen Decomposition

- **Spectral Theorem:** If A is a **real symmetric** $N \times N$ matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors
 - Let us analyze the **special/simple case** where **all real eigenvalues** = λ
 - Then, for all v in R^N , linear operator A 's action on v simply scales v by factor λ
 - We can exactly model such an operator A simply by a diagonal matrix $\lambda I_{N \times N}$
 - Or model A as $A = Q(\lambda I_{N \times N})Q^T$,
where Q is any orthogonal basis for R^N with real-valued column vectors
 - Thus, $Q^T A Q$ leads to a diagonal matrix
 - So, the N columns of Q are N real-valued eigenvectors of A
 - Of course, Q isn't unique

Linear Algebra – Eigen Decomposition

- **Spectral Theorem:** If A is a **real symmetric** $N \times N$ matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors

- For real-symmetric A ,
eigenvectors corresponding to distinct (real) eigenvalues
are orthogonal
- Proof:

- Let A have eigenvector v_1 with real eigenvalue λ_1
- Let A have eigenvector v_2 with real eigenvalue $\lambda_2 \neq \lambda_1$
- Then, $\lambda_1 v_1^T v_2^* = (\lambda_1 v_1)^T v_2^* = (A v_1)^T v_2^*$
- $= v_1^T (A v_2)^*$ (because A is symmetric)
- $= v_1^T (A v_2)^*$ (because A is real)
- $= v_1^T (\lambda_2 v_2)^* = v_1^T \lambda_2 v_2^*$ (because λ_2 is real)
- $= \lambda_2 v_1^T v_2^*$
- Because $\lambda_2 \neq \lambda_1$, we get v_1 orthogonal to v_2

Notation:

$(\cdot)^T$ = transpose

Linear Algebra – Eigen Decomposition

- **Spectral Theorem:** If A is a **real symmetric** $N \times N$ matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors

We show that, for A , we can pick N **real-valued** eigenvectors v_n , each corresponding to a real eigenvalue λ_n

If A has a complex eigenvector $v \in \mathbb{C}^N$ for a real eigenvalue $\lambda \in \mathbb{R}$, then let $v = a + ib$, where $a \in \mathbb{R}^N$ and $b \in \mathbb{R}^N$

Then, having $Av = \lambda v$ and real λ and real A ,

we get $Aa = \lambda a$ and $Ab = \lambda b$, where a and b are real-valued eigenvectors

Linear Algebra – Eigen Decomposition

- **Spectral Theorem:** If A is a **real symmetric** $N \times N$ matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors
 - How do we know that real-valued eigenvectors exist for A ?
 - Let real-positive scalar $b := \max_{y \in \mathbb{R}^N : \|y\|_2=1} \|Ay\|_2$,
 - Let real-valued vector $x := \arg \max_{y \in \mathbb{R}^N : \|y\|_2=1} \|Ay\|_2$
 - If there are multiple such unit-norm vectors y , then we pick one of them as x
 - Thus, action of A scales norm of any vector by at-most b times
 - Then $b^2 = (Ax)^T(Ax) = x^T(AAx)$
 - RHS can take value b^2 only if x is parallel to AAx (and then $AAx = b^2x$)
 - Then we can claim that:
 1. Either Ax is parallel to x ; so $v_1 := x$ is a **real-valued eigenvector**, with **real eigenvalue** b
 2. Or $v_1 := Ax+bx$ (**non-zero**) is a **real-valued eigenvector**, with **real eigenvalue** b , because $A(Ax + bx) = AAx + bAx = b^2x + bAx = b(bx + Ax)$

Linear Algebra – Eigen Decomposition

- **Spectral Theorem:** If A is a **real symmetric NxN** matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors

- How do we know that real-valued eigenvectors exist for A ?

- **Repeat the following:**

- Let real-positive scalar $c := \max_{y \in R^N, y^\top v_1 = 0 : \|y\|_2 = 1} \|Ay\|_2 ,$

- Let real-valued vector $x := \arg \max_{y \in R^N, y^\top v_1 = 0 : \|y\|_2 = 1} \|Ay\|_2$

- Then $c^2 = (Ax)^\top (Ax) = x^\top (AAx)$

- RHS can take value c^2 only if x is parallel to AAx (and then AAx = c^2)

- Then we can claim that:

1. Either Ax is parallel to x; so x is a real-valued eig.vector (**orthogonal to v_1**) (eig.value c)
2. Or $(Ax+cx)$ is a real-valued eig.vector of A (**orthogonal to v_1 , as both Ax and x are**), because $A(Ax + cx) = AAx + cAx = c^2x + cAx = c(cx + Ax)$

Linear Algebra – Eigen Decomposition

- Every NxN real symmetric **positive definite** (SPD) matrix M (e.g., covariance matrix C) has an eigen-decomposition with all eigenvalues as **positive**
- Proof:
 - Let eigen decomposition for **real symmetric** matrix M be: $M = Q D Q^T$
 - Where Q is real orthogonal and D is real diagonal
 - Then, $v^T M v = v^T Q D Q^T v = u^T D u$, where $u := Q^T v$ (simply “rotated” v)
 - For a **PD** matrix M, $v^T M v$ must be positive for every non-zero v
 - So, $u^T D u$ must be positive for every non-zero u
 - So, all values on diagonal of D must be positive

Multivariate Gaussian – Level Sets

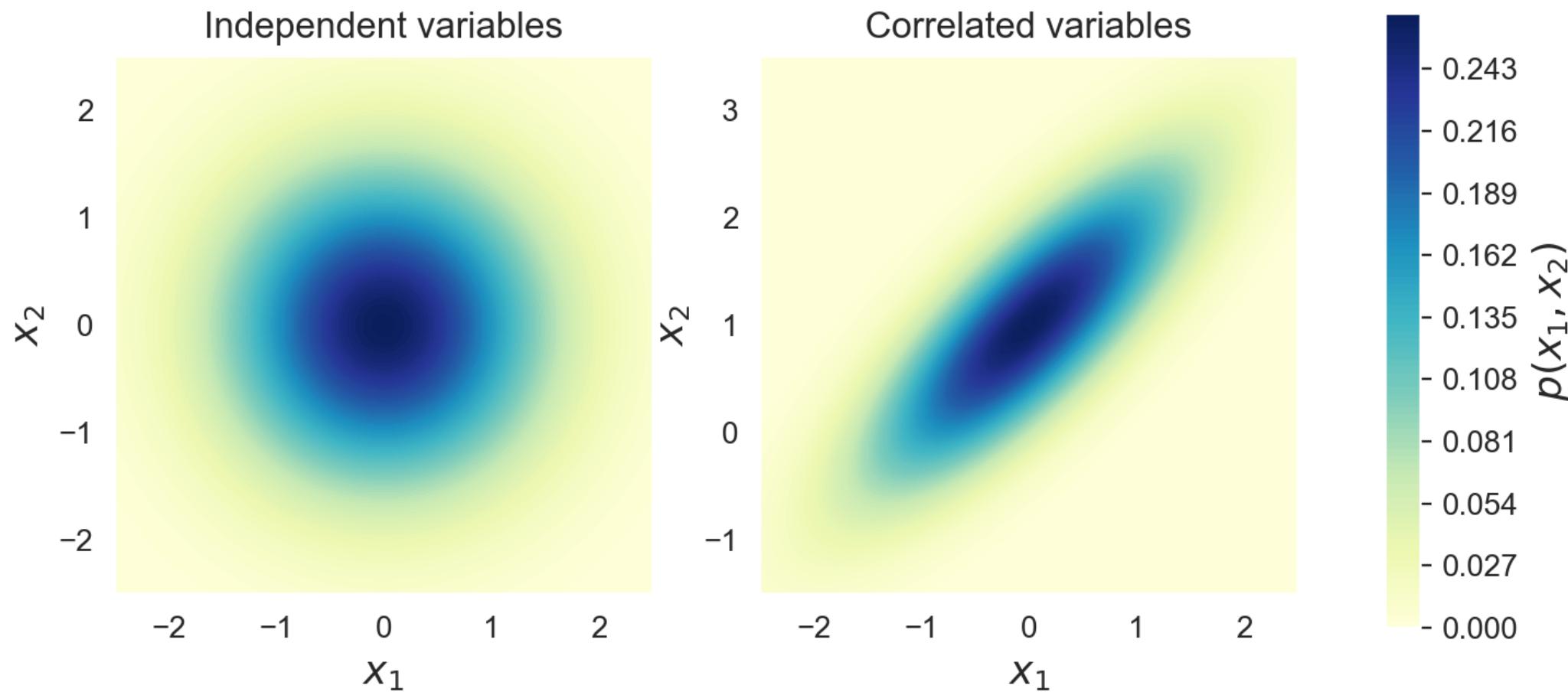
- If $X = A W$ is a multivariate Gaussian,
then $Y = X + \mu$ is a multivariate Gaussian with

$$p(y) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$$

- What are the **level sets** of this PDF ?
- Let $C = Q D Q^\top$. Then, $C^{-1} = Q D^{-1} Q^\top$ that is also SPD
- Each level set satisfies $(y-\mu)^\top C^{-1} (y-\mu) = a$, where $a >= 0$
 - Because C^{-1} is SPD; ‘ a ’ becomes zero iff $y=\mu$
- So, $(y-\mu)^\top Q D^{-1} Q^\top (y-\mu) = a$
- Change to roto-reflected coordinate system represented by orthogonal basis Q
 - Where y maps to $y' = Q^\top y$, and μ maps to $\mu' = Q^\top \mu$
- Then, $(y'-\mu')^\top D^{-1} (y'-\mu') = a$, which is a hyper-ellipsoid:
 - In roto-reflected coordinate system, center is at μ' and axes are along cardinal axes
 - Whose half-lengths of axes are square root of diagonal elements in D^{-1}

Multivariate Gaussian – Level Sets

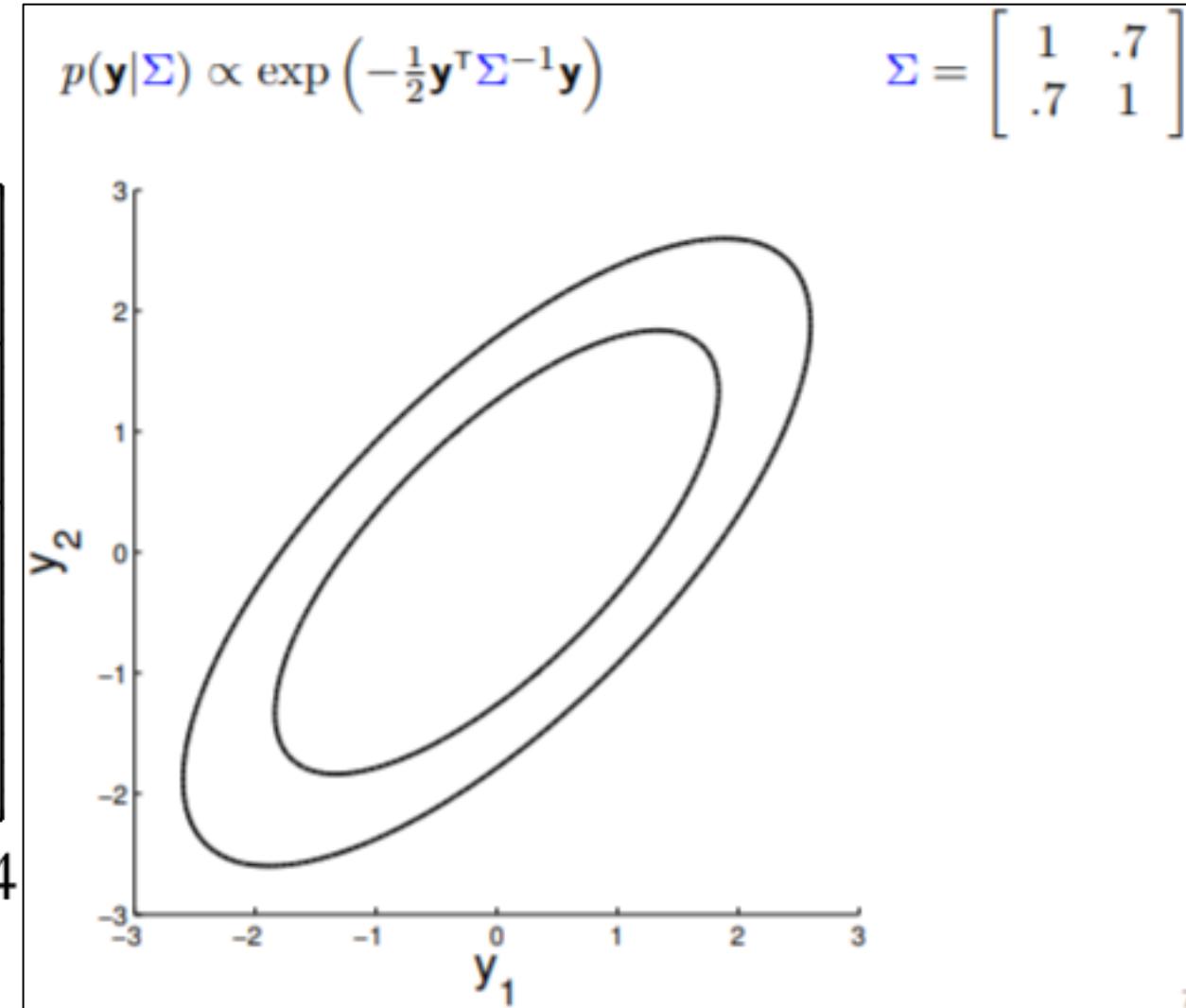
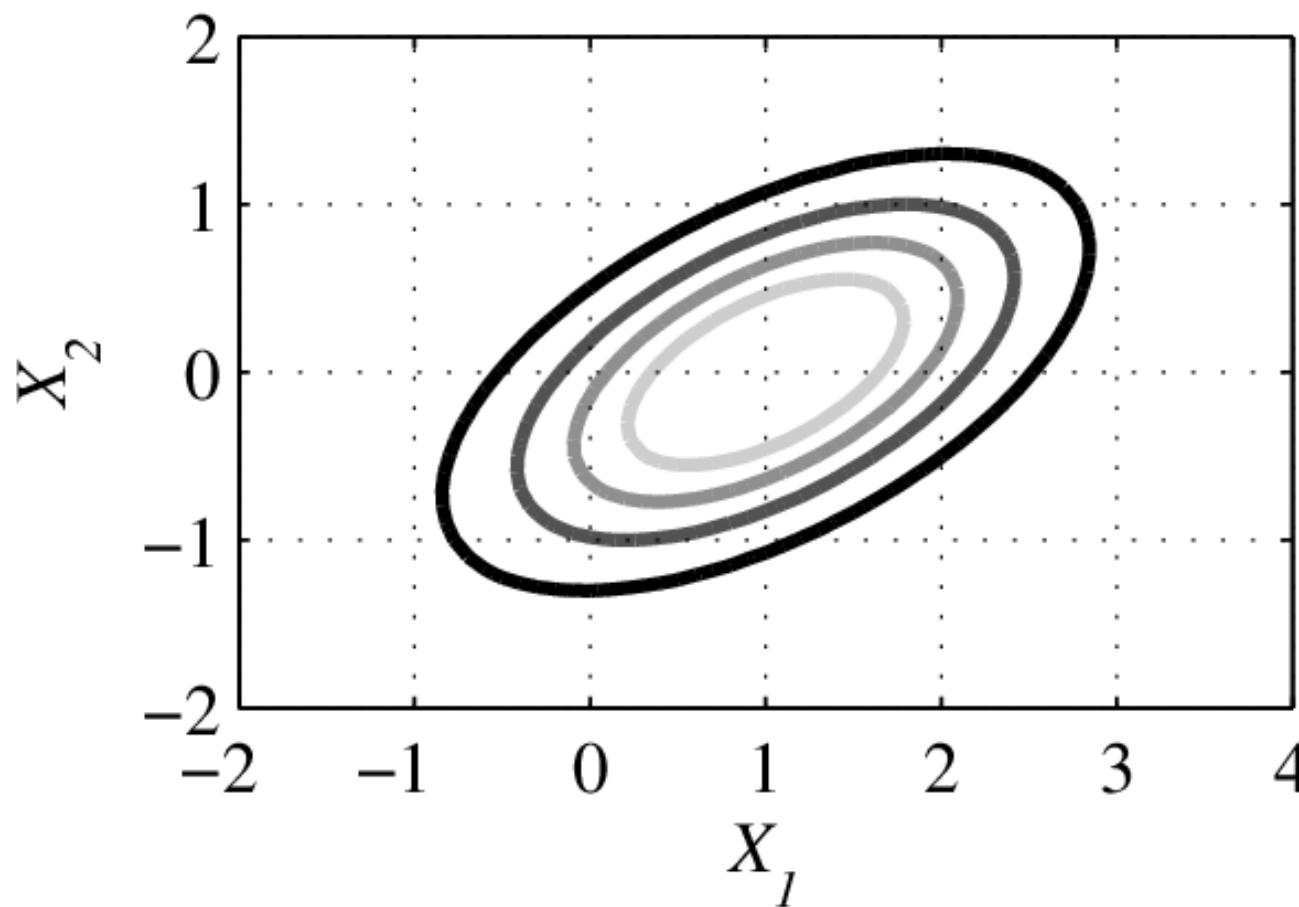
- If $X = AW$ is a multivariate Gaussian, then $Y = X + \mu$ is a multivariate Gaussian with $p(y) = \frac{1}{(2\pi)^D/2|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$
- What are the **level sets** of this PDF ?



Multivariate Gaussian – Level Sets

- If $X = A W$ is a multivariate Gaussian, then $Y = X + \mu$ is a multivariate Gaussian with $p(y) = \frac{1}{(2\pi)^D/2|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$

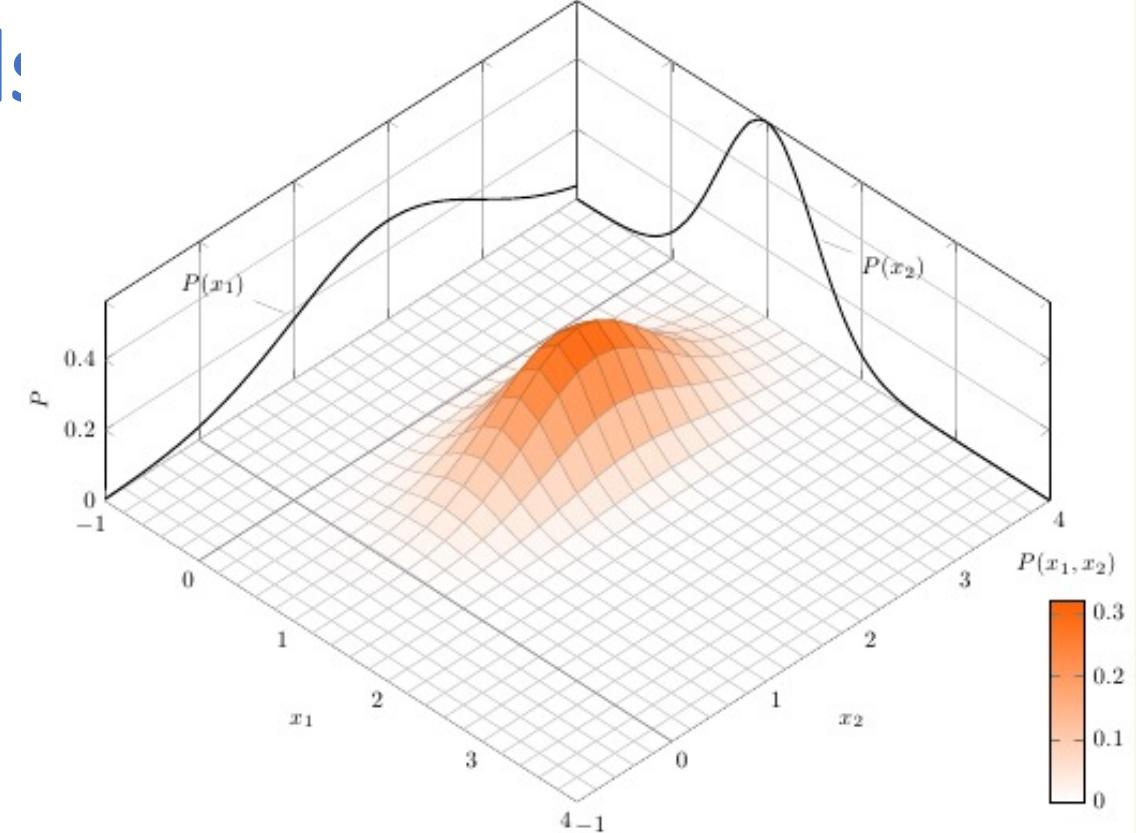
- What are the **level sets** of this PDF ?



Multivariate Gaussian – Marginals and Conditionals

Multivariate Gaussian – Marginal:

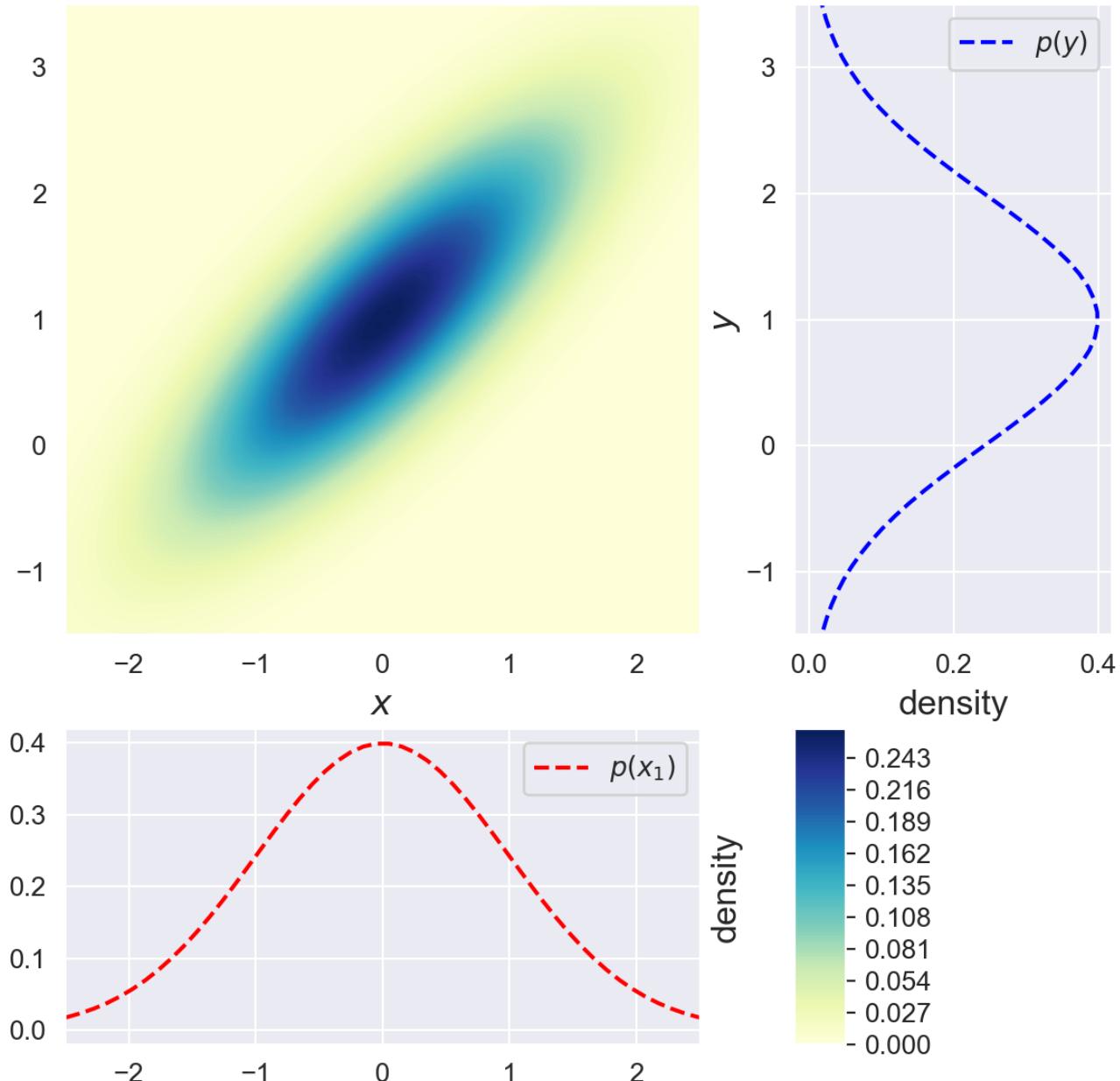
- Marginal PDFs
- Property: The 1D marginal PDF of multivariate Gaussian X , for any single variable, is (univariate) Gaussian
- Proof:
 - From the definition, we know that:
 - (1) $X_d = \mu_d + \sum_n A_{dn} W_n$, where W_n are i.i.d. standard Normal
 - (2) transformations of scaling and/or translation on a univariate Gaussian RV lead to another univariate Gaussian RV
 - (3) sum of 2 independent univariate Gaussian RVs leads to another univariate Gaussian RV



Multivariate Gaussian – Marginals

- **Marginal PDFs**
- Property: The 1D marginal PDF of multivariate Gaussian X , for any single variable, is (univariate) Gaussian

Marginal distributions

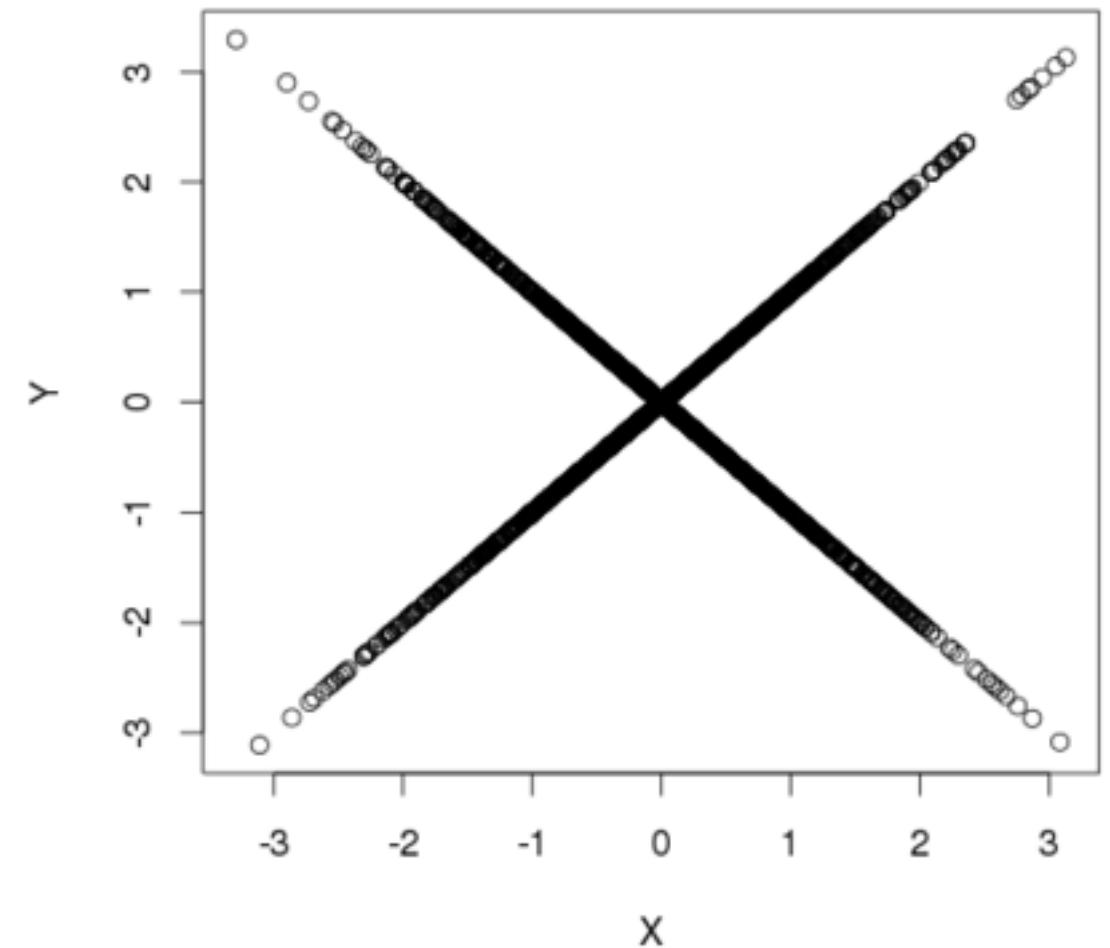


Multivariate Gaussian – Marginals

- Marginal PDFs
- Property: Marginal PDFs of multivariate Gaussian X in N dimensions, over any chosen **subset** of the variables (subset size $M < N$), are (multivariate) Gaussian
- Proof:
 - Choose transformation B as a projection matrix of size $M \times N$, where $M < N$
 - Each row has all zeros except a 1 at one position
 - e.g., row $[1 \ 0 \ \dots \ 0]$ will select the first component of X
 - If we consider multivariate Gaussian $X := AW + \mu$, where A is invertible, then $BX = (BA)W + (B\mu)$
 - Note: Because A is invertible (full rank), BA has rank M
 - By definition, BX is also multivariate Gaussian
 - Mean = $B\mu$, Covariance = $(BA)(BA)^T = BAA^T B^T = BCB^T = C'$, where C' is a square sub-matrix of C corresponding to the chosen M variables

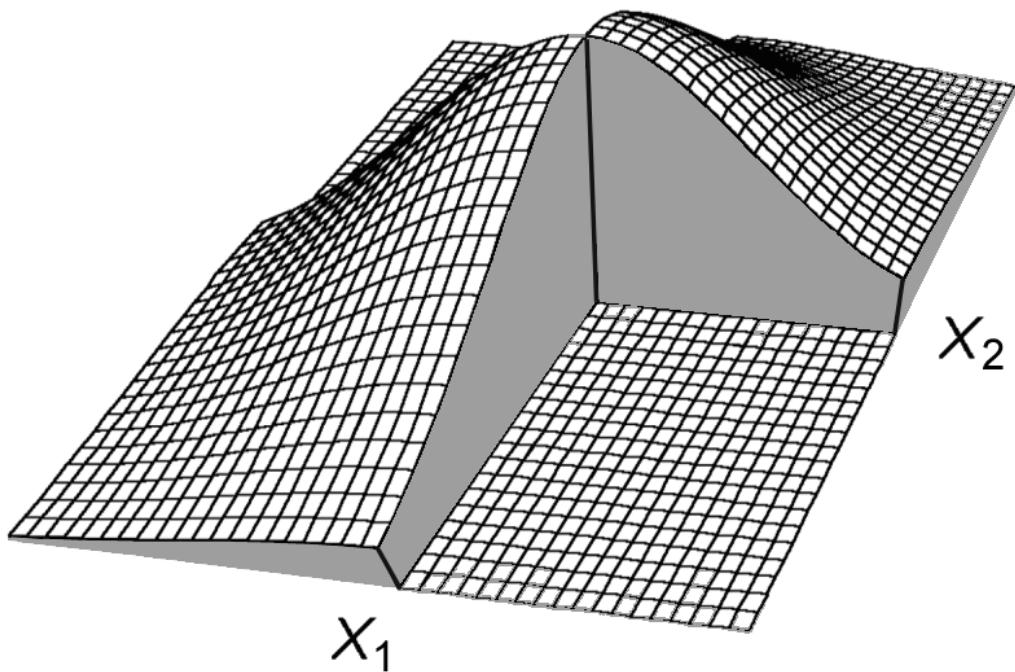
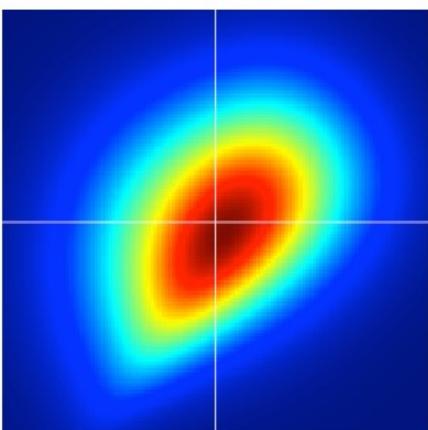
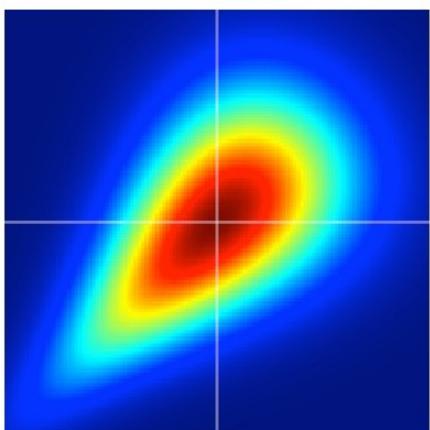
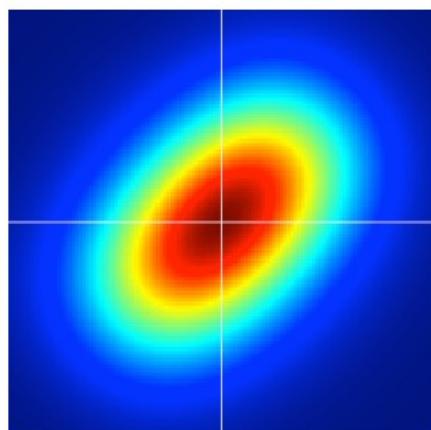
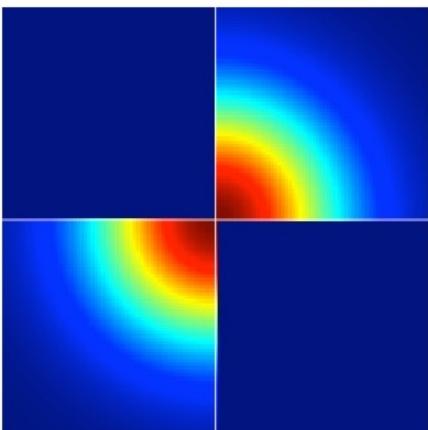
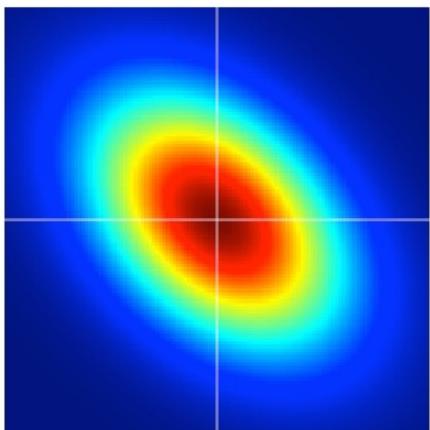
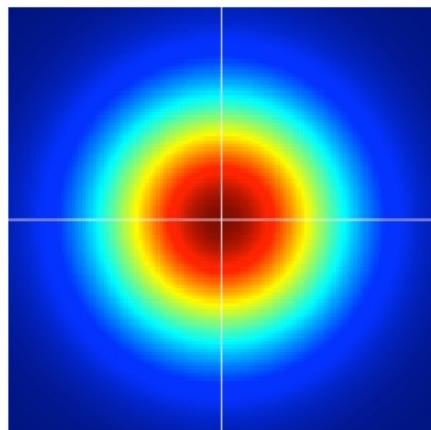
Multivariate Gaussian – Marginals

- Marginal PDFs being Gaussian doesn't imply joint PDF is multivariate Gaussian
- Example
 - Let X be a standard Normal
 - Let $Y = X(2B - 1)$ where B is Bernoulli with parameter 0.5
- More examples
 - https://en.wikipedia.org/wiki/Normally_distributed_and_uncorrelated_does_not_imply_independent



Multivariate Gaussian – Marginals

- Marginal PDFs being Gaussian doesn't imply joint PDF is multivariate Gaussian
 - Only top-row left, top-row middle, bottom-row left are bivariate Gaussian
 - All marginals are Gaussian

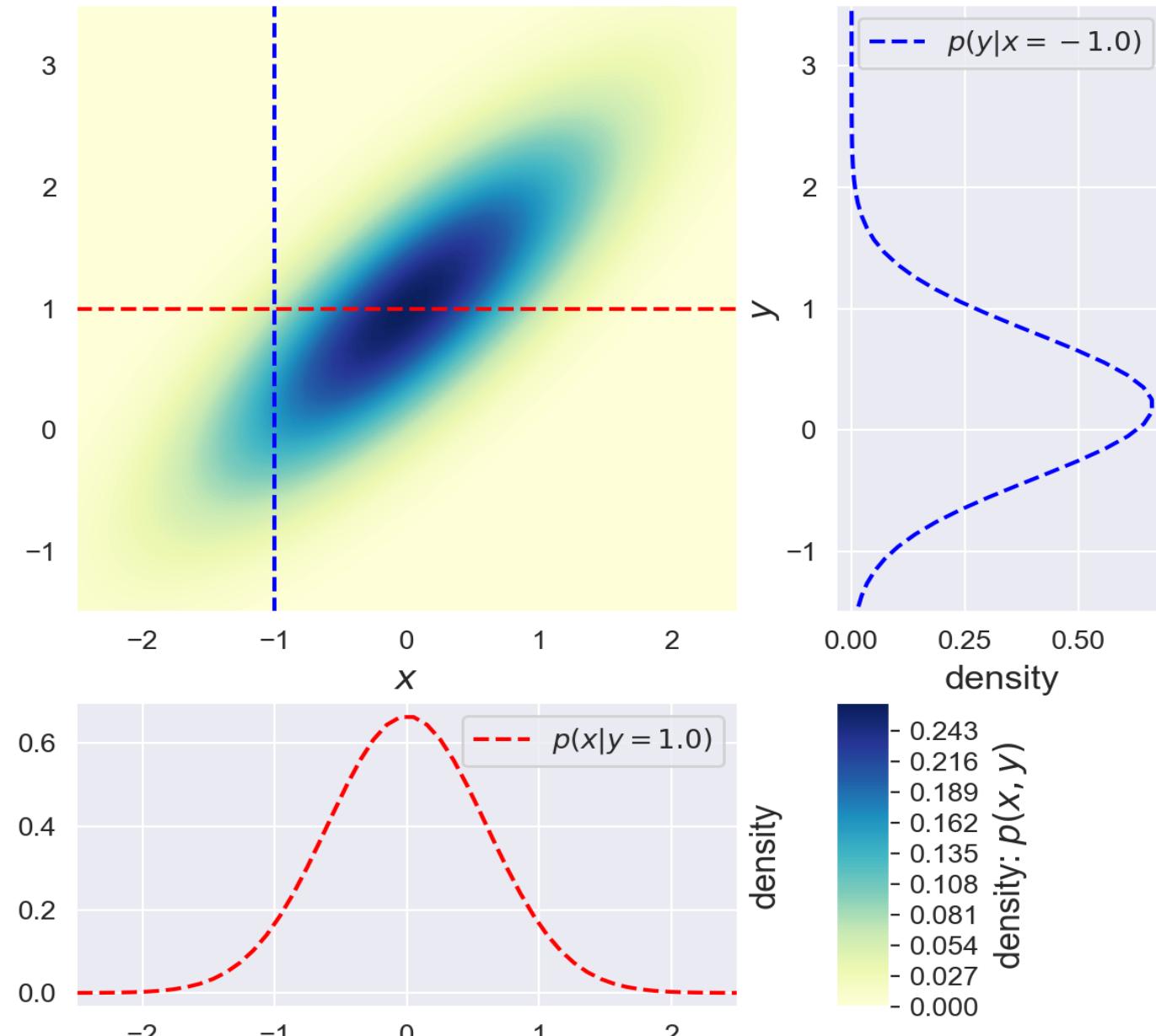


Multivariate Gaussian – Conditionals

- **Conditional PDFs**

- If multivariate Gaussian X is partitioned into X_1 and X_2 , then conditional PDF $P(X_1|X_2=x_2)$ is also a multivariate Gaussian
- $P(X_1|X_2=x_2) = P(X_1, X_2=x_2) / P(X_2=x_2)$

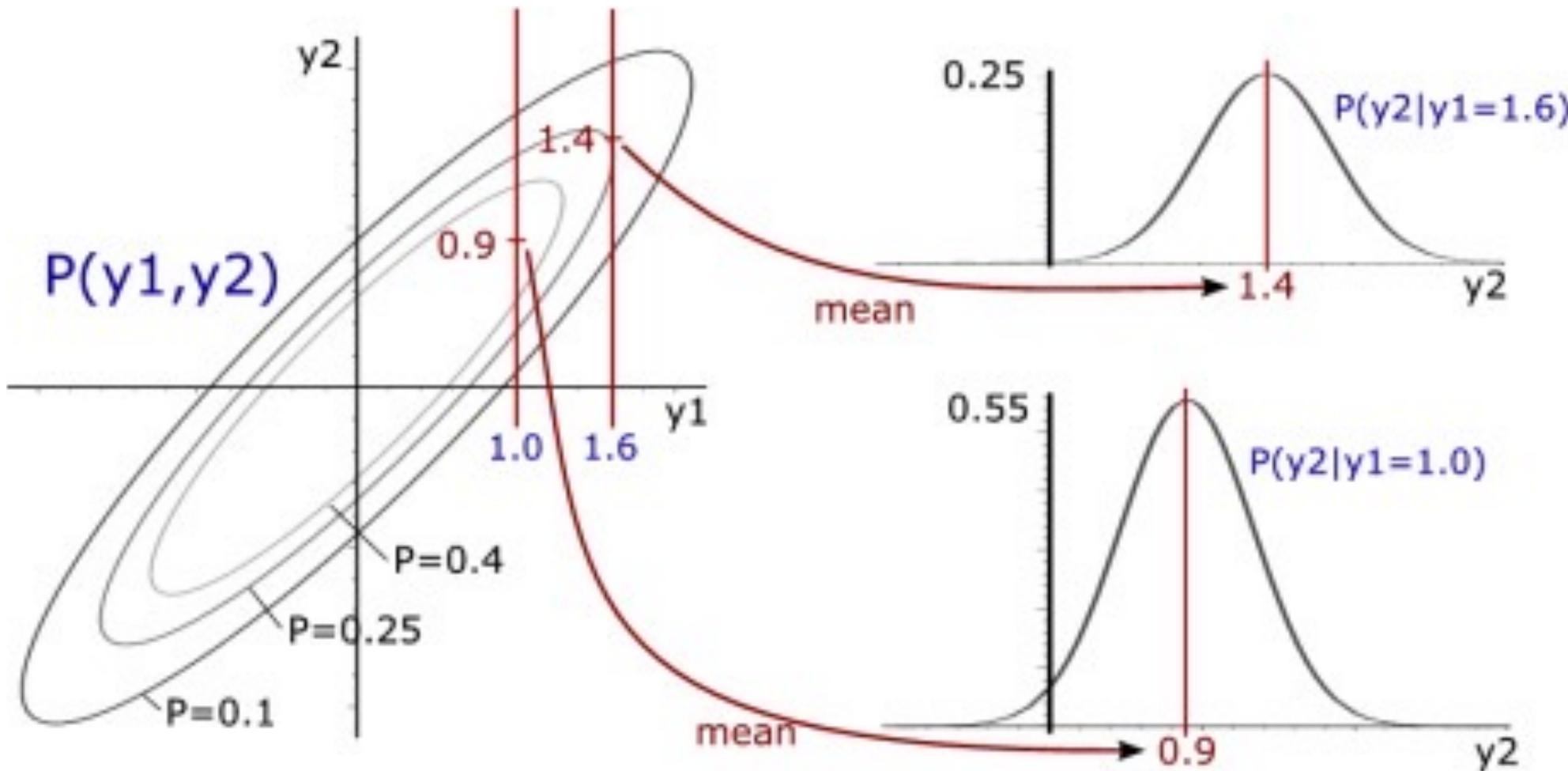
Conditional distributions



Multivariate Gaussian – Conditionals

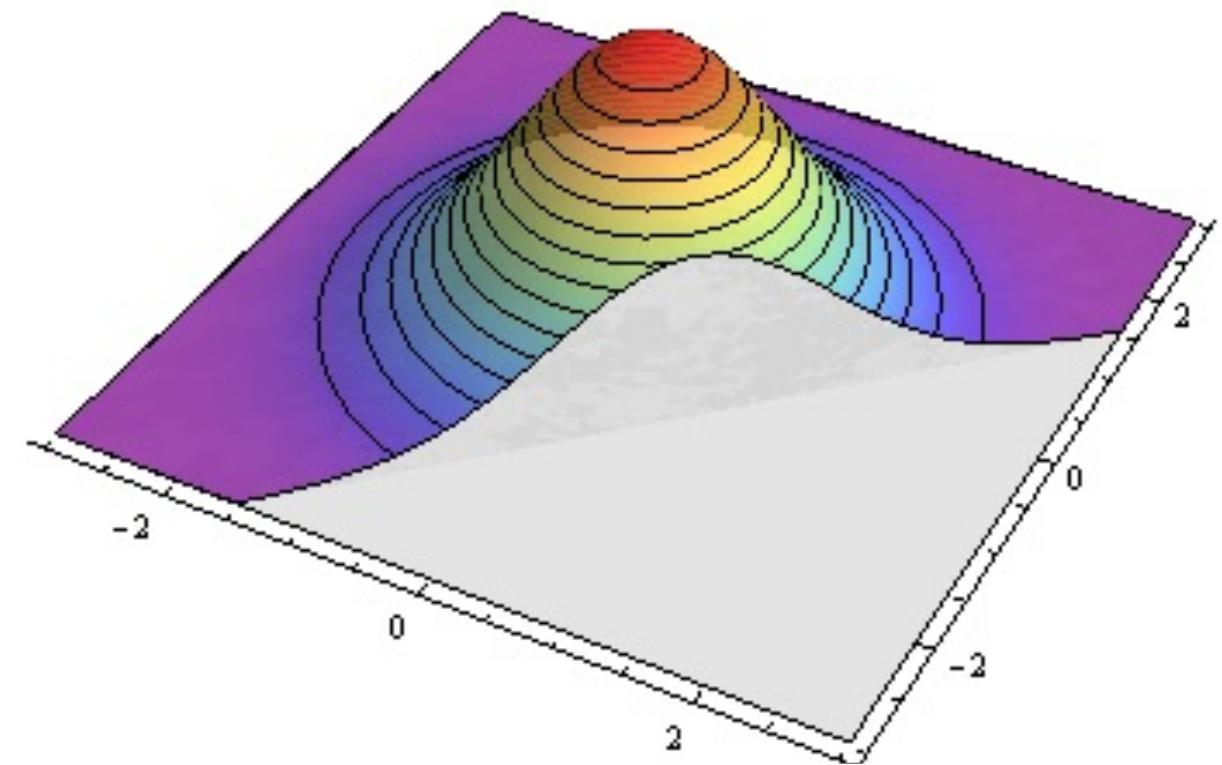
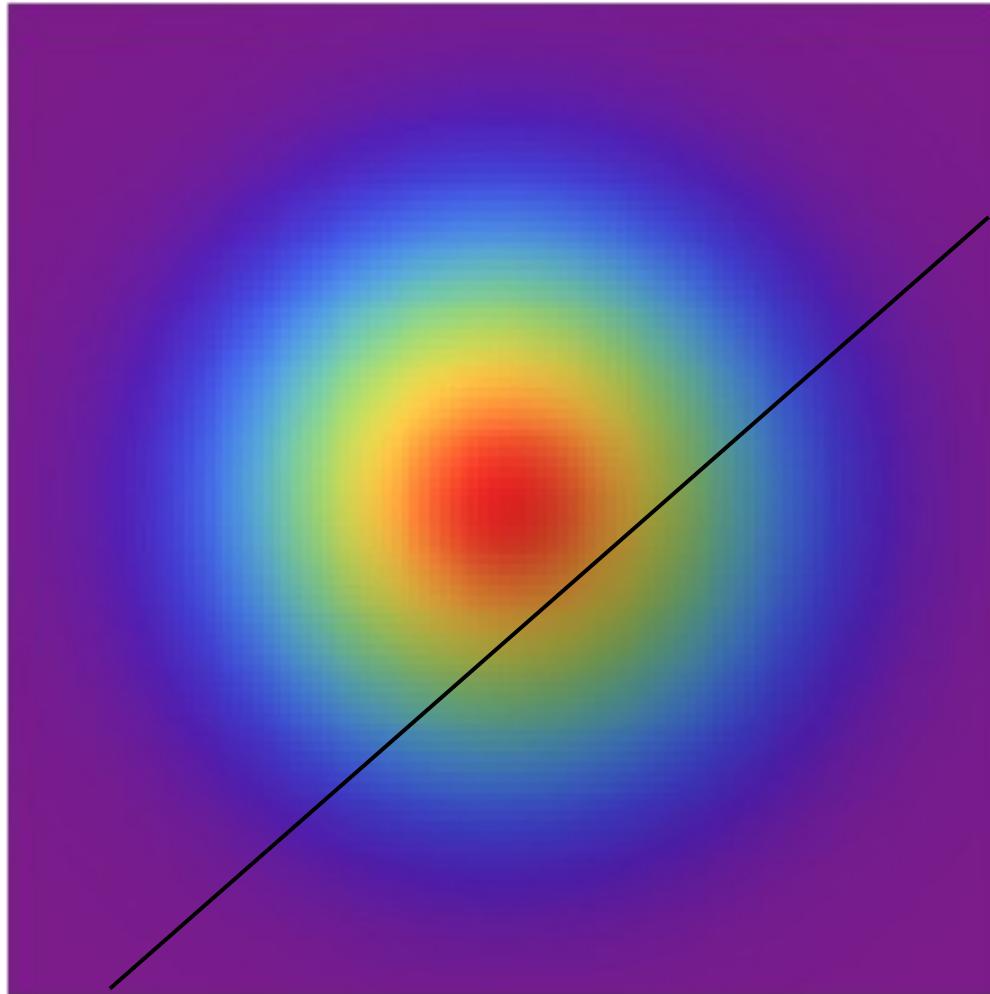
- **Conditional PDFs**

- If multivariate Gaussian X is partitioned into X_1 and X_2 ,
then the conditional PDF $P(X_1 | X_2=x_2)$ is also a multivariate Gaussian



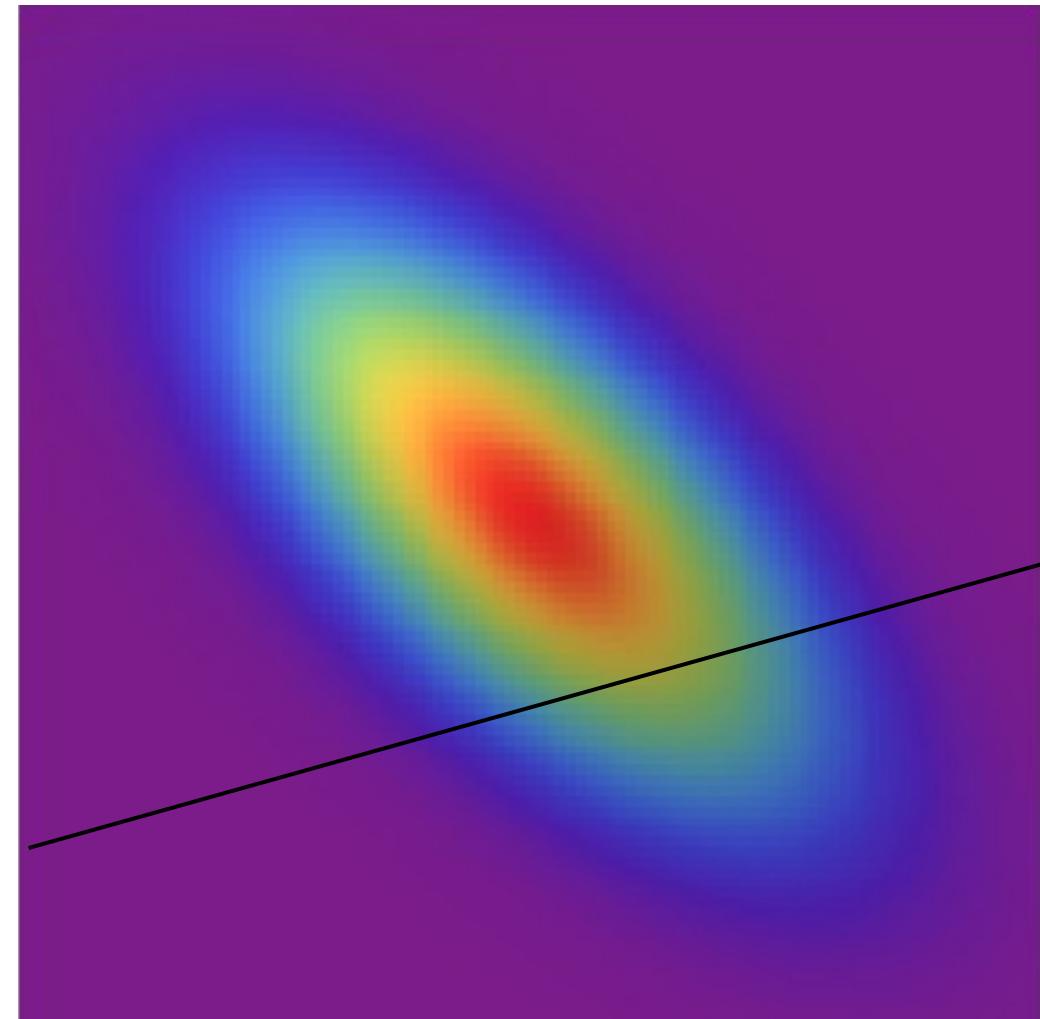
Multivariate Gaussian – Conditionals

- “Conditional” PDFs
 - What about this way of slicing ?
 - Yes, profile on the line has a Gaussian shape



Multivariate Gaussian – Conditionals

- “Conditional” PDFs
 - What about this way of slicing ?
 - Yes, profile on the line has a Gaussian shape



Multivariate Gaussian – ML Estimation

Multivariate Gaussian – ML Estimation

- Data: $\{y_1, \dots, y_N\}$
$$p(y) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$$
- Take log-likelihood function
- ML estimate (MLE) for mean vector (= sample mean)
 - Take derivative with respect to (w.r.t.) μ , and assign to zero. Solve.
 - Quadratic form $a^\top B a$
$$= \sum_i \sum_j a_i a_j B_{ij}$$
$$\frac{\partial}{\partial \mu} (\mu - x)^\top C^{-1}(\mu - x) = 2C^{-1}(\mu - x)$$
 - Partial derivative w.r.t. a_k
$$= \sum_j a_j B_{kj} + \sum_i a_i B_{ik}$$
$$= 2 \sum_j B_{kj} a_j \text{ (because } B \text{ is symmetric)}$$
$$= 2 \text{ (k-th row of } B \text{ * column-vector } a)$$
- Scalar function, say $f(a)$, of multiple scalar variables in column-vector ‘ a ’
 - Jacobian df/da will be a row vector of the same length as ‘ a ’
 - Change in function value (df) = derivative (df/da) * change in variable (da)
 - Can be reshaped/rearranged into a column vector of the same shape as ‘ a ’

Multivariate Gaussian – ML Estimation

- Data: $\{y_1, \dots, y_N\}$
$$p(y) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$$
- Take log-likelihood function
- MLE for covariance matrix (= sample covariance; uncorrected/biased)
 - Take derivative w.r.t. C , and assign to zero. Solve.
 - Need partial derivatives w.r.t. C_{ij}
 - Scalar function, say $f(C)$, of multiple scalar variables in C
 - Consider a (column)-vectorized form of C
 - Jacobian df/dC will be a row vector of the same length as (column)-vectorized C
 - Can be reshaped/rearranged into a matrix of the same shape as C

$$\frac{d}{dC} (x - \mu)^\top C^{-1} (x - \mu) = -C^{-\top} (x - \mu) (x - \mu)^\top C^{-\top}$$

$$\frac{d}{dC} \log(|C|) = \frac{1}{|C|} |C| C^{-\top} = C^{-\top}$$

Multivariate Gaussian – ML Estimation

- “Matrix Calculus”

- <http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html>
- https://en.wikipedia.org/wiki/Matrix_calculus
- <http://www.matrixcalculus.org/>

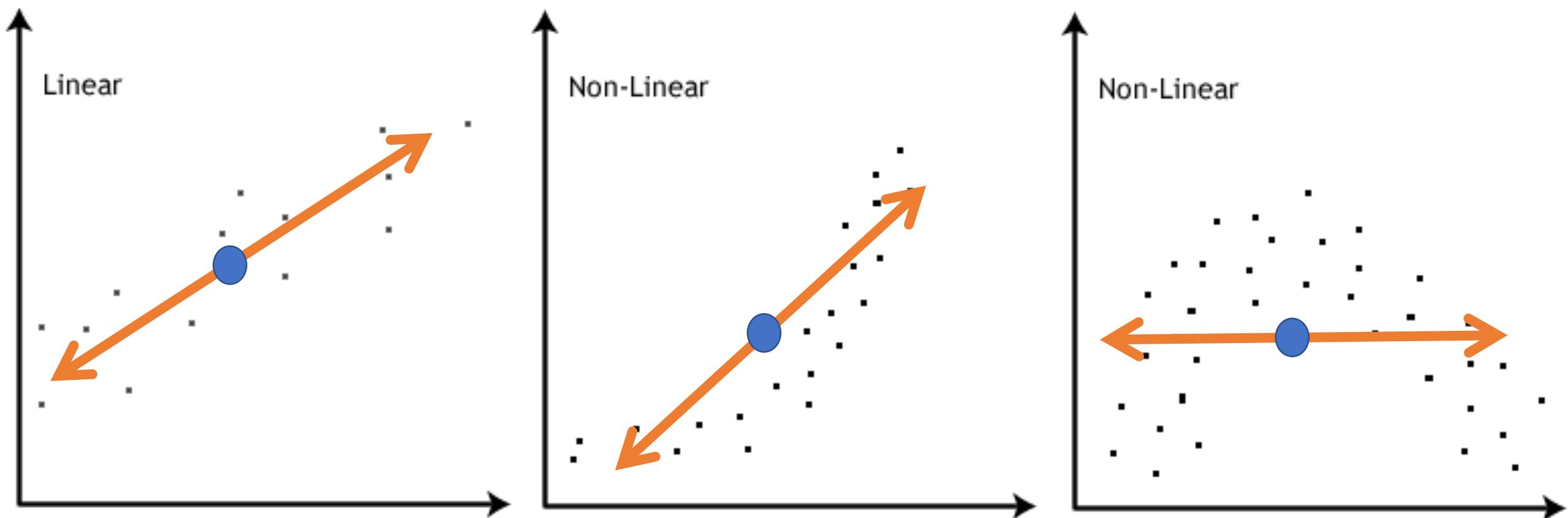
Principal Component Analysis (PCA)

Principal Component Analysis (PCA)

- Principal Component Analysis (PCA)
 - What is it about ?
 - What does it tell us about the distribution underlying the data ?
 - What does it tell us about the distribution underlying the data,
when the data is known to have a multivariate Gaussian distribution ?
 - Applications

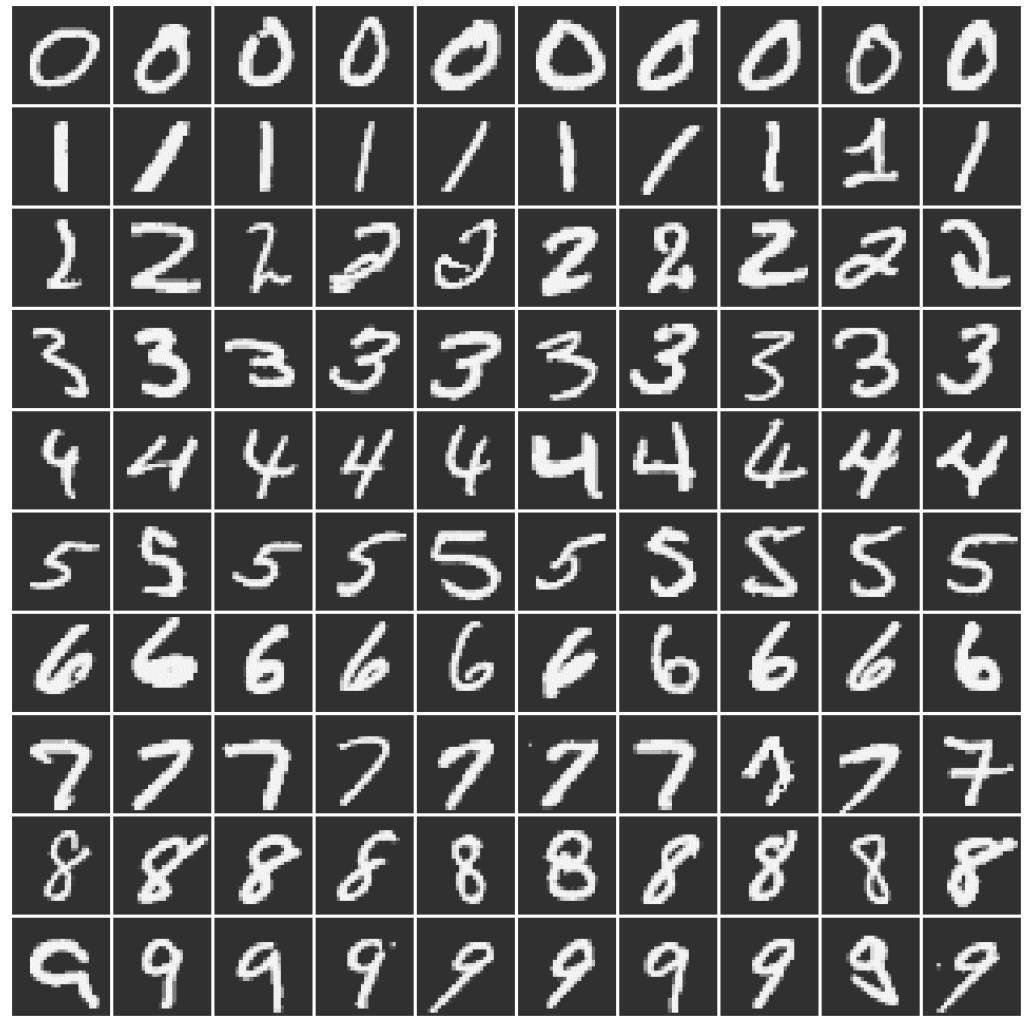
Principal Component Analysis (PCA)

- “Modes of variation”
 - Set of vectors (directions and magnitudes) that are used to depict the variation in a population or sample, around the mean



Principal Component Analysis (PCA)

- “Modes of variation”
 - Set of vectors (directions and magnitudes) that are used to depict the variation in a population or sample, around the mean
 - Can we do it for the distribution of images of each digit ?



Principal Component Analysis (PCA)

- Directions of maximal variance
 - Consider a general multivariate random variable X with PDF $P(X)$
 - **We aren't assuming it to be a Gaussian yet**

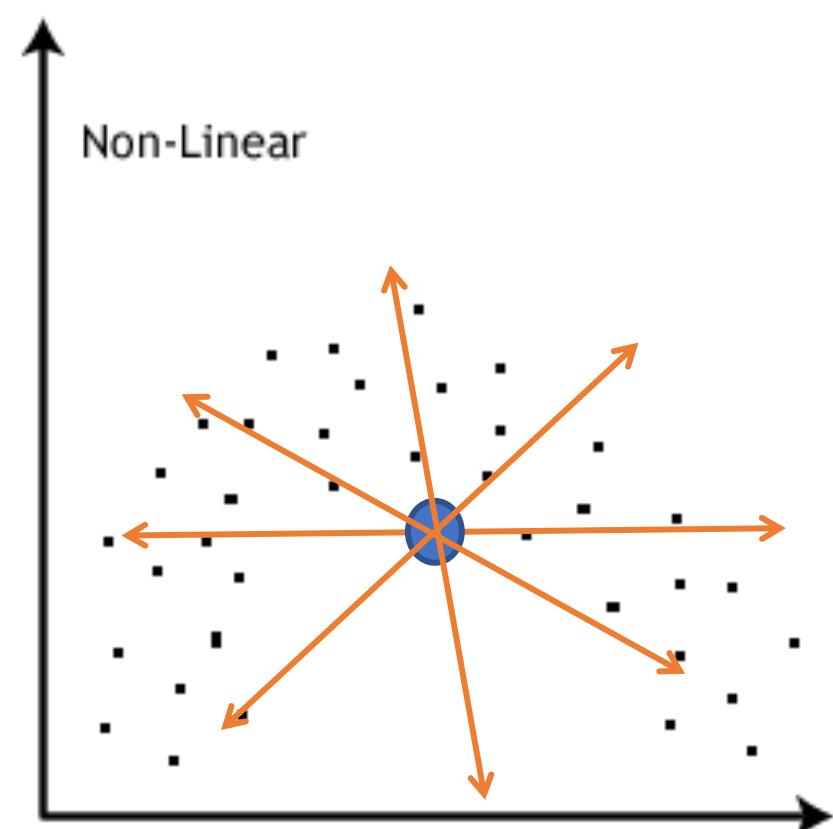
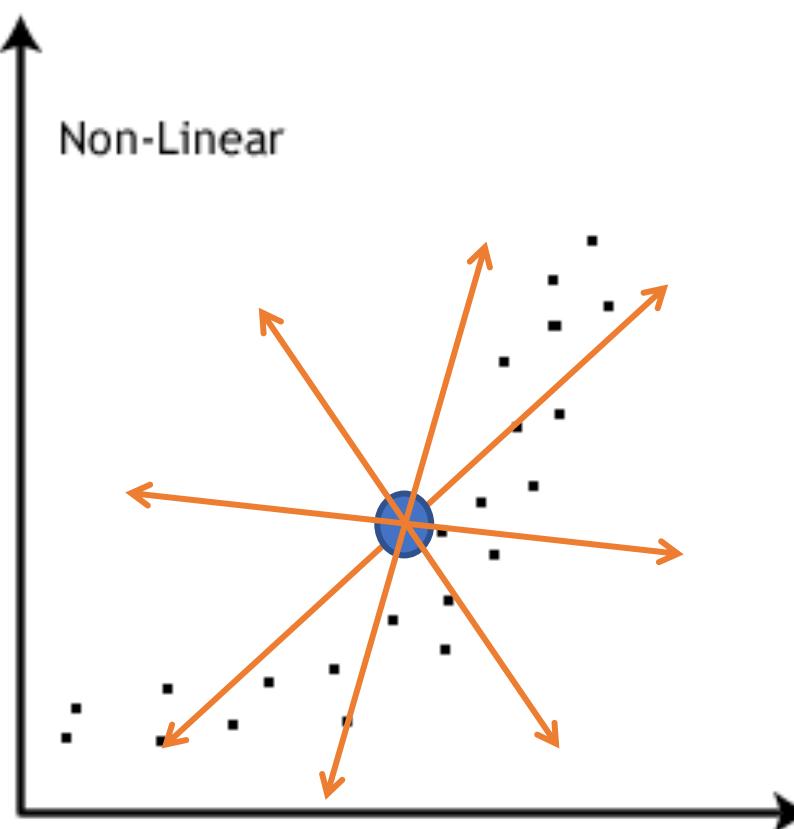
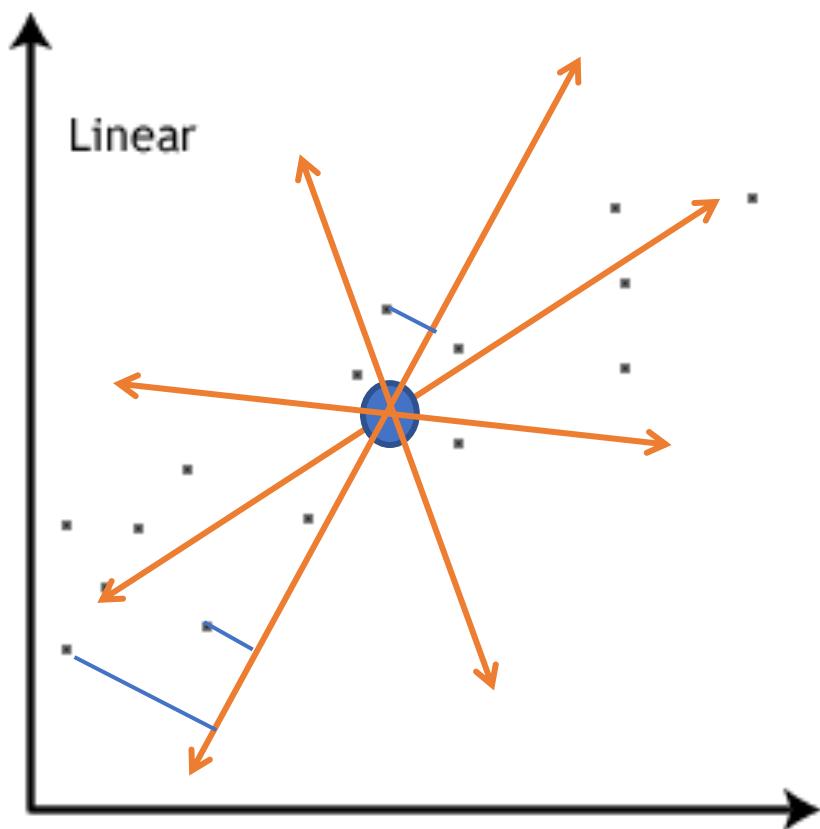
Consider observed multivariate data $\{x_i \in \mathbb{R}^D\}_{i=1}^N$
drawn from some PDF $P(X)$ with some mean μ and some covariance matrix C

As $N \rightarrow \infty$ for the data, sample mean $\rightarrow \mu$, and sample covariance $\rightarrow C$

Principal Component Analysis (PCA)

- Directions of maximal variance

Find the “direction” v (i.e., $\| v \|_2 = 1$) such that the data projected on the 1D space indicated by (i) mean μ and (ii) direction v that passes through μ has the maximal variance



Principal Component Analysis (PCA)

- Directions of maximal variance

For finite N , we perform the analysis in a shifted coordinate frame where the sample mean $\sum_{i=1}^N x_i/N = 0$

Assume the sample mean for $\{x_i \in \mathbb{R}^D\}_{i=1}^N$ is at the origin

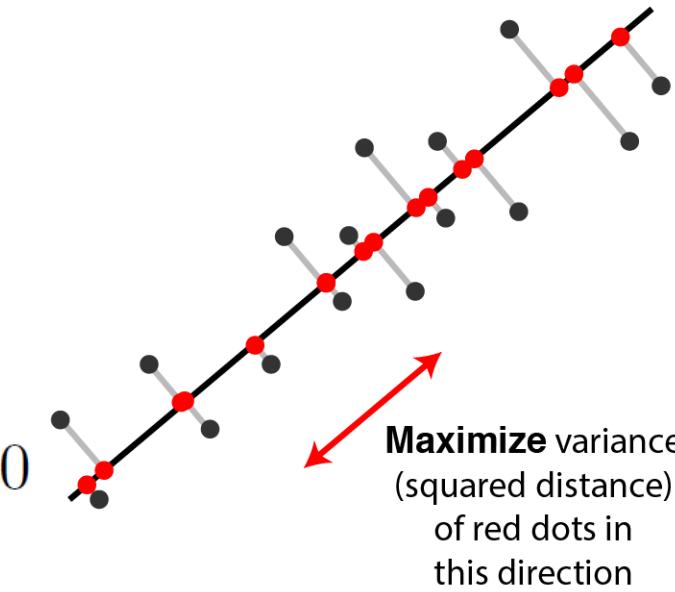
Projected data are $\langle x_i, v \rangle v$

Sample mean of projected data, in the 1D space, is $\sum_i \langle x_i, v \rangle v = 0$

Distance of projected data from the sample mean (i.e., origin) is

$$\|\langle x_i, v \rangle v\|_2 = |\langle x_i, v \rangle|$$

Sample variance of projected data, in the 1D space, is $\sum_i \langle x_i, v \rangle^2 / N$

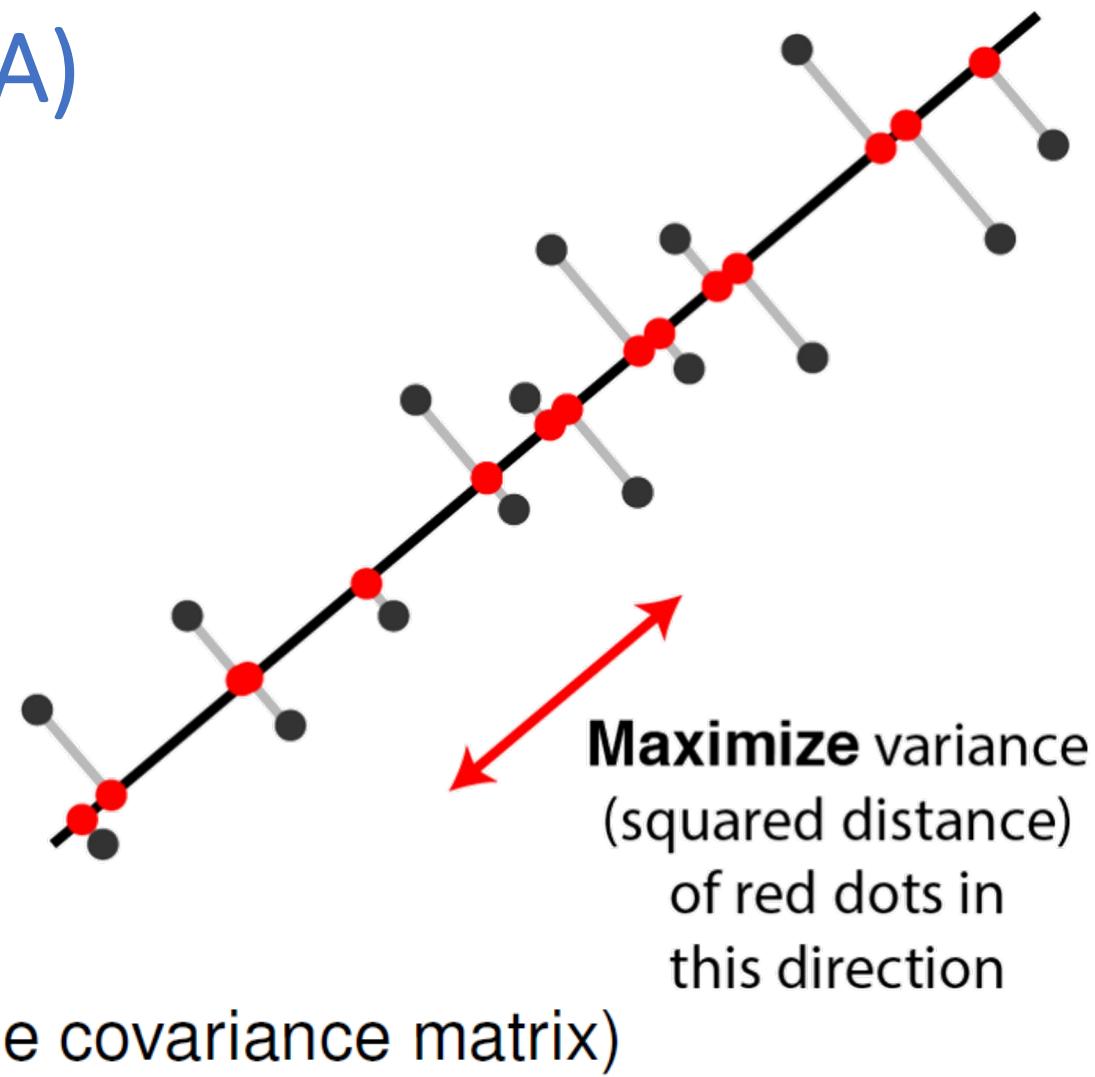


Principal Component Analysis (PCA)

- Directions of maximal variance

Optimal direction

$$\begin{aligned} &= \arg \max_{v: \|v\|_2=1} \sum_i \langle x_i, v \rangle^2 / N \\ &= \arg \max_{v: \|v\|_2=1} \sum_i (x_i^\top v)^2 / N \\ &= \arg \max_{v: \|v\|_2=1} \sum_i (x_i^\top v)^\top (x_i^\top v) / N \\ &= \arg \max_{v: \|v\|_2=1} \sum_i v^\top x_i x_i^\top v / N \\ &= \arg \max_{v: \|v\|_2=1} v^\top (\sum_i x_i x_i^\top / N) v \\ &= \arg \max_{v: \|v\|_2=1} v^\top C v \text{ (where } C \text{ is the sample covariance matrix)} \end{aligned}$$

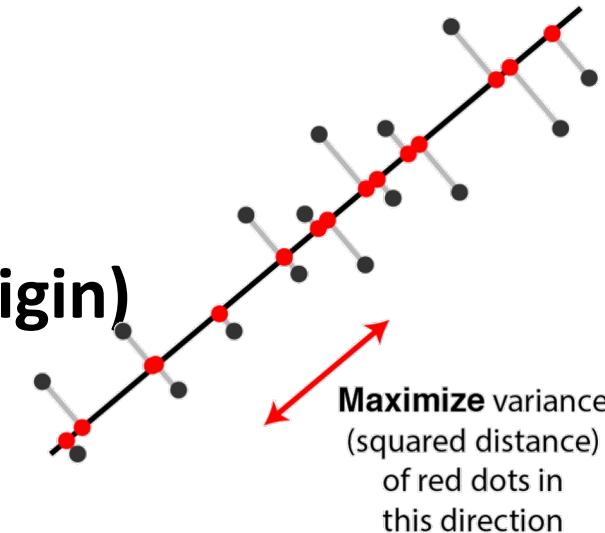


Principal Component Analysis (PCA)

- Directions of maximal variance

- When covariance matrix C is diagonal (sample mean at origin)**

- Let d -th element on diagonal of C be C_{dd}
- Let d -th element in vector ' v ' be v^d



Without loss of generality, let the diagonal elements be sorted in descending order, i.e., $C_{11} \geq C_{22} \geq \dots$

The optimal direction is $\arg \max_{v: \|v\|_2=1} \sum_d C_{dd}(v^d)^2$

The unit-norm constraint on v is: $\sum_d (v^d)^2 = 1$

- Objective function is a convex combination of $\{C_{dd}\}$ with weights $\{(v^d)^2\}$

The “objective function” $\sum_d C_{dd}(v^d)^2$ is maximized when $v^1 = 1$, and $v^d = 0$ for $d = 2, \dots, D$

The maximal variance then is C_{11} that is the eigenvalue corresponding to the principal eigenvector $[1, 0, \dots, 0]^\top$

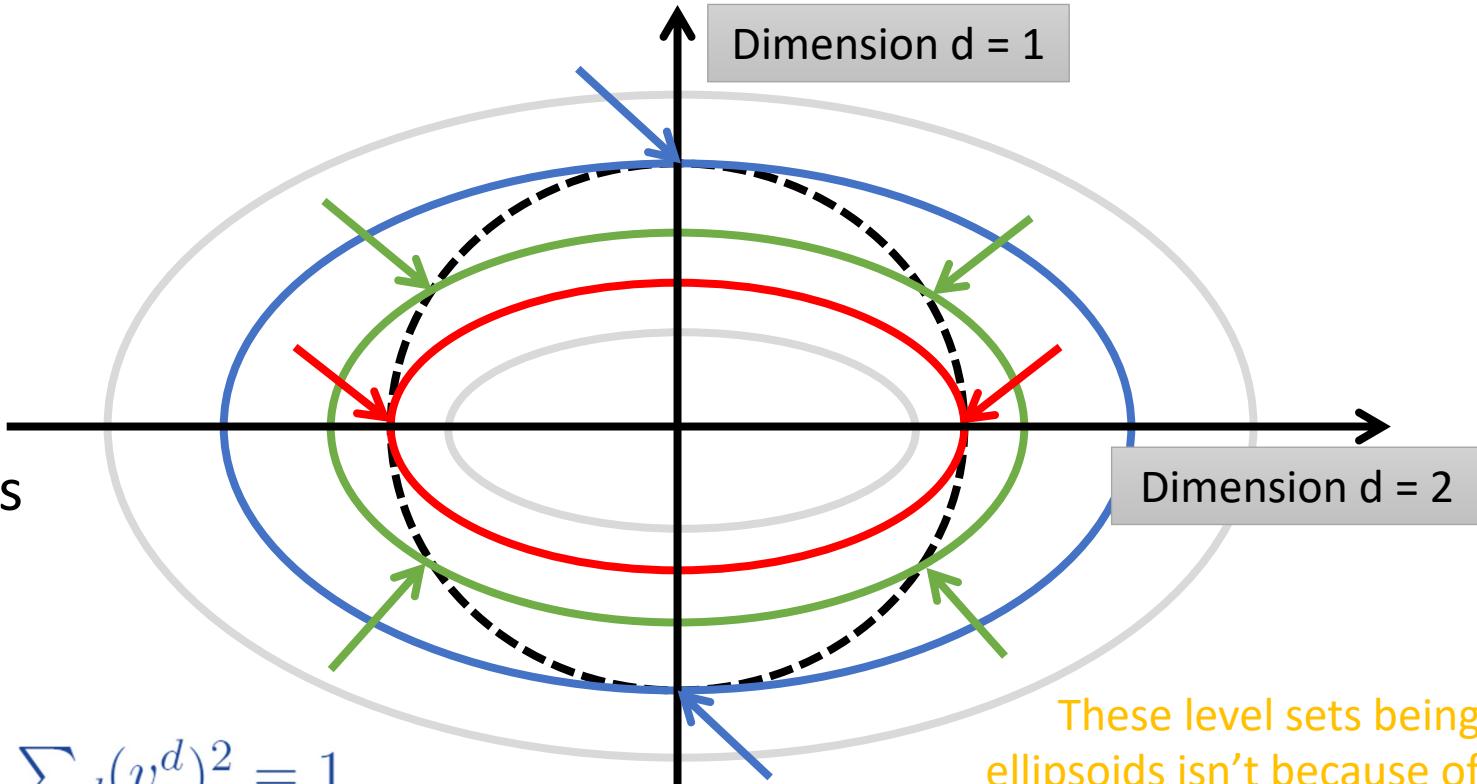
Principal Component Analysis (PCA)

- Directions of maximal variance

- When covariance matrix C is diagonal (and sample mean at origin):

- Minor axis corresponds to dimension d with smallest $1/C_{dd}$, i.e., largest C_{dd}

- The point on hypersphere that maximizes objective function lies at the end of the minor axis of one of the hyper-ellipsoids



“Constraint set” is the hyper-sphere $\sum_d (v^d)^2 = 1$

Level sets of objective function $\sum_d C_{dd} (v^d)^2$ are hyper-ellipsoids with axes lengths proportional to $1/\sqrt{C_{dd}}$

Principal Component Analysis (PCA)

- Directions of maximal variance

- When covariance matrix C is diagonal (and sample mean at origin):

Now find the 2nd direction u that is:

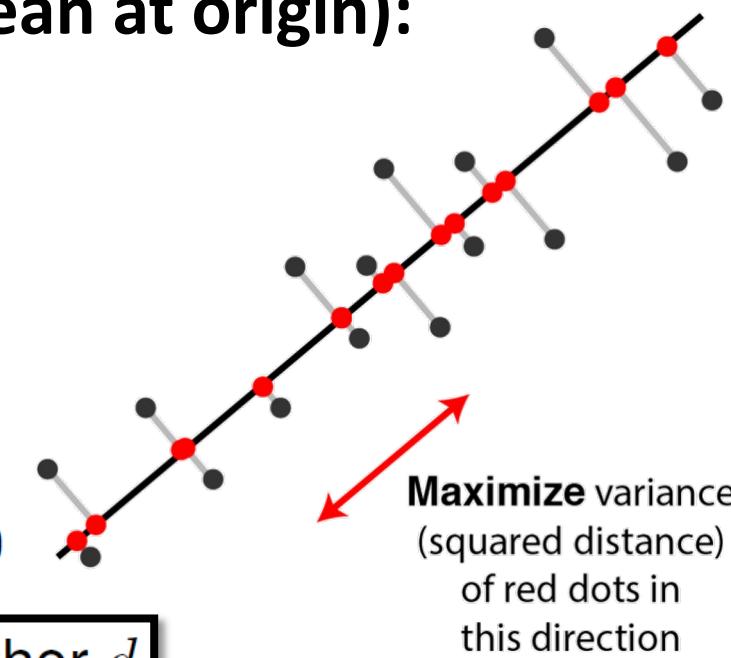
- (i) orthogonal to v and
- (ii) maximizes the variance of the data projected onto it

Optimal direction

$$= \arg \max_{u: \|u\|_2=1, u \perp v} \sum_i \langle x_i, u \rangle^2$$

$$= \arg \max_{u: \|u\|_2=1, u \perp v} \sum_d C_{dd} (u^d)^2, \text{ where we know } C_{dd} \geq 0$$

This is maximized when $u^d = 1$ for $d = 2$, and $u^d = 0$ for all other d



- Second mode of variation is second cardinal axis (another eigenvector). Variance along that mode = second-largest eigenvalue = C_{22}
- Similar arguments hold for 3rd, 4th, ... directions
- Thus, for any $P(X)$ with a **diagonal** covariance matrix C , modes of variation are **cardinal directions** that **maximize variance** of projected data

Principal Component Analysis (PCA)

- Directions of maximal variance
 - For a general SPD covariance matrix C (and sample mean at origin):

Then, $C = Q\Lambda Q^\top$, where the diagonal of Λ has sorted values (high to low)

Then, $\max_v v^\top Cv = \max_v v^\top Q\Lambda Q^\top v = \max_{u:=Q^\top v} u^\top \Lambda u$

Thus, the principal mode of variation is given by $u = [1, 0, \dots, 0]^\top$ (in the Q basis)

that is equivalent to $v = Qu = \text{first column of } Q$ (in the original basis)

The maximal variance then is Λ_{11} that is the eigenvalue corresponding to the principal eigenvector

Similarly, the remaining modes of variation will be the other columns of Q

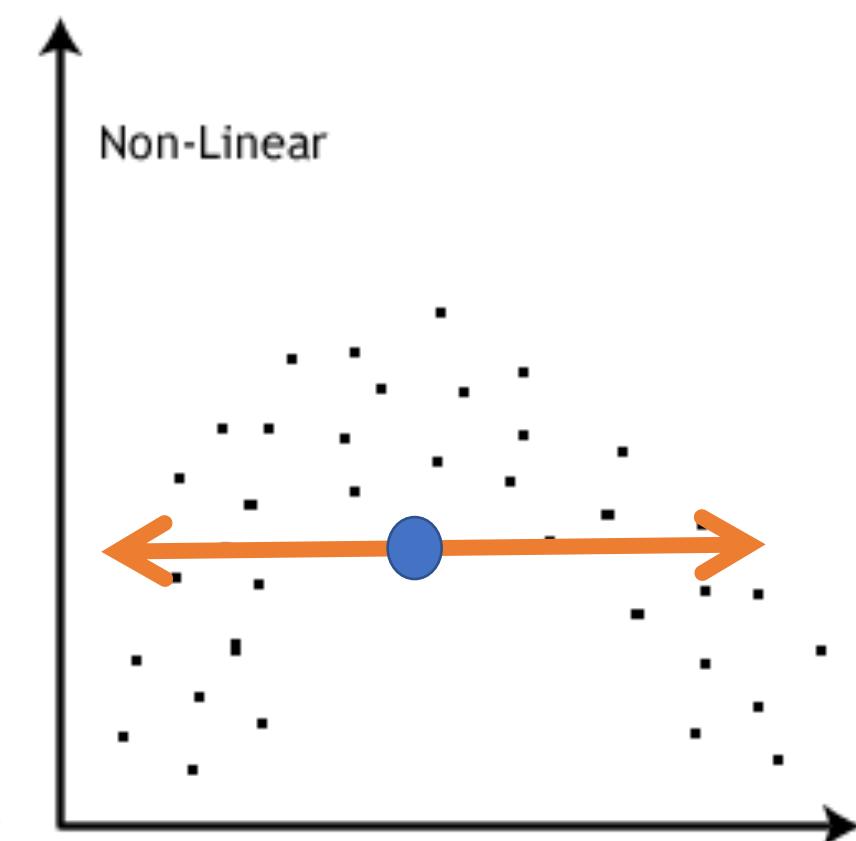
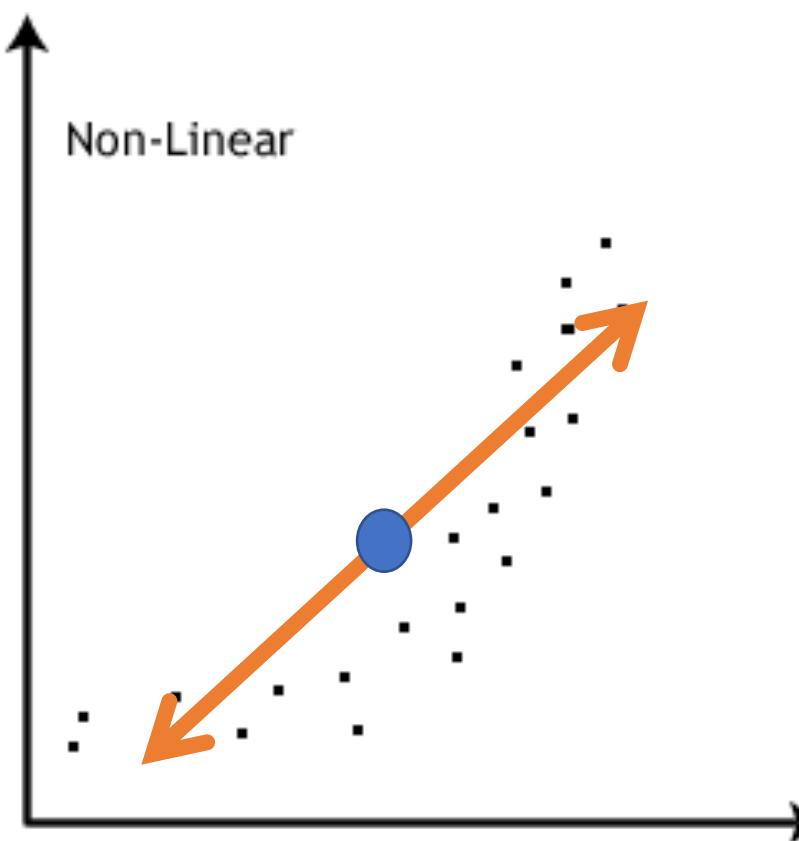
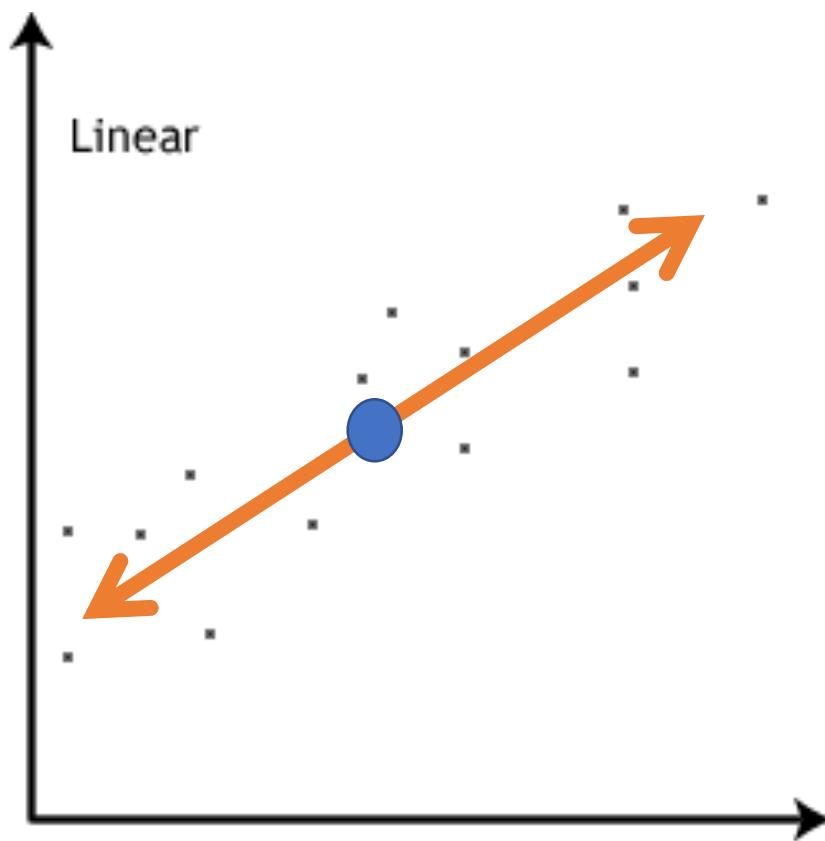
Principal Component Analysis (PCA)

- Directions of maximal variance

- Data



- What do the eigenvectors look like ? Compute them and see for yourself.



Principal Component Analysis (PCA)

- Example 1: `rng(0); N = 1e5; data = randn (2,N)`

```
>> rng(0); N=1e5; data=(randn(2,N)); C=cov(data'), [V,D]=eig(C)
```

C =

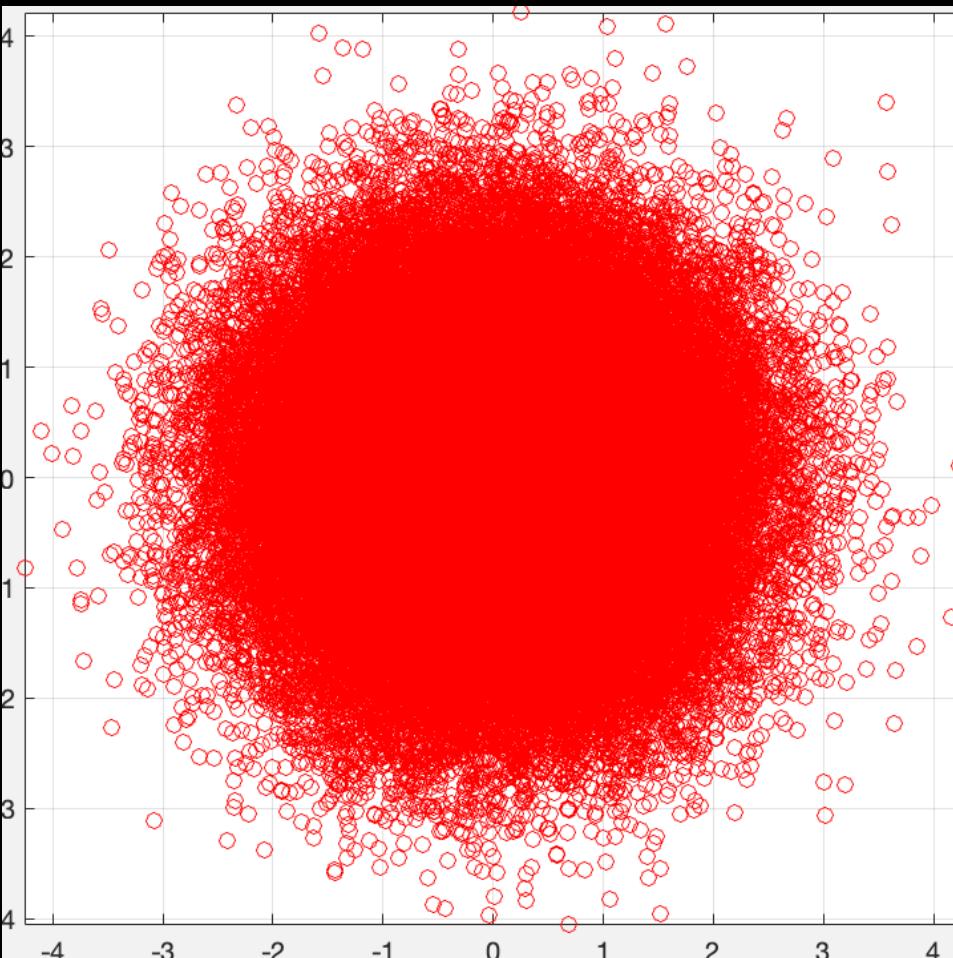
```
1.0007 -0.0023  
-0.0023 0.9966
```

V =

```
-0.4112 -0.9115  
-0.9115 0.4112
```

D =

```
0.9956 0  
0 1.0018
```



Principal Component Analysis (PCA)

- Example 2: `rng(0); N = 1e5; data = rand (2,N)`

```
>> rng(0); N=1e5; data=rand(2,N)); C=cov(data'), [V,D]=eig(C)
```

C =

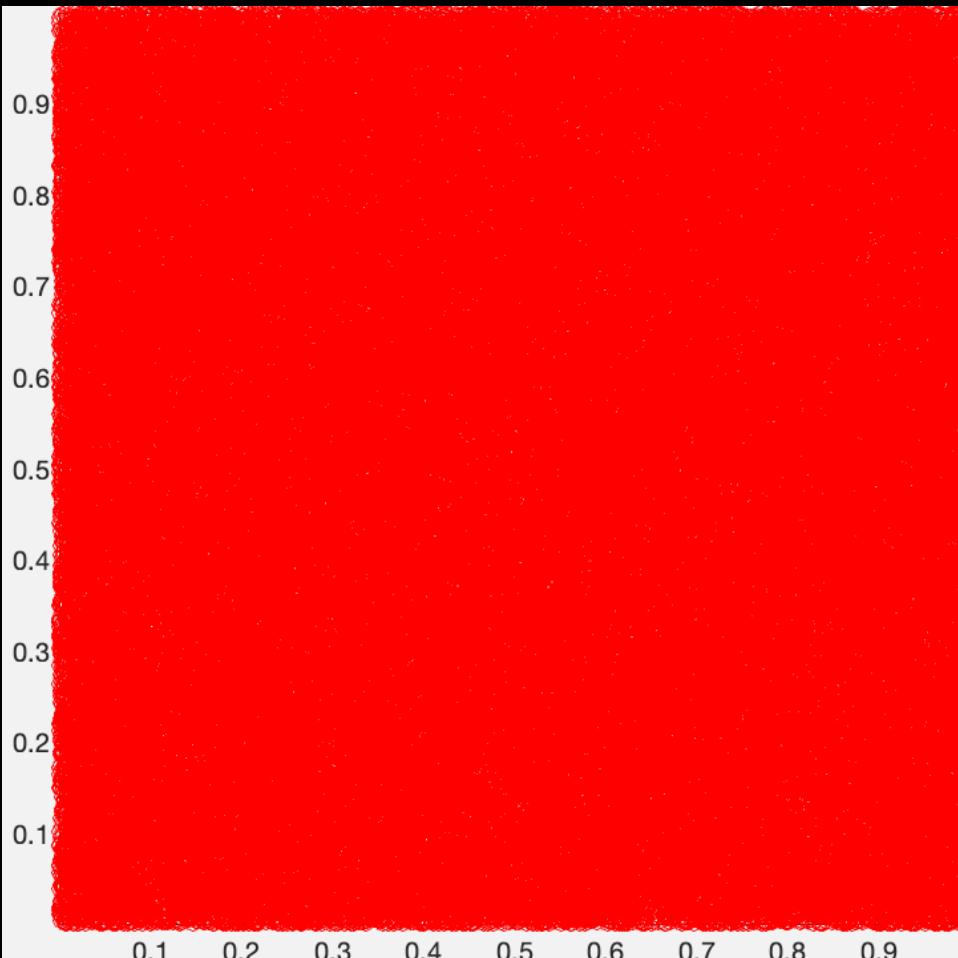
```
0.0835 -0.0001  
-0.0001 0.0831
```

V =

```
-0.1710 -0.9853  
-0.9853 0.1710
```

D =

```
0.0831 0  
0 0.0835
```



Principal Component Analysis (PCA)

- Example 3: `rng(0); N = 1e5; data = [[1 -1];[1 1]]/sqrt(2) * (rand (2,N))`

```
>> rng(0); N=1e5; data=[[1 -1];[1 1]]/sqrt(2)*(rand(2,N)); C=cov(data'), [V,D]=eig(C)
```

C =

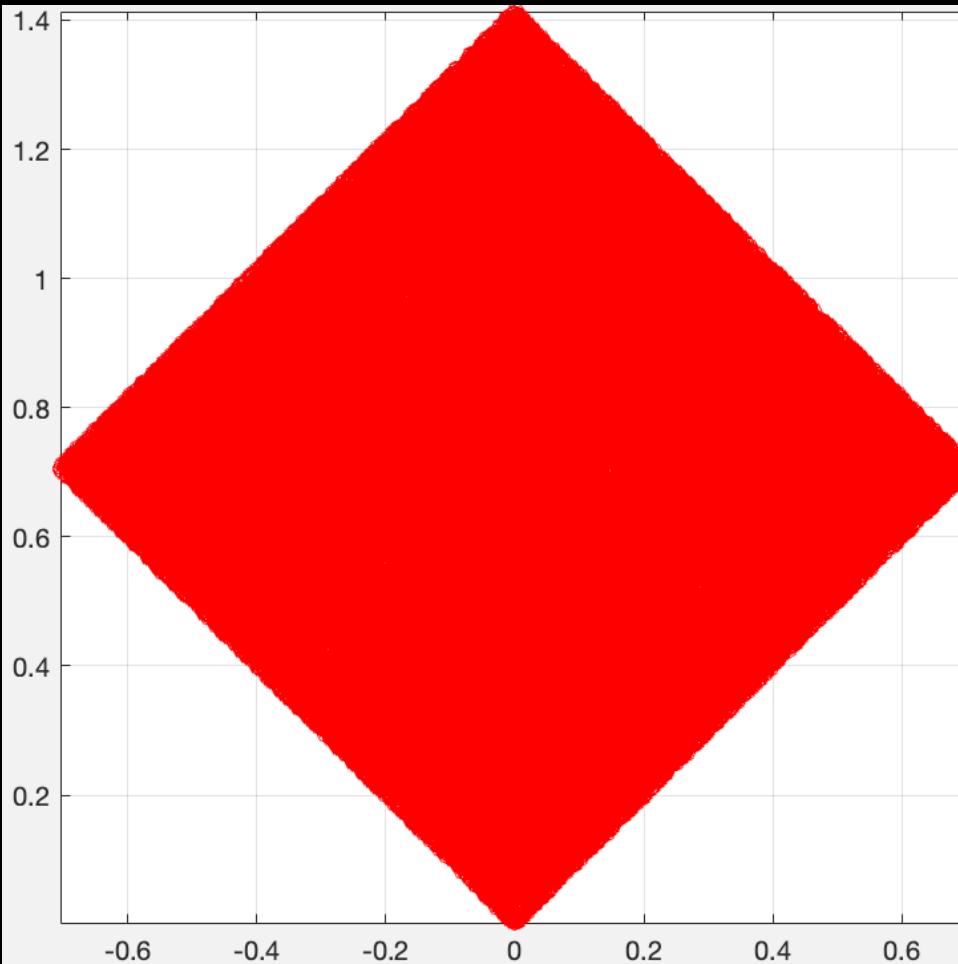
```
0.0834    0.0002  
0.0002    0.0832
```

V =

```
0.5758   -0.8176  
-0.8176   -0.5758
```

D =

```
0.0831      0  
0           0.0835
```



Principal Component Analysis (PCA)

- Example 4: `rng(0); N = 1e5; data = exprnd (1,[2,N])`

```
>> rng(0); N=1e5; data=(exprnd(1, [2,N])); C=cov(data'), [V,D]=eig(C)
```

C =

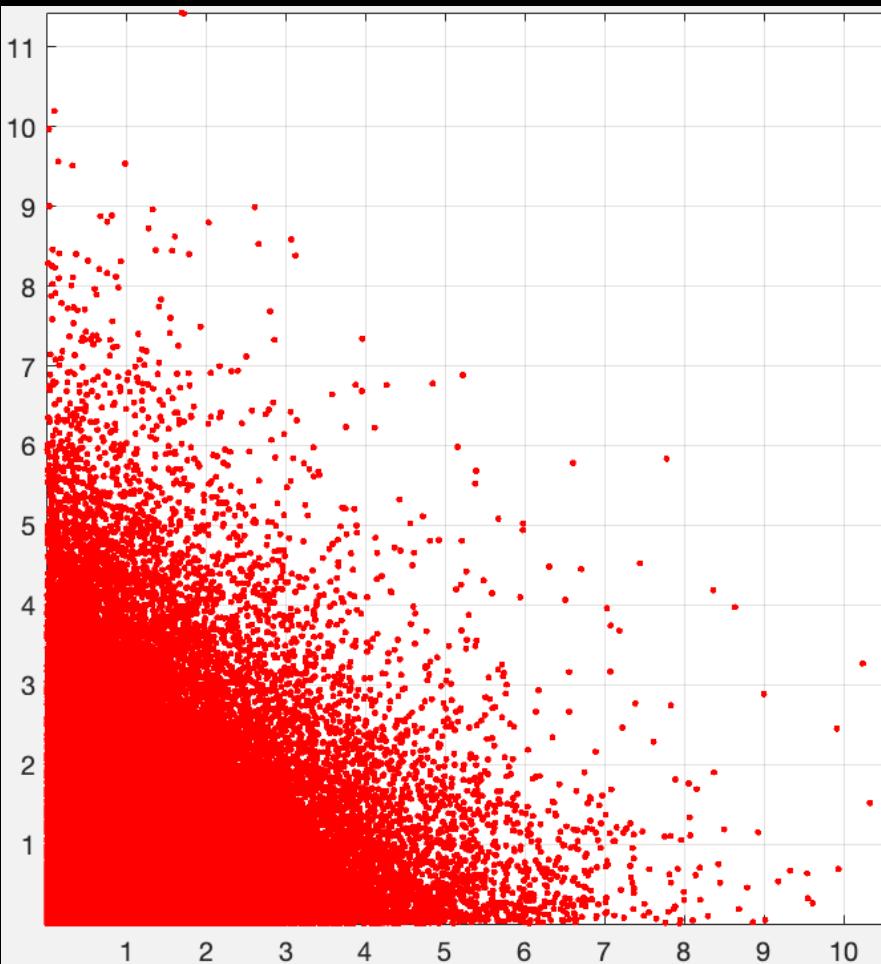
```
1.0030    0.0001  
0.0001    0.9991
```

V =

```
0.0147   -0.9999  
-0.9999  -0.0147
```

D =

```
0.9991      0  
0          1.0030
```



Principal Component Analysis (PCA)

- Example 5: $\text{rng}(0); N=1\text{e}5; \text{data} = [[1 -1]; [1 1]]/\sqrt{2} * \text{exprnd}(1, [2, N])$

```
>> rng(0); N=1e5; data= [[1 -1]; [1 1]]/sqrt(2) * (exprnd(1, [2,N])); C=cov(data'), [V,D]=eig(C)
close
```

C =

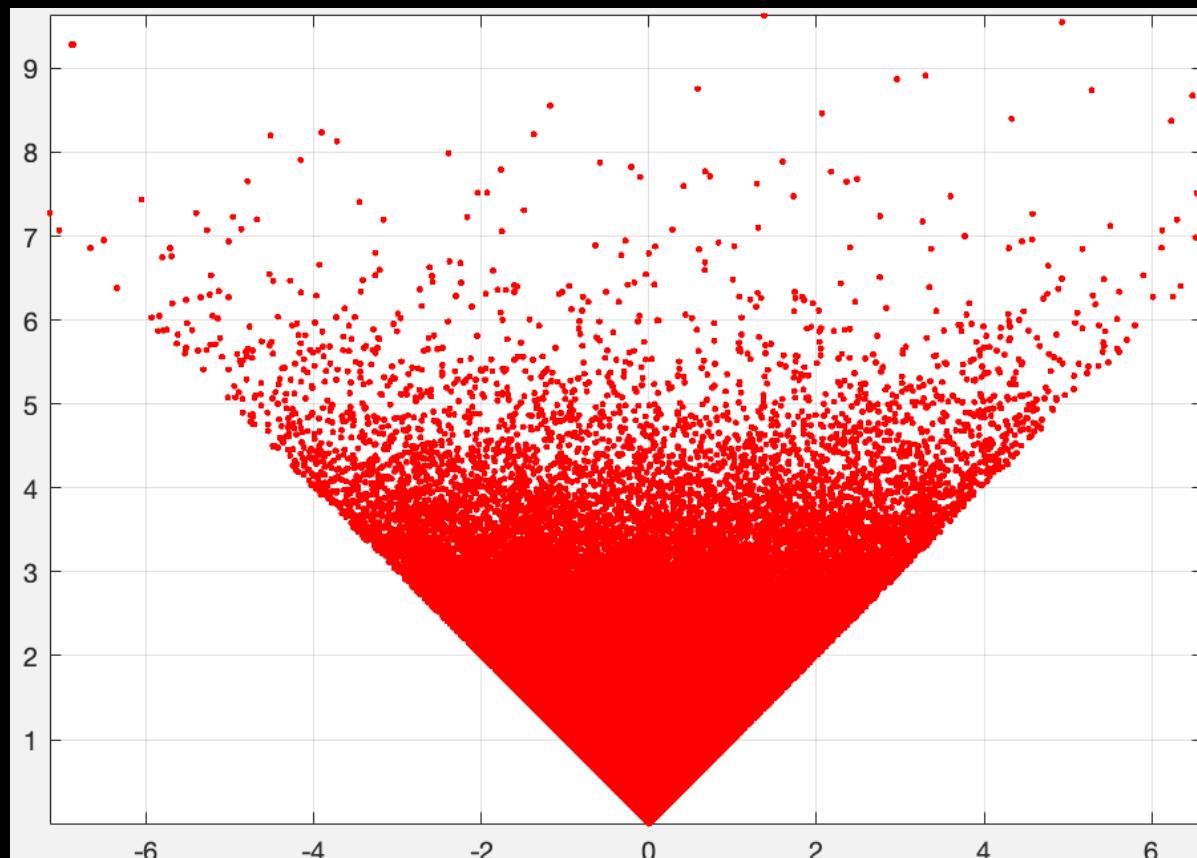
$$\begin{bmatrix} 1.0010 & 0.0020 \\ 0.0020 & 1.0011 \end{bmatrix}$$

V =

$$\begin{bmatrix} -0.7174 & 0.6966 \\ 0.6966 & 0.7174 \end{bmatrix}$$

D =

$$\begin{bmatrix} 0.9991 & 0 \\ 0 & 1.0030 \end{bmatrix}$$



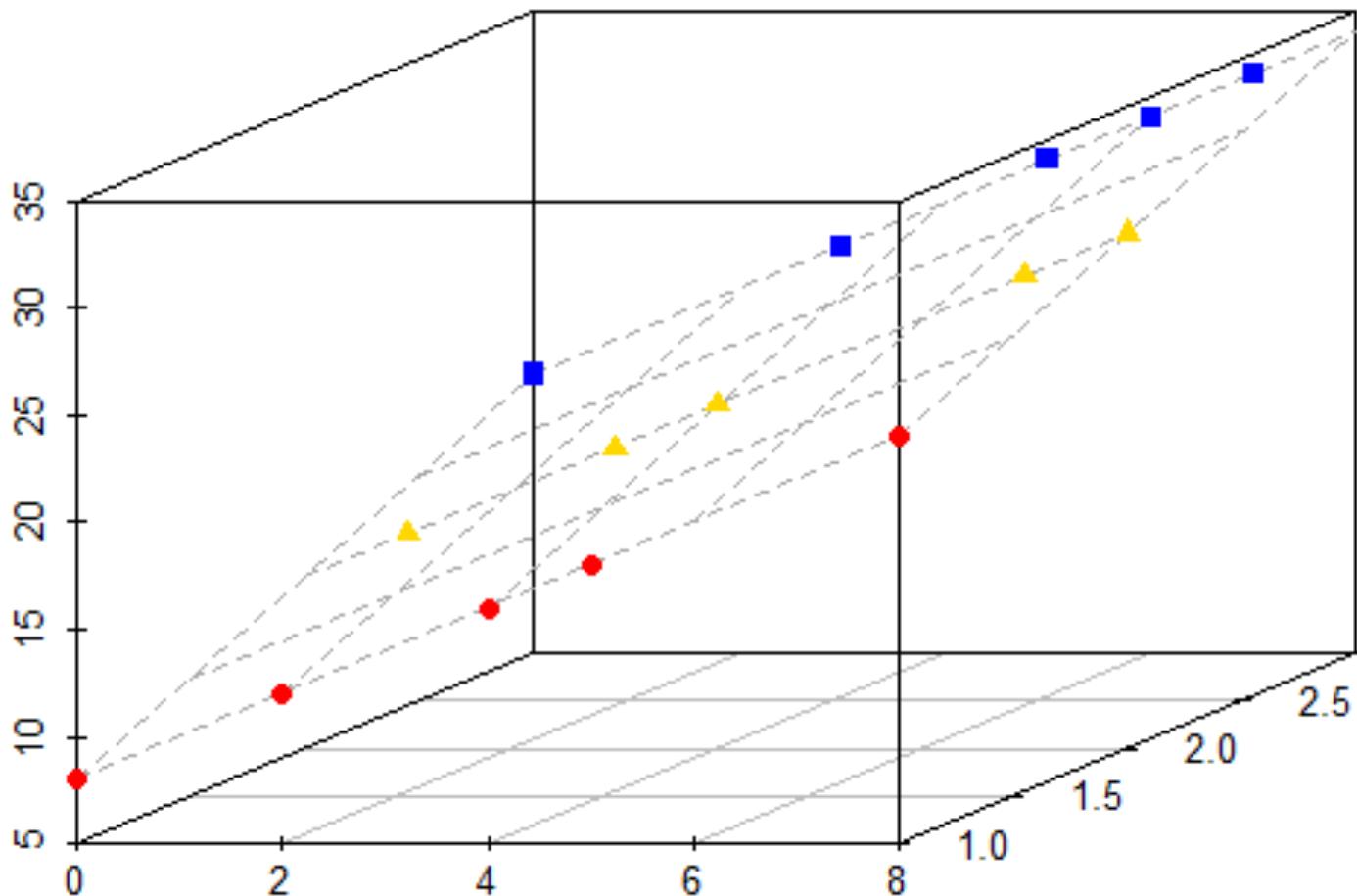
Principal Component Analysis (PCA)

- What happens to covariance matrix when we rotate the data ?
- In general
 - Let $C_X := E[(X - \mu_X)(X - \mu_X)^T] = E[XX^T] - \mu_X\mu_X^T$
 - Let $Y := RX$ (for any invertible R)
 - Then, $\mu_Y := E[Y] = E[RX] = RE[X] = R\mu_X$
 - Then, $C_Y := E[(Y - \mu_Y)(Y - \mu_Y)^T]$
 - $= E[YY^T] - \mu_Y\mu_Y^T$
 - $= E[RXX^TR^T] - R\mu_X\mu_X^TR^T$
 - $= R(E[XX^T] - \mu_X\mu_X^T)R^T$
 - $= RC_XR^T$
- If $C_X = aI$ and R is orthogonal, then $C_Y = RaIR^T = aI = C_X$
 - Thus, eigenvalues don't change, and eigenvectors can be any orthogonal basis
 - Thus, directions of maximal variance aren't unique

Principal Component Analysis (PCA)

- **Spaces of maximal variance**

- What if we want to find multi-D lower-dimensional spaces that maximize “*total dispersion/variance*” ?
 - Total dispersion/variance is empirical average of squared distance from mean



Principal Component Analysis (PCA)

- **Spaces of maximal variance**

- What if we want to find multi-D lower-dimensional spaces that maximize “*total dispersion/variance*”?

- Total dispersion/variance is empirical average of squared distance from mean

- **When covariance matrix C is diagonal (and sample mean is at origin):**

Let a 2D space be defined by (i) mean μ and (ii) orthogonal directions v_1 and v_2 through μ

Then the goal is to find $\arg \max_{v_1, v_2} v_1^\top C v_1 + v_2^\top C v_2$

under the constraints $\|v_1\| = 1$, $\|v_2\| = 1$, and $\langle v_1, v_2 \rangle = 0$

The objective function is

$$\begin{aligned} & v_1^\top C v_1 + v_2^\top C v_2 \\ &= \sum_{d=1}^D C_{dd}(v_1^d)^2 + \sum_{d=1}^D C_{dd}(v_2^d)^2 \\ &= \sum_{d=1}^D C_{dd}[(v_1^d)^2 + (v_2^d)^2] \end{aligned}$$

Optimal direction

$$\begin{aligned} &= \arg \max_{v: \|v\|_2=1} \sum_i \langle x_i, v \rangle^2 / N \\ &= \arg \max_{v: \|v\|_2=1} \sum_i (x_i^\top v)^2 / N \\ &= \arg \max_{v: \|v\|_2=1} \sum_i (x_i^\top v)^\top (x_i^\top v) / N \\ &= \arg \max_{v: \|v\|_2=1} \sum_i v^\top x_i x_i^\top v / N \\ &= \arg \max_{v: \|v\|_2=1} v^\top (\sum_i x_i x_i^\top / N) v \\ &= \arg \max_{v: \|v\|_2=1} v^\top C v \text{ (where } C \text{ is the)} \end{aligned}$$

Principal Component Analysis (PCA)

- **Spaces of maximal variance**

- What if we want to find multi-D lower-dimensional spaces that maximize “*total dispersion/variance*” ?
- **When covariance matrix C is diagonal (and sample mean at origin):**

Now, consider v_1 and v_2 as the first two columns vectors of a basis (of D vectors) for the complete D -dimensional Euclidean space

Then, for any dimension d , we have $(v_1^d)^2 + (v_2^d)^2 + \dots + (v_D^d)^2 = 1$ (each row in the basis has unit norm)

Thus, for any d , we have the constraint $(v_1^d)^2 + (v_2^d)^2 \leq 1$

Another constraint is: $\sum_{d=1}^D [(v_1^d)^2 + (v_2^d)^2] = 1 + 1 = 2$ (each column, i.e., v_1 and v_2 , has unit norm)

Principal Component Analysis (PCA)

- **Spaces of maximal variance**

- What if we want to find multi-D lower-dimensional spaces that maximize “*total dispersion/variance*” ?
- **When covariance matrix C is diagonal (and sample mean at origin):**

Let us define a vector a such that its d -th component $a^d := (v_1^d)^2 + (v_2^d)^2$

Then, the new objective function is $\sum_{d=1}^D C_{dd}a^d$

and the new constraints are: (i) $\forall d, 0 \leq a^d \leq 1$ and (ii) $\sum_d a^d = 2$ (two)

- Now this problem is similar to what we had solved before
- Like before, we have a “convex” combination of $\{C_{dd}\}$ with weights $\{0 \leq a^d \leq 1\}$
- So, increase a^1 to its limit (i.e., 1) and then increase a^2 to its limit (i.e., 1)

The solution for this problem is $a^1 = a^2 = 1$ and all other $a^d = 0$ (because $C_{11} \geq C_{22} \geq \dots$)

Principal Component Analysis (PCA)

- **Spaces of maximal variance**

- What if we want to find multi-D lower-dimensional spaces that maximize “*total dispersion/variance*” ?
- **When covariance matrix C is diagonal (and sample mean at origin):**

The solution for this problem is $a^1 = a^2 = 1$ and all other $a^d = 0$ (because $C_{11} \geq C_{22} \geq \dots$)

That implies:

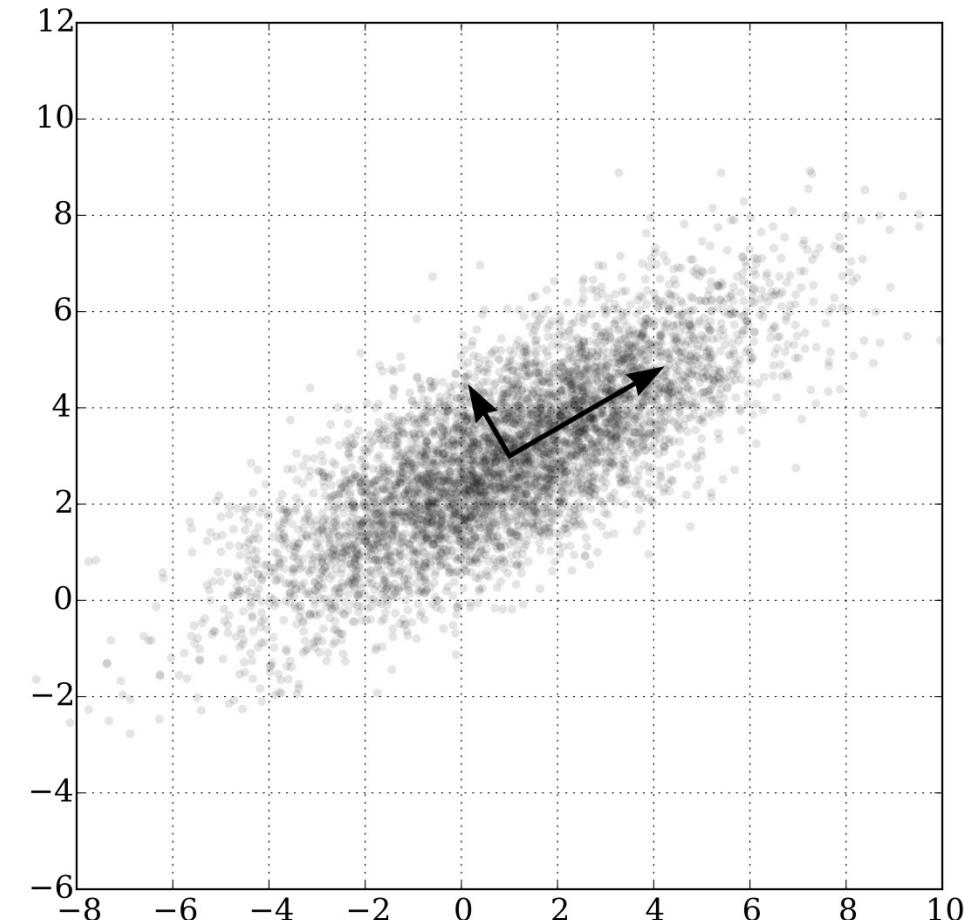
- (i) $v_1^d = v_2^d = 0$ for all $d > 2$; so that optimal 2D space is the one spanned by the first two cardinal axes, and
- (ii) to maximize $\sum_{d=1}^D C_{dd}[(v_1^d)^2 + (v_2^d)^2]$ (where $C_{11} \geq C_{22}$), we choose $v_1^1 = 1, v_2^1 = 0$ and $v_1^2 = 0, v_2^2 = 1$

Thus, total variance in the (optimal) 2D space is $\sum_{d=1}^D C_{dd}[(v_1^d)^2 + (v_2^d)^2] = C_{11} + C_{22}$ that is the sum of the top two eigenvalues

- Any orthogonal basis spanning the space spanned by first 2 cardinal axes will be a solution
- Similar arguments will hold for **lower-dimensional spaces of dimensions 3, 4, ..., D-1**
- Similar arguments will also hold for a **general SPD covariance matrix C**

Principal Component Analysis (PCA)

- PCA applied to data from a multivariate Gaussian distribution
 - Consider X is multivariate Gaussian
 - If $X := AW + b$, then:
 - Principal modes of variation are **directions** given by **eigenvectors** of covariance matrix $C := AA^T$
 - Principal modes of variation are along axes of hyper-ellipsoids that are level sets of $P(X)$
 - **Variances** along principal modes of variation are the **eigenvalues** of C
 - If $X := RSW + b$, then:
 - Principal modes of variation are **column vectors** of **orthogonal** matrix R i.e., **eigenvectors** of $C = RS^2R^T$
 - Variances along principal modes of variation are the **eigenvalues** of C , i.e., **diagonal** elements in S^2

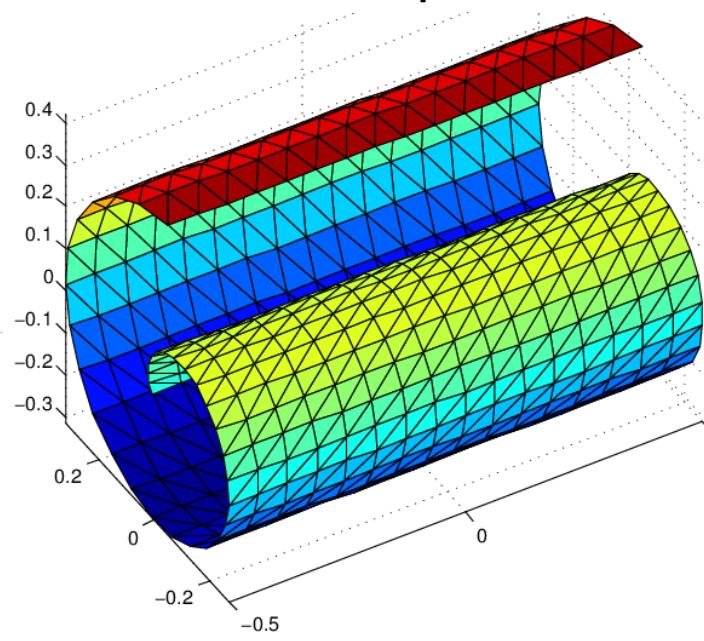
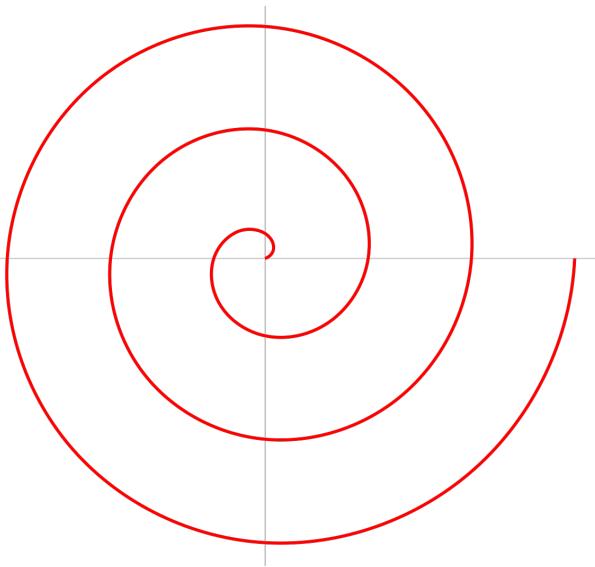


Principal Component Analysis (PCA)

- Applications: **Dimensionality reduction**

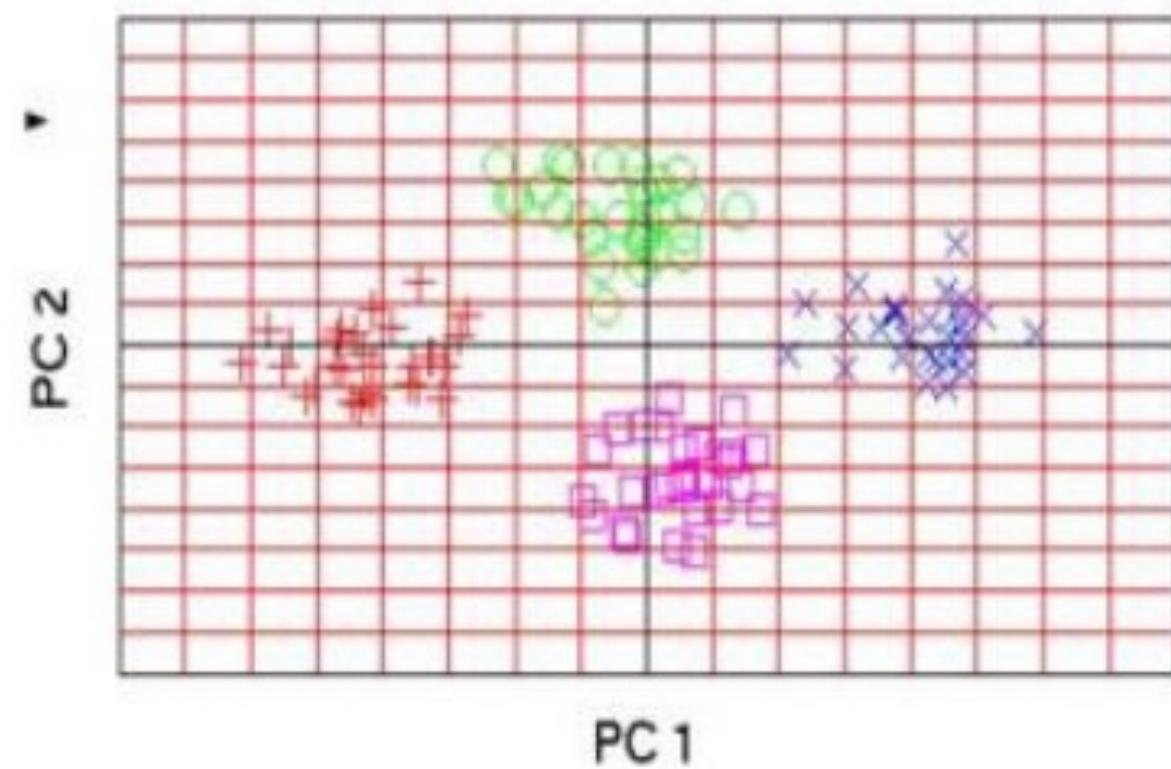
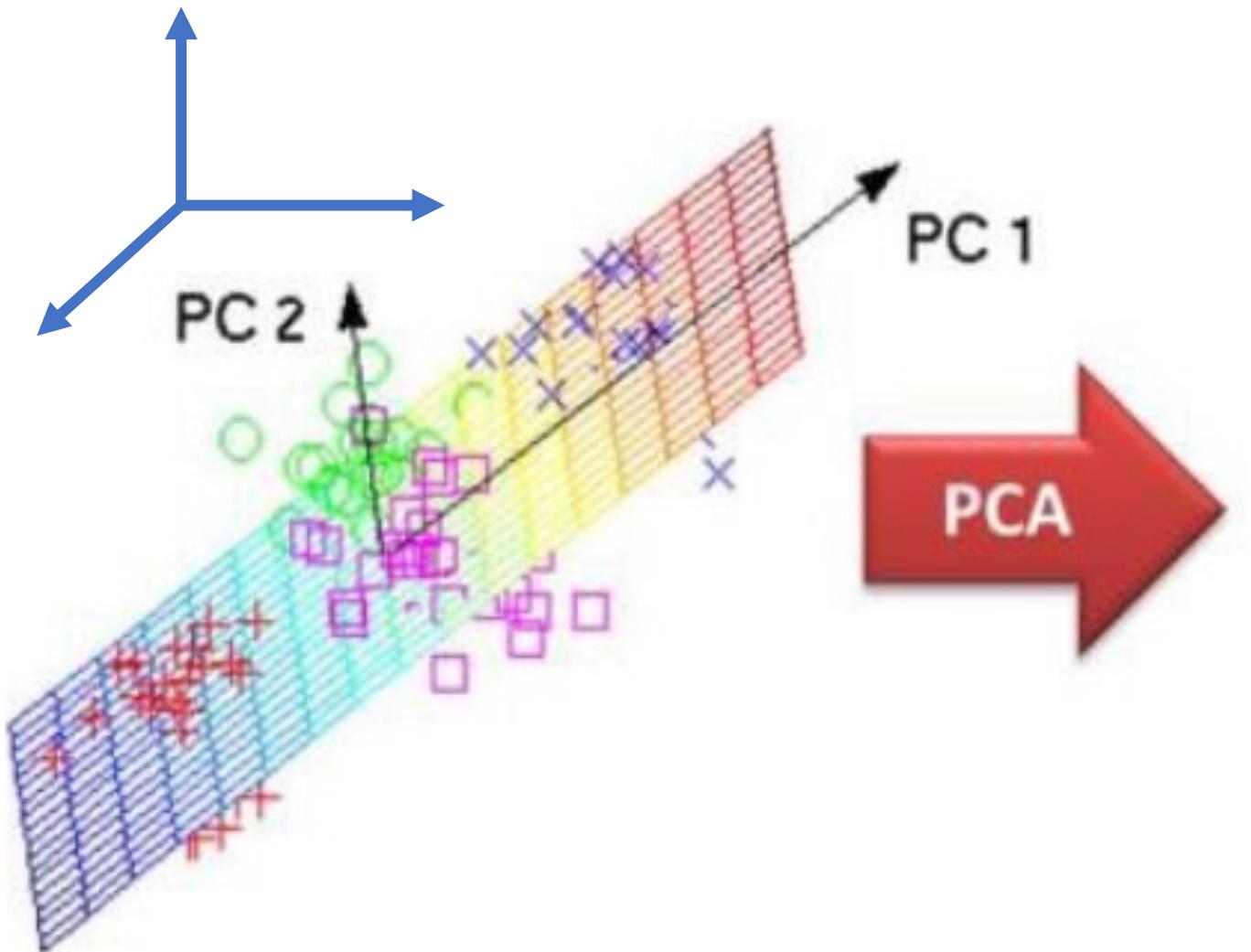
- Intrinsic dimension: Minimum number of variables (degrees of freedom) required to represent the signal

- Consider a multivariate random vector X of N scalar variables: $x = (x_1, \dots, x_N)$
- Consider a function $g(\cdot)$, and $M < N$ scalar variables a_1, \dots, a_M such that every $x \sim P(X)$ can be written as $x = g(a_1, \dots, a_M)$ for some a_1, \dots, a_M , then signal X needs only M variables for representation
- Here, “intrinsic dimension” of X is M , instead of the “representation dimension” = N



Principal Component Analysis (PCA)

- Applications: Dimensionality reduction

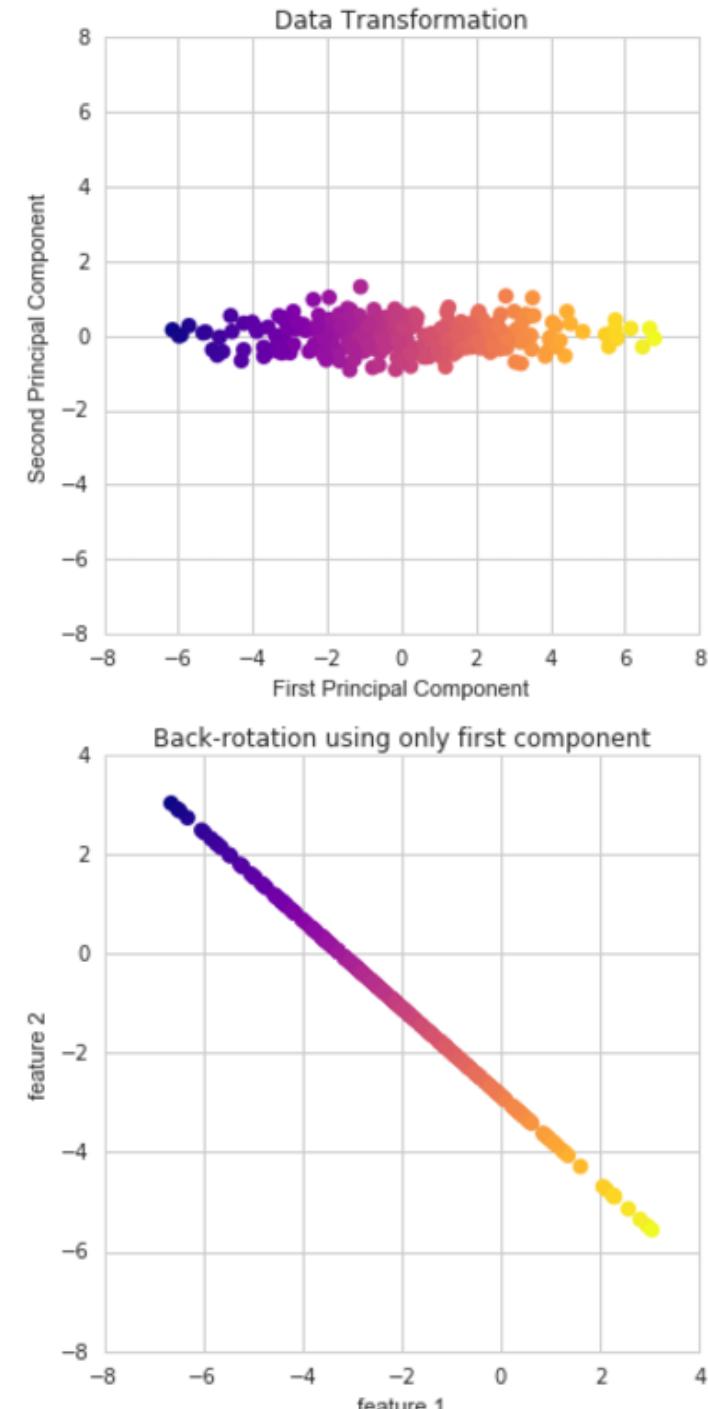
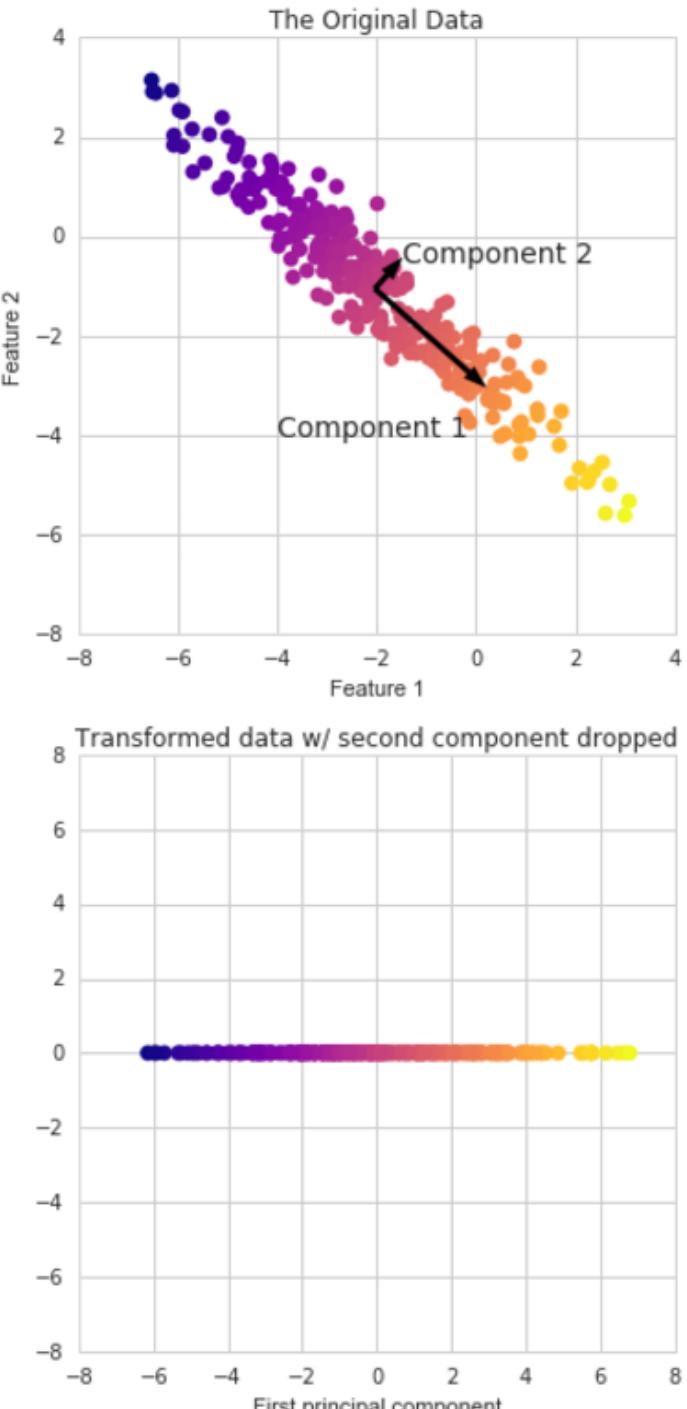


Principal Component Analysis (PCA)

- Applications: Dimensionality reduction
 - Acquired data is corrupted with errors
 - e.g., measurement errors
 - Such errors make the signal representation seem to be of a dimension higher than intrinsic dimension
 - Dimensionality reduction:
Transformation of data
from a higher-dimensional space into a lower-dimensional space
so that
lower-D representation (ideally close to its intrinsic dimension)
retains some meaningful properties of original data,
- PCA can perform **linear** dimensionality reduction

Principal Component Anal

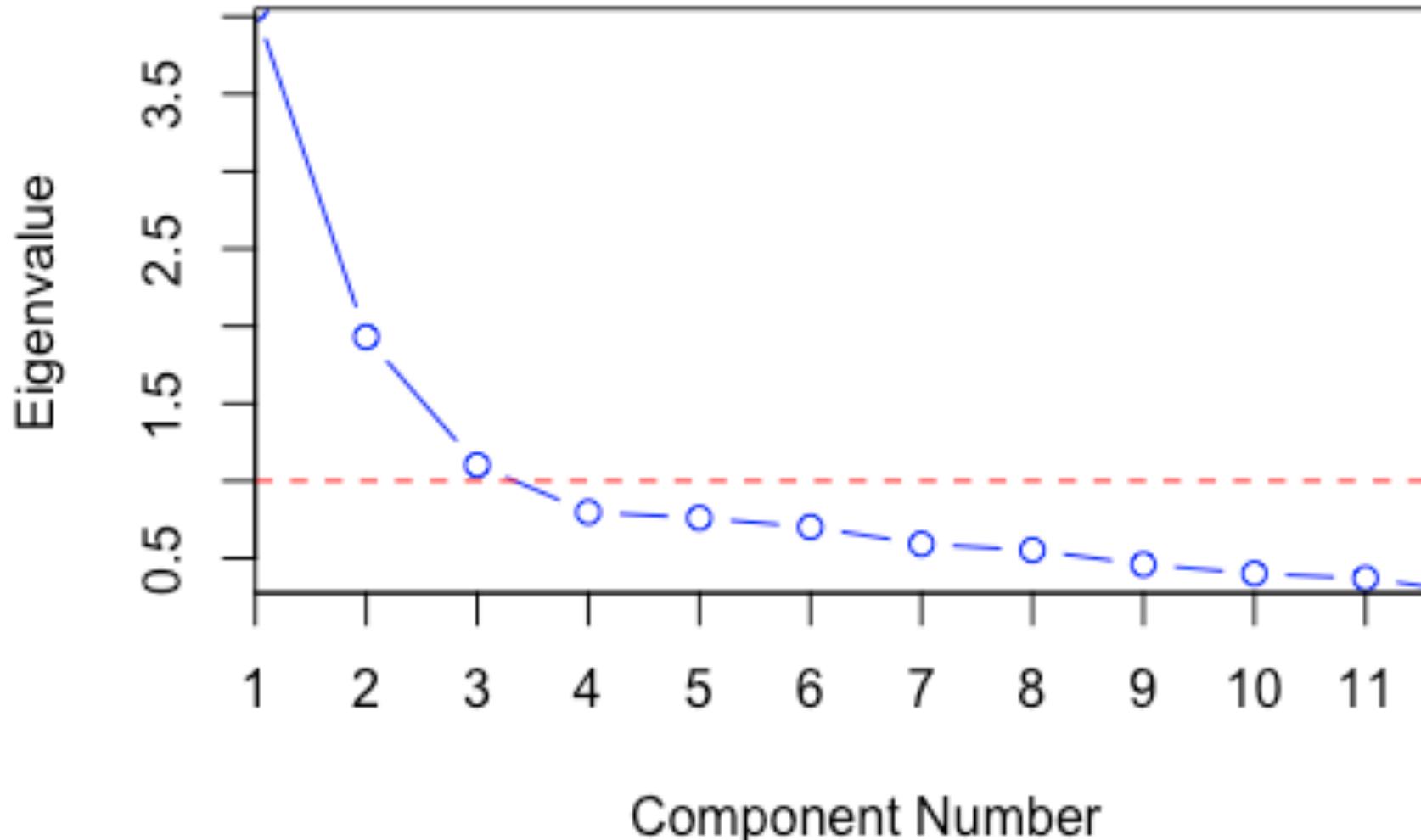
- Applications:
Dimensionality reduction
 - Using PCA
 - X may be N dimensional
 - PCA finds an M-dimensional space that captures most of the variability (total dispersion) in the data



Principal Component Analysis (PCA)

- Applications: Dimensionality reduction

- Using PCA
- X may be N dimensional
- PCA can find an M -dimensional space (often when $M \ll N$) that captures most of the variability (total dispersion) in the data

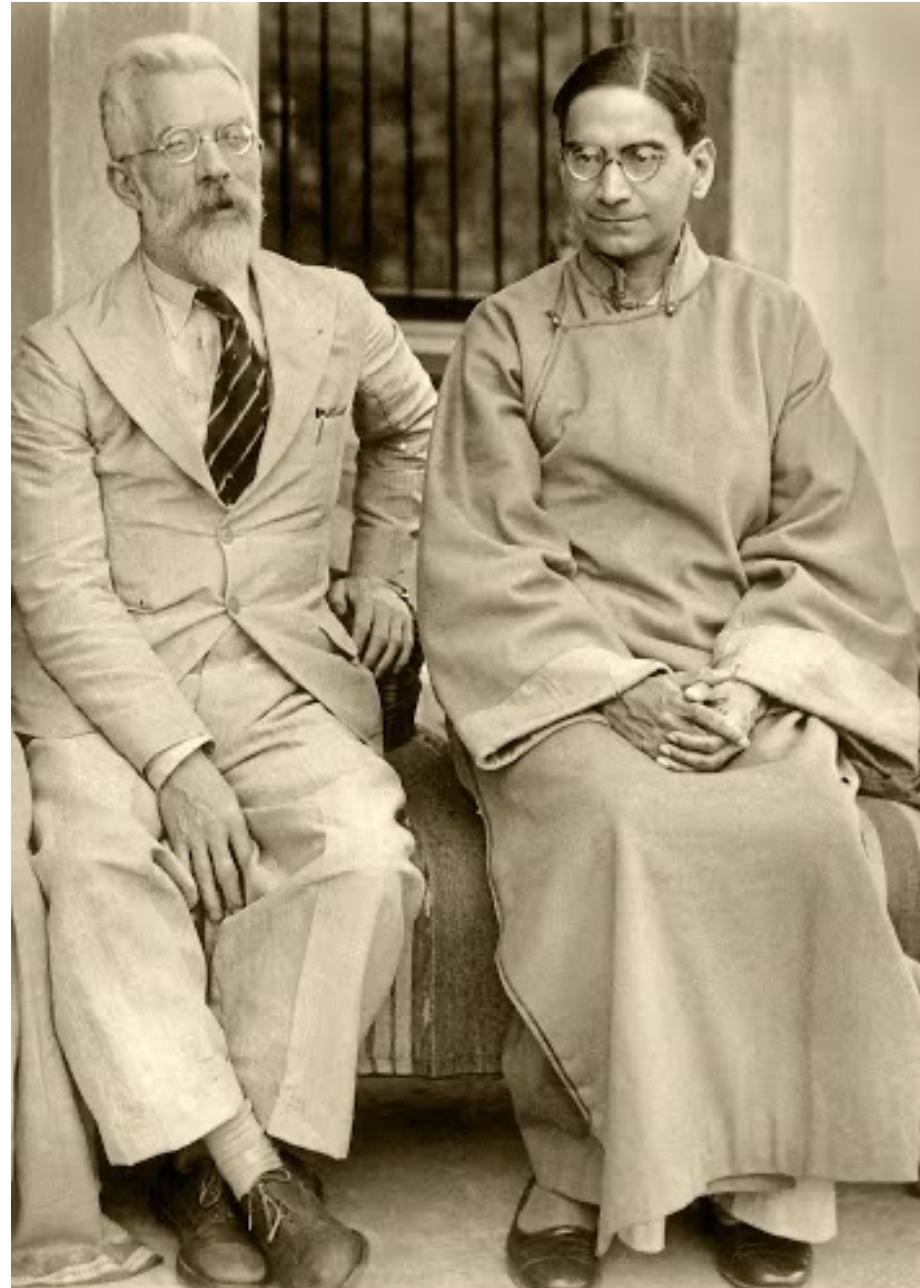


Multivariate Gaussian – Mahalanobis Distance

Multivariate Gaussian – Mahalanobis Distance

$$p(y) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$$

- Term $(y-\mu)^\top C^{-1} (y-\mu)$ appearing in exponent
= squared Mahalanobis distance
of point y from mean μ
- $d(y, \mu; C)^2 := (y-\mu)^\top C^{-1} (y-\mu)$ (where C is SPD)
- Prasanta Chandra Mahalanobis
founded
Indian Statistical Institute (ISI) in Kolkata



Multivariate Gaussian – Mahalanobis Distance

- $d(y, \mu; C)^2 := (y - \mu)^\top C^{-1} (y - \mu)$ (where C is SPD)
- Generalizes Euclidean distance in a multidimensional space
- When C is Identity:
 - Mahalanobis distance = Euclidean distance
- When C is diagonal:
 - Mahalanobis distance rescales “units” along each dimension based on standard deviation of the marginal along that dimension
- **A level set of a Multivariate Gaussian PDF is the locus of points with equal Mahalanobis distance from the mean**

$$p(y) = \frac{1}{(2\pi)^{D/2}|C|^{0.5}} \exp(-0.5(y - \mu)^\top C^{-1} (y - \mu))$$

Multivariate Gaussian – Mahalanobis Distance

- $d(y, \mu; C)^2 := (y - \mu)^T C^{-1} (y - \mu)$ (where C is SPD)
- Property: The Mahalanobis distance is a “distance metric”
- Proof:
 - A distance metric is a function $d(\dots) \rightarrow R$ that needs to satisfy 3 properties:
 - (1) identity of indiscernibles: $d(x, y) = 0$ iff $x = y$
 - (2) symmetry: $d(x, y) = d(y, x)$
 - (3) triangle inequality: $d(x, y) \leq d(x, z) + d(z, y)$
 - These imply non-negativity (i.e., $d(x, y) \geq 0$, for all x, y):
$$0 = d(x, x) \leq d(x, y) + d(y, x) = 2 d(x, y)$$
 - In our case of SPD matrix C :
 - C being SPD implies: $d(x, y; C) \geq 0$ for all x, y
 - C being SPD implies: $d(x, y; C) = 0$ iff $x = y$
 - Because of the specific quadratic form of $d(\dots; C)^2$: $d(x, y; C) = d(y, x; C)$

Multivariate Gaussian – Mahalanobis Distance

- Property: The Mahalanobis distance is a true distance metric
- Proof (when covariance matrix \mathbf{C} is diagonal):

3) Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z$

Let $u := x - z$ and $v := z - y$

Then $u + v = x - y$

$$\text{LHS} = \sqrt{(u + v)^\top C^{-1}(u + v)}$$

$$\text{RHS} = \sqrt{u^\top C^{-1}u} + \sqrt{v^\top C^{-1}v}$$

- Showing LHS \leq RHS is equivalent to showing $\text{LHS}^2 \leq \text{RHS}^2$

$$\text{LHS}^2 = (u + v)^\top C^{-1}(u + v)$$

$$= \sum_d (u_d + v_d)^2 / \sigma_d^2 \text{ (assuming } C \text{ is diagonal)}$$

$$= \sum_d u_d^2 / \sigma_d^2 + \sum_d v_d^2 / \sigma_d^2 + 2 \sum_d u_d v_d / \sigma_d^2$$

$$\begin{aligned}\text{RHS}^2 &= u^\top C^{-1}u + v^\top C^{-1}v + 2\sqrt{u^\top C^{-1}u}\sqrt{v^\top C^{-1}v} \\ &= \sum_d u_d^2 / \sigma_d^2 + \sum_d v_d^2 / \sigma_d^2 + 2\sqrt{\sum_d u_d^2 / \sigma_d^2} \sqrt{\sum_d v_d^2 / \sigma_d^2}\end{aligned}$$

Multivariate Gaussian – Mahalanobis Distance

- Property: The Mahalanobis distance is a true distance metric
- Proof (when covariance matrix \mathbf{C} is diagonal):

3) Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z$

$$\begin{aligned}\text{LHS}^2 &= (u + v)^\top C^{-1}(u + v) \\ &= \sum_d (u_d + v_d)^2 / \sigma_d^2 \text{ (assuming } C \text{ is diagonal)} \\ &= \sum_d u_d^2 / \sigma_d^2 + \sum_d v_d^2 / \sigma_d^2 + 2 \sum_d u_d v_d / \sigma_d^2\end{aligned}$$

$$\begin{aligned}\text{RHS}^2 &= u^\top C^{-1}u + v^\top C^{-1}v + 2\sqrt{u^\top C^{-1}u}\sqrt{v^\top C^{-1}v} \\ &= \sum_d u_d^2 / \sigma_d^2 + \sum_d v_d^2 / \sigma_d^2 + 2\sqrt{\sum_d u_d^2 / \sigma_d^2}\sqrt{\sum_d v_d^2 / \sigma_d^2}\end{aligned}$$

The first 2 terms in LHS and RHS are same !

Let $a_d = u_d / \sigma_d$ and $b_d = v_d / \sigma_d$

Last term in LHS = $2\langle a, b \rangle$

Last term in RHS = $2 \| a \| \| b \|$

Now, we know that $\langle a, b \rangle \leq |\langle a, b \rangle|$ (holds for any scalar)

And the Cauchy-Schwartz inequality tells us that $|\langle a, b \rangle| \leq \| a \| \| b \|$ for any $a, b \in \mathbb{R}^D$

Multivariate Gaussian – Mahalanobis Distance

- Property: The Mahalanobis distance is a true distance metric
- Proof (for a **general covariance matrix C**):

3) Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z$

For a non-diagonal C, write $C = Q\Lambda Q^\top$ and define $u := Q^\top(x - z)$ and $v := Q^\top(z - y)$ and proceed as before.

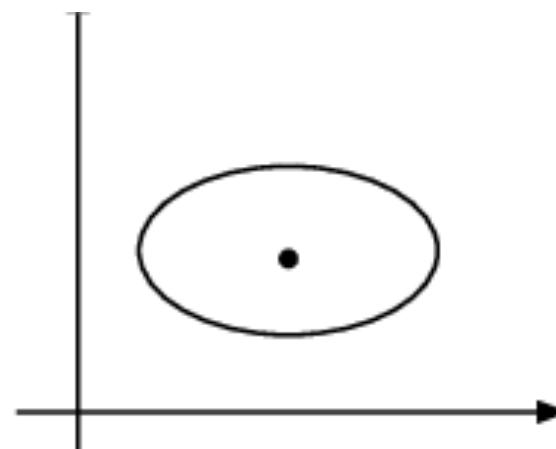
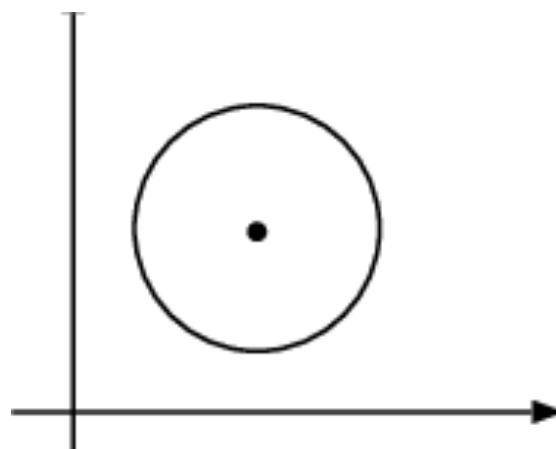
Multivariate Gaussian – Mahalanobis Distance

- A level set of a Multivariate Gaussian is the locus of points with the same Mahalanobis distance from the mean
- Scaling the coordinate frame: $X := SW$
 - How does the Mahalanobis distances change (w.r.t. case when $C = \text{Identity}$) ?
 - How do the level sets change ?

Let $X := SW$, where S is a diagonal matrix that rescales the units along each coordinate axes

Then, what is the covariance matrix ? $A = S$. Thus, $C = AA^\top = SS^\top = S^2$

Then, Mahalanobis distance between x and the mean (origin) is $x^\top C^{-1}x = x^\top S^{-2}x$



Multivariate Gaussian – Mahalanobis Distance

- A level set of a Multivariate Gaussian is the locus of points with the same Mahalanobis distance from the mean
- Scaling + “Rotating” (proper + improper) coordinate frame: $Y := USW$
 - How does the Mahalanobis distances change (w.r.t. case when $C = \text{Identity}$) ?
 - How do the level sets change ?

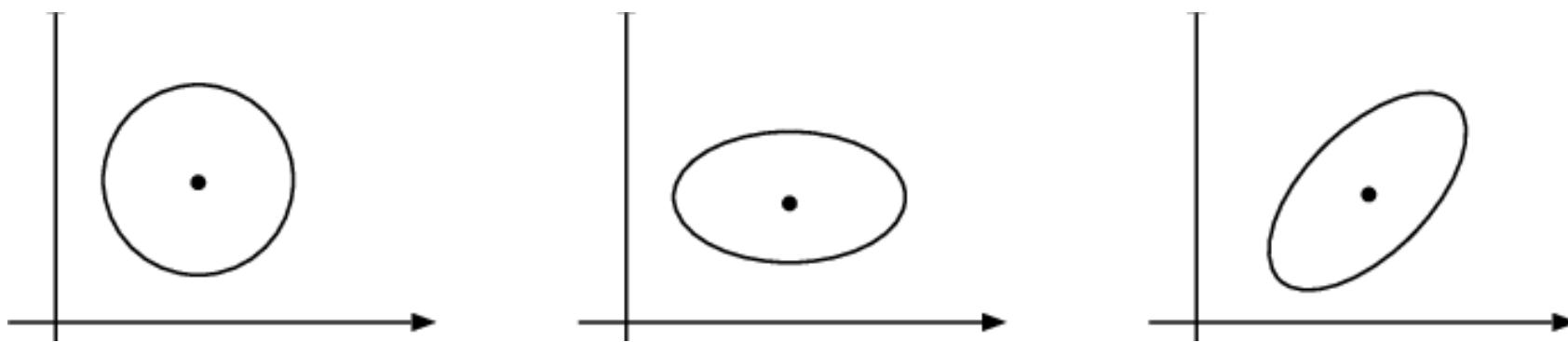
Let $Y := UX = USW$, where U is a rotation matrix that rotates the coordinate frame

Then, what is the covariance matrix ? $A = US$. Thus, $C = AA^\top = (US)(US)^\top = US^2U^\top$

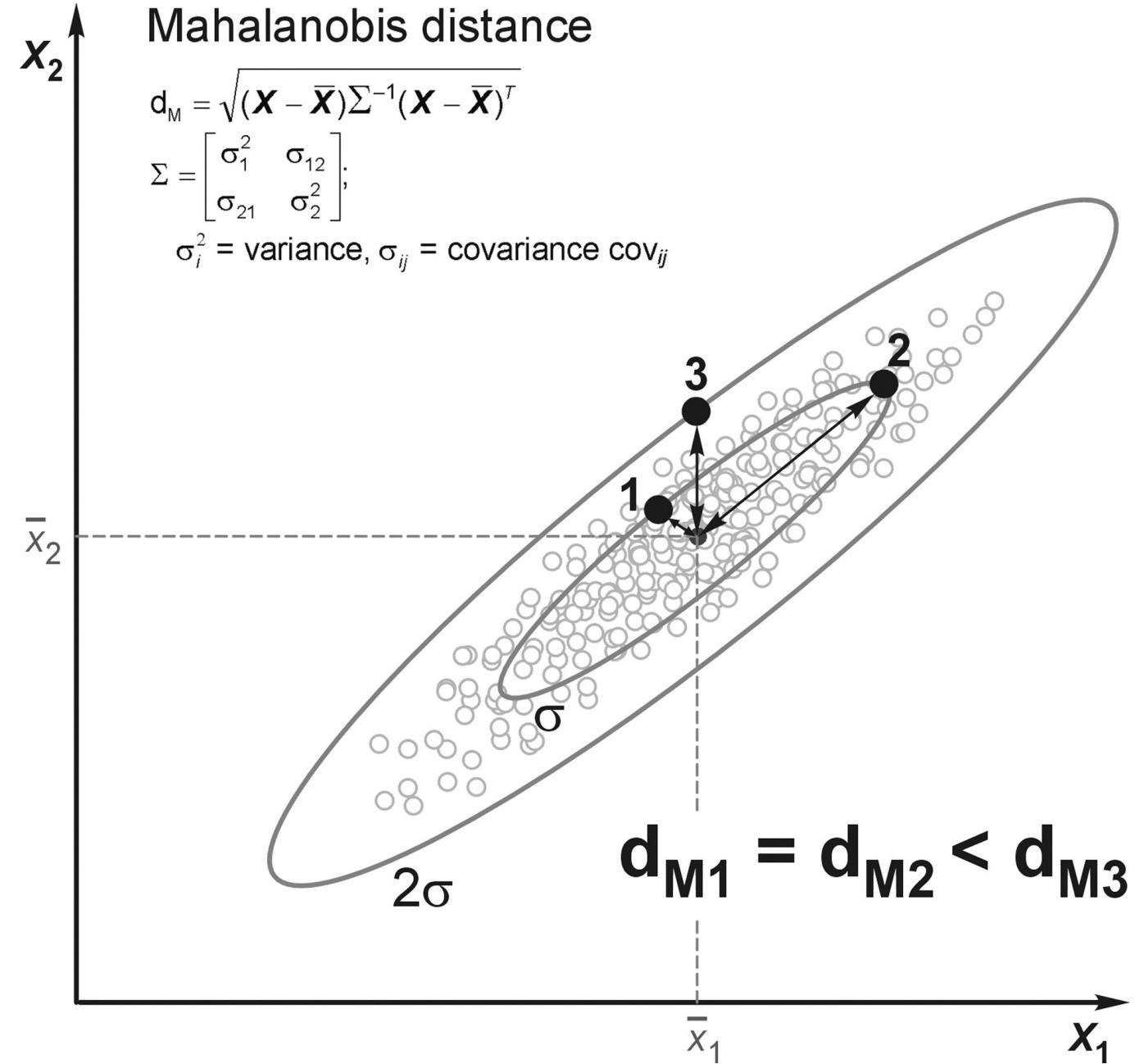
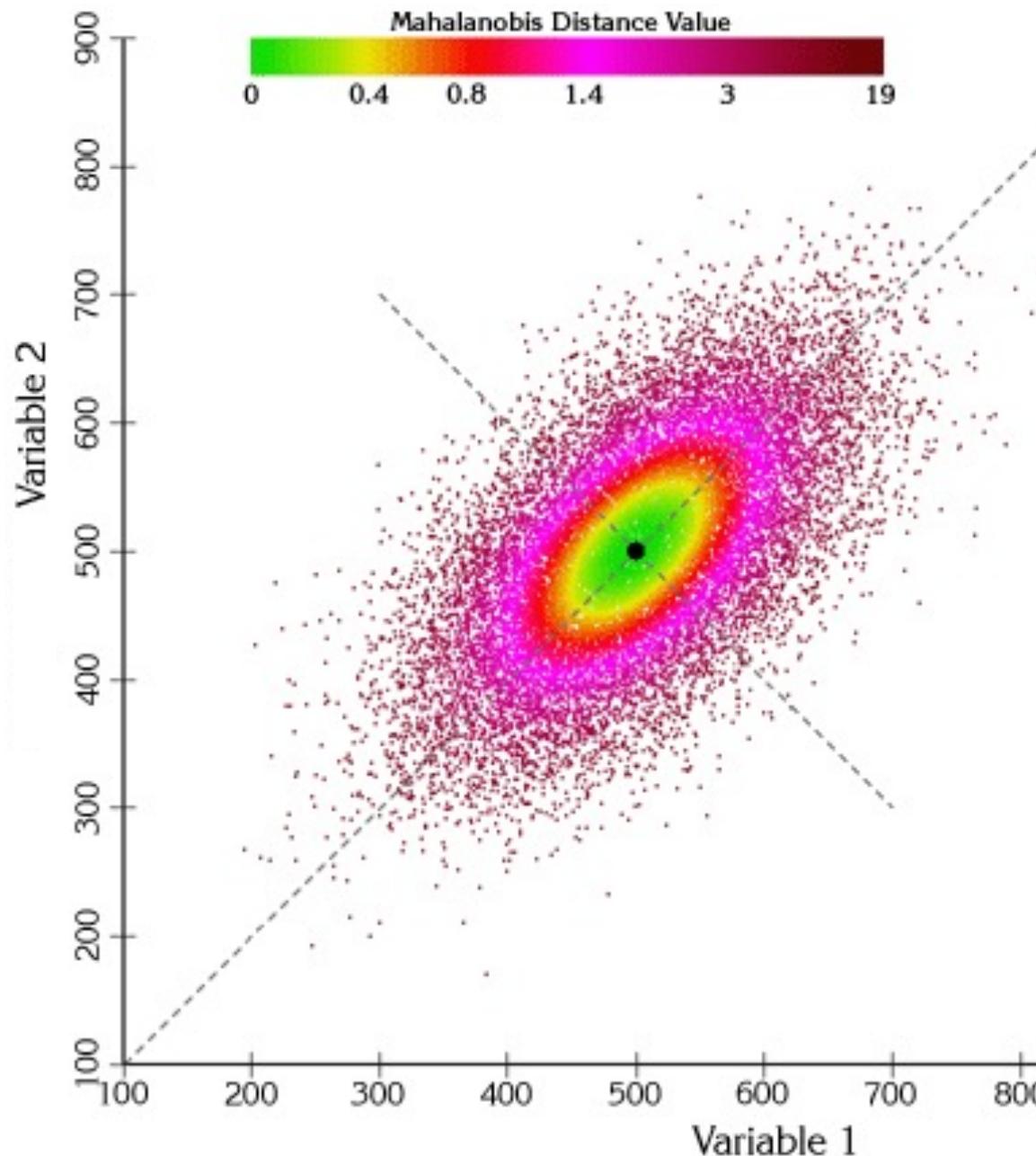
Then, Mahalanobis distance between $y := Ux$ and the mean (origin) is

$y^\top C^{-1}y = (Ux)^\top (US^{-2}U^\top)(Ux) = x^\top S^{-2}x$, which is the same as before !

Thus, rotating the data x simply rotates the iso-probability contours of $P(X)$.



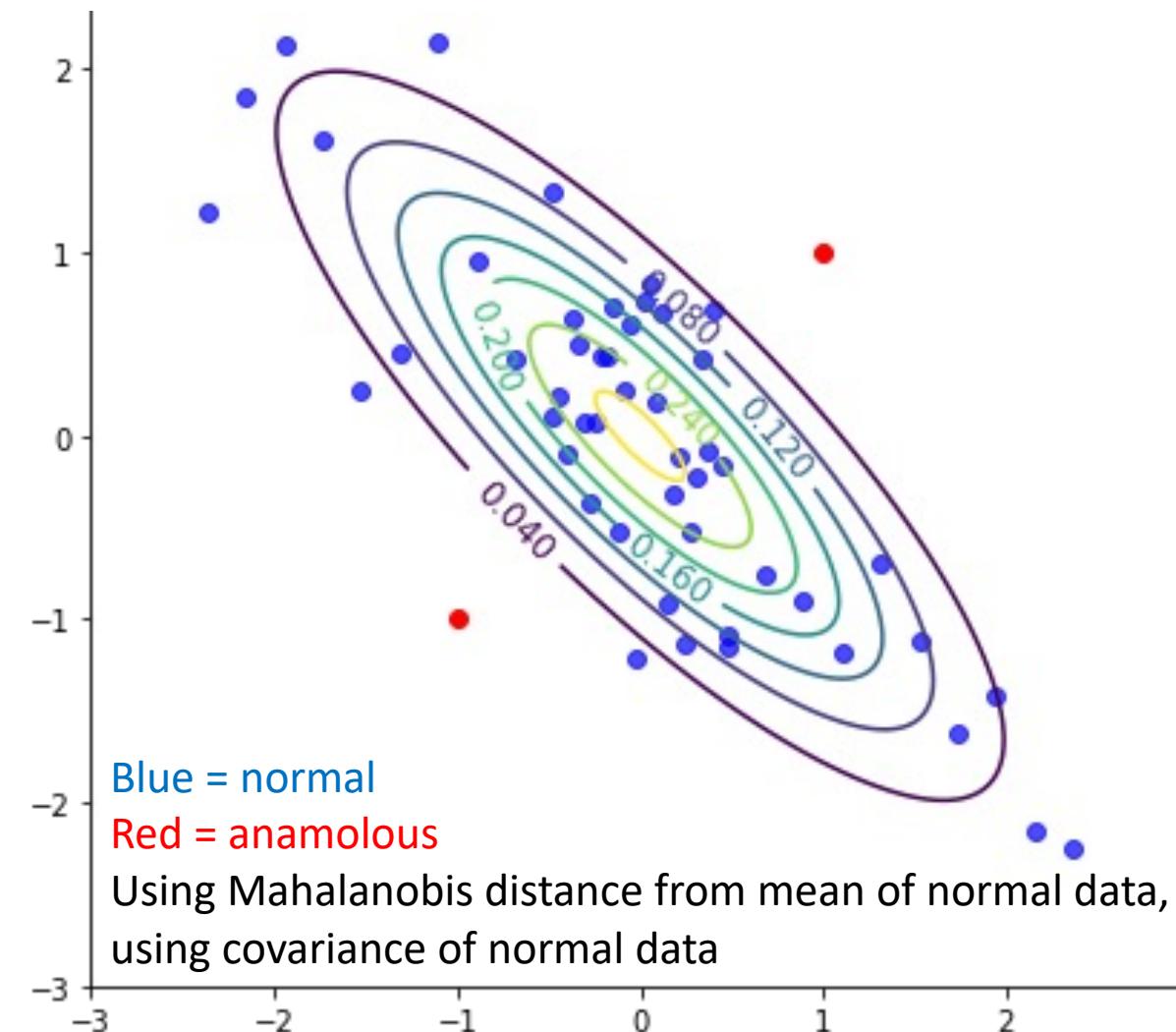
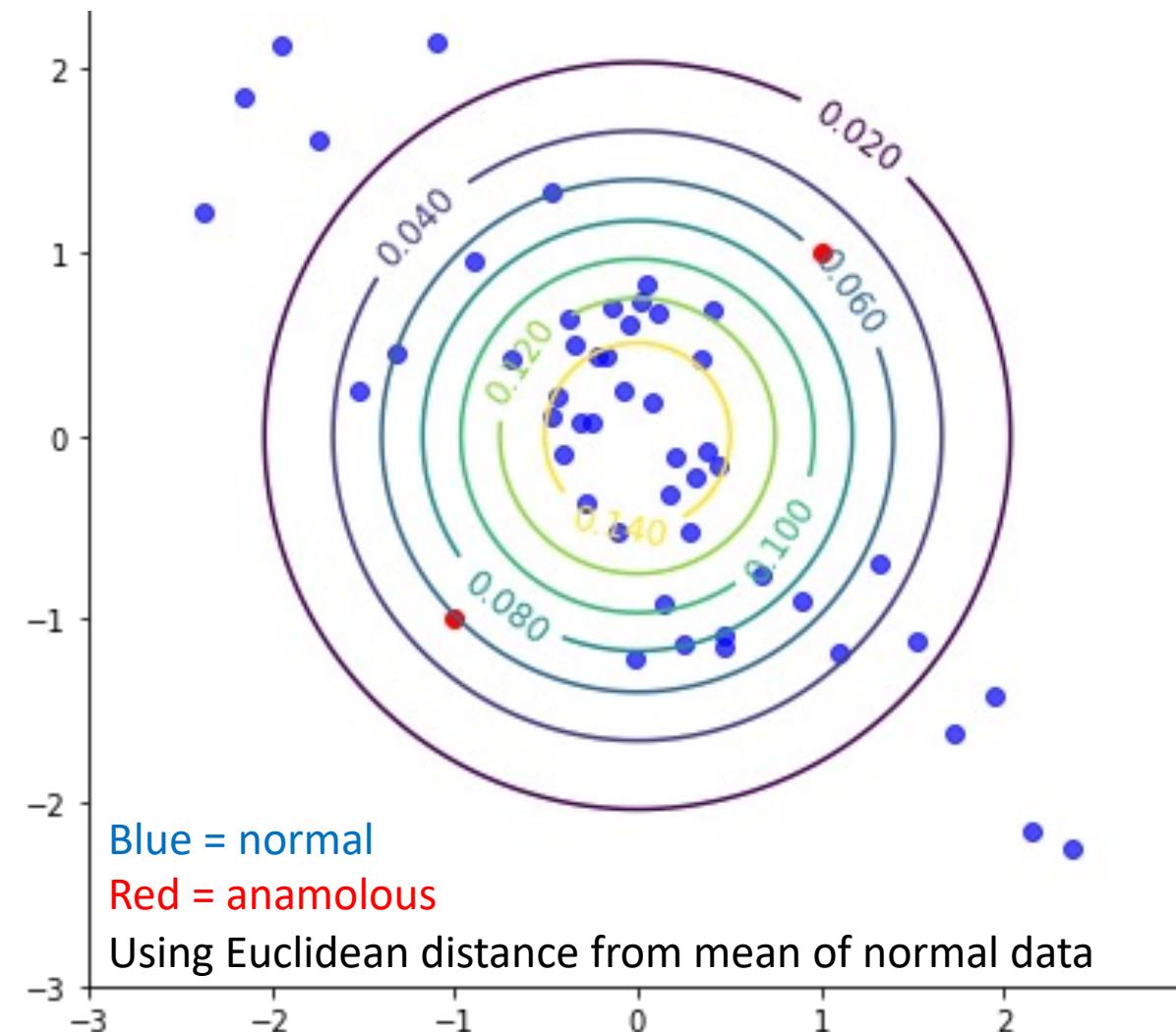
Multivariate Gaussian – Mahalanobis Distance



Multivariate Gaussian – Applications

Multivariate Gaussian – Applications

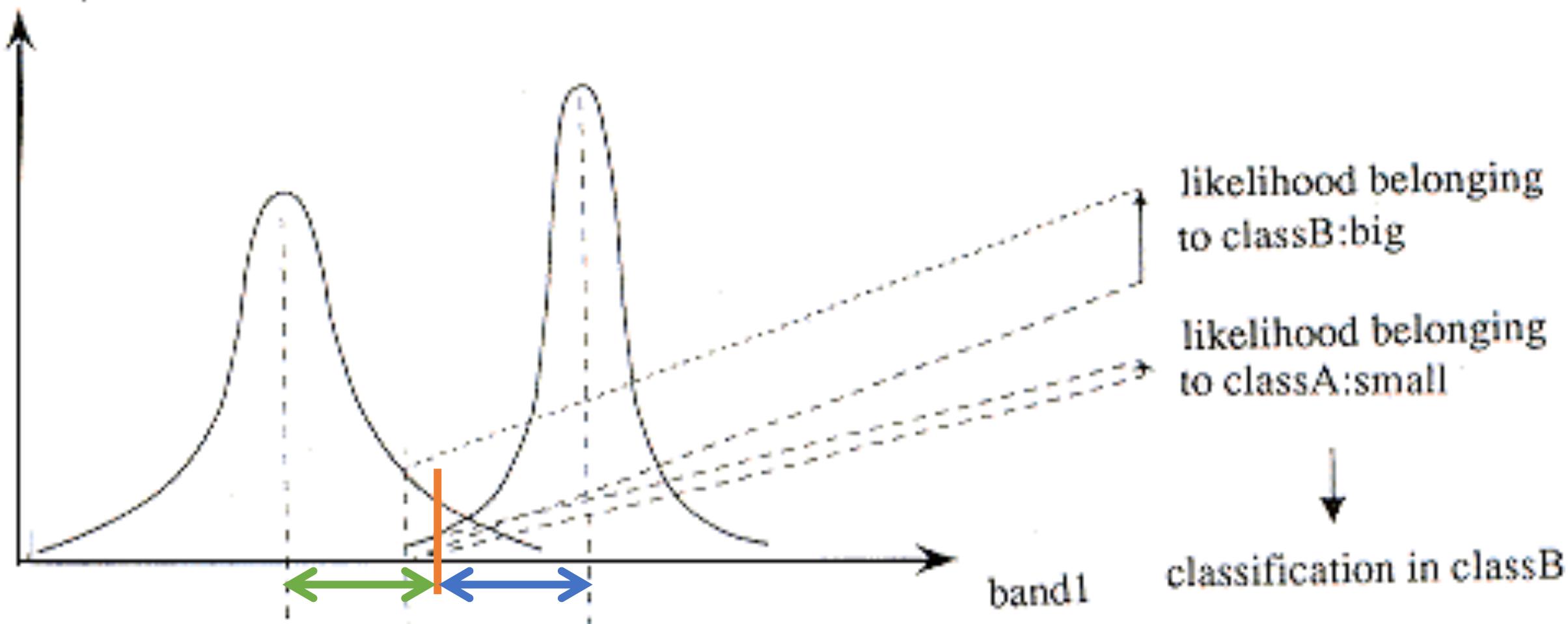
- Multivariate Gaussian (Mahalanobis distance) for **anomaly detection**



Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood **classification (in 1D)**

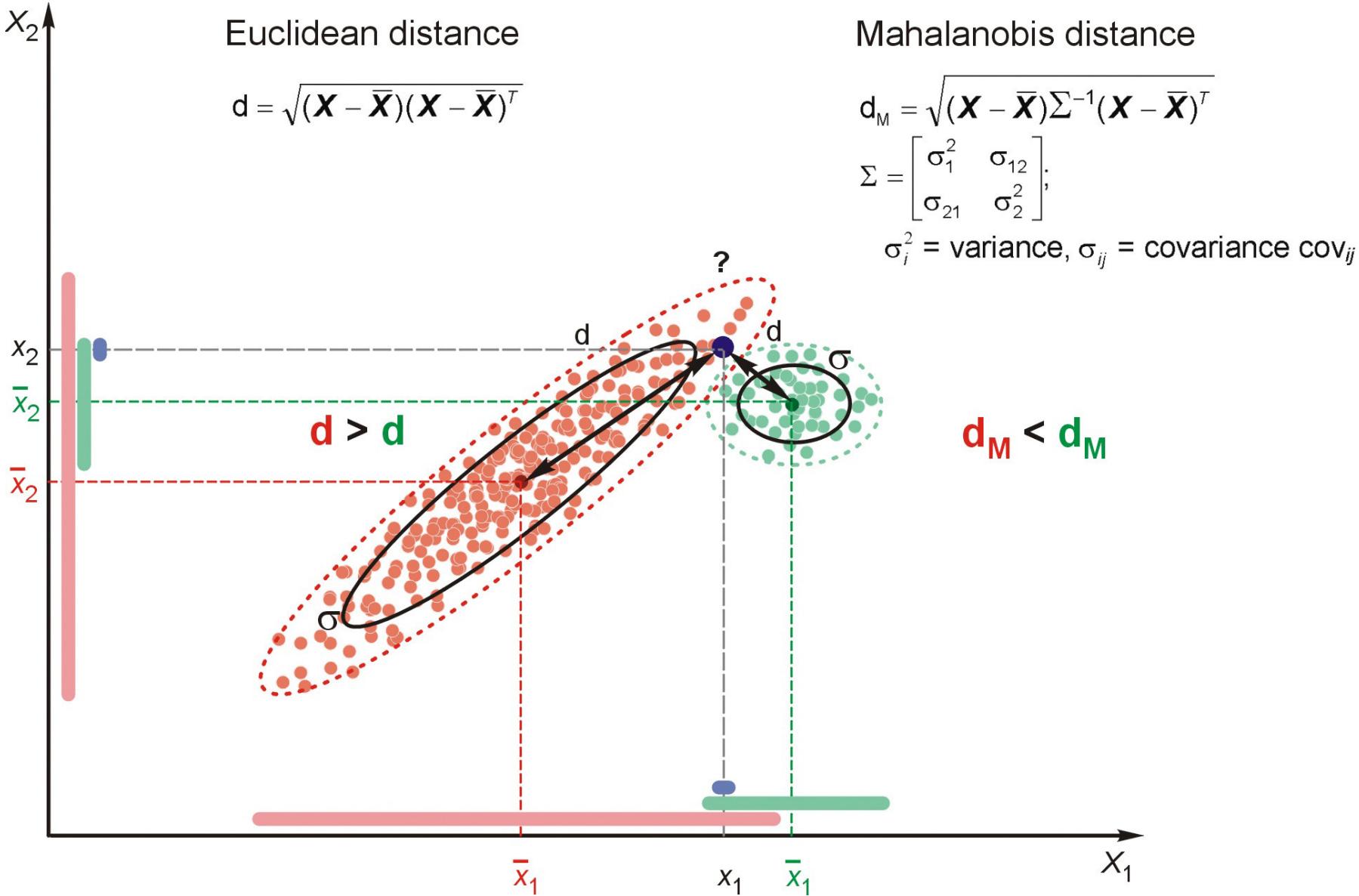
probability density



Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood **classification** (in 2D)

- If $\det(C_1) = \det(C_2)$, then likelihood-based classification is equivalent to Mahalanobis-distance based classification

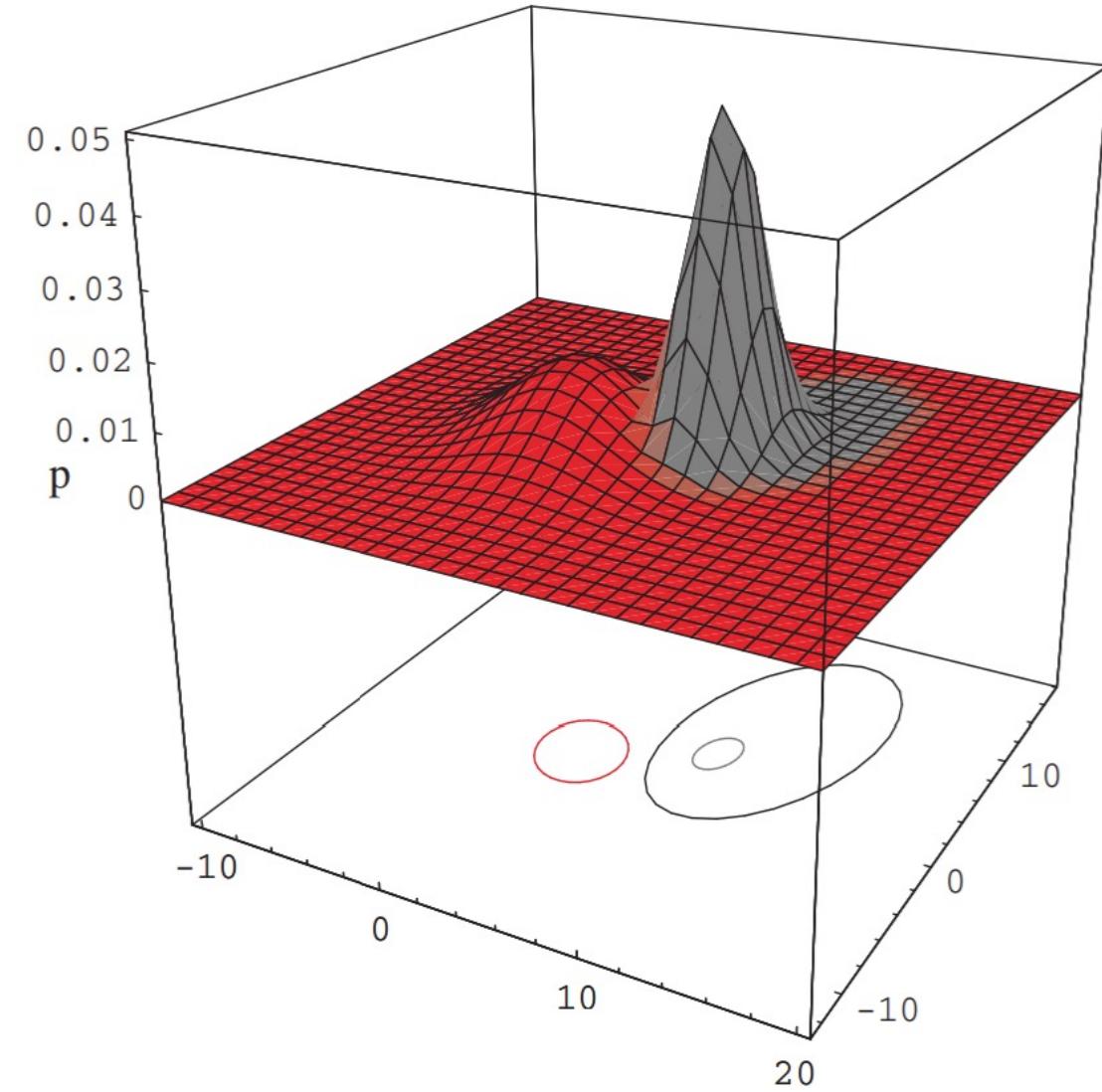
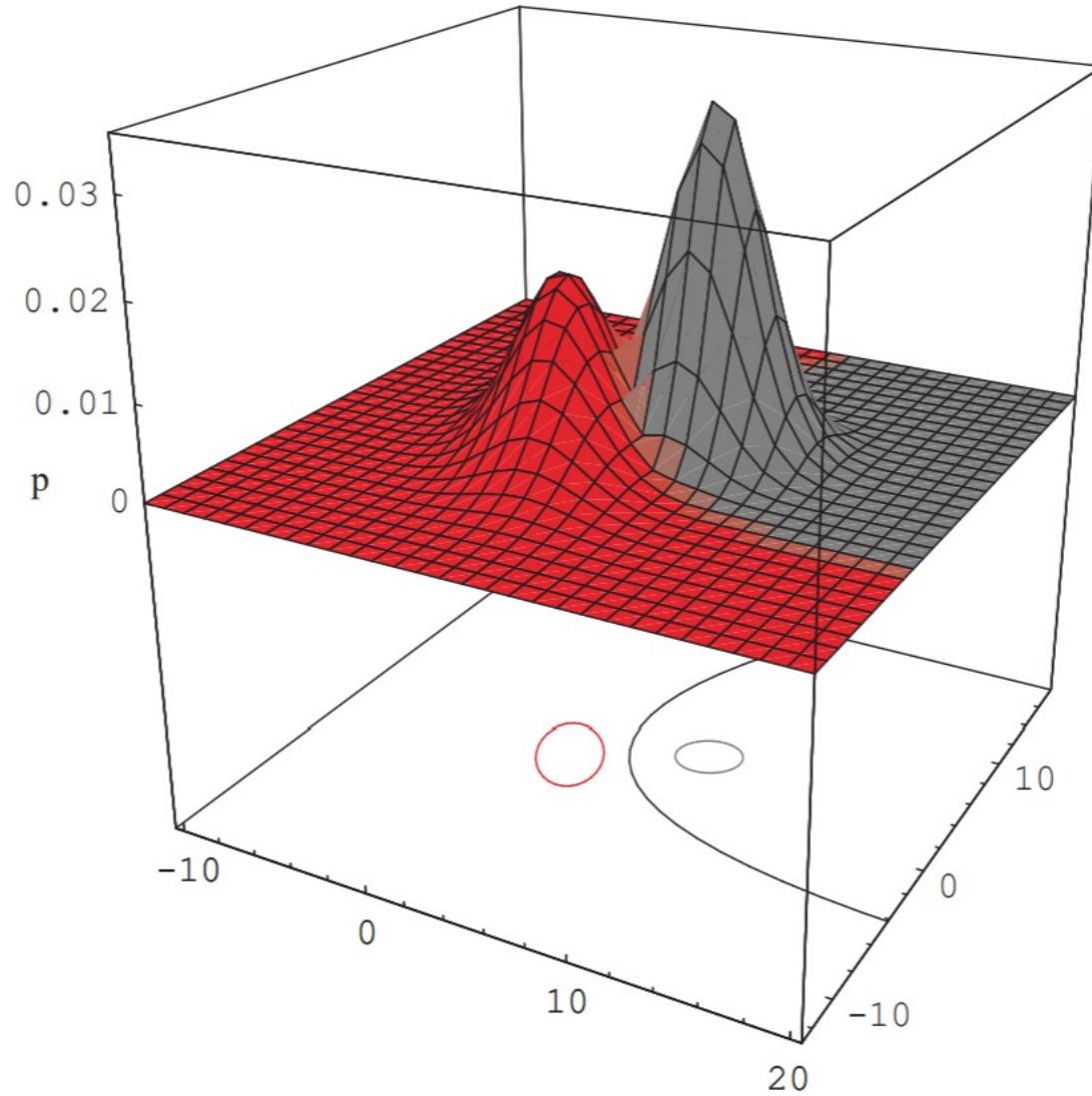


Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood classification
 - How do decision boundaries look like ?
 - $P(x|Class_1) = G(x; m_1, C_1)$
 - $P(x|Class_2) = G(x; m_2, C_2)$
 - Decision surface comprises all points ‘x’ at which likelihoods are equal
 - $\{ x : P(x|Class_1) = P(x|Class_2) \}$
 - $\{ x : 0 = \log (P(x|Class_1) / P(x|Class_2)) \}$
 - At any point in the domain ‘x’, the log likelihood-ratio is:
 $\log (P(x|Class_1) / P(x|Class_2))$
=
 $- 0.5 (x-m_1)^T C_1^{-1} (x-m_1) - 0.5 \log (\det (C_1))$
 $+ 0.5 (x-m_2)^T C_2^{-1} (x-m_2) + 0.5 \log (\det (C_2))$
 - In general, decision surface is a hyper-quadric

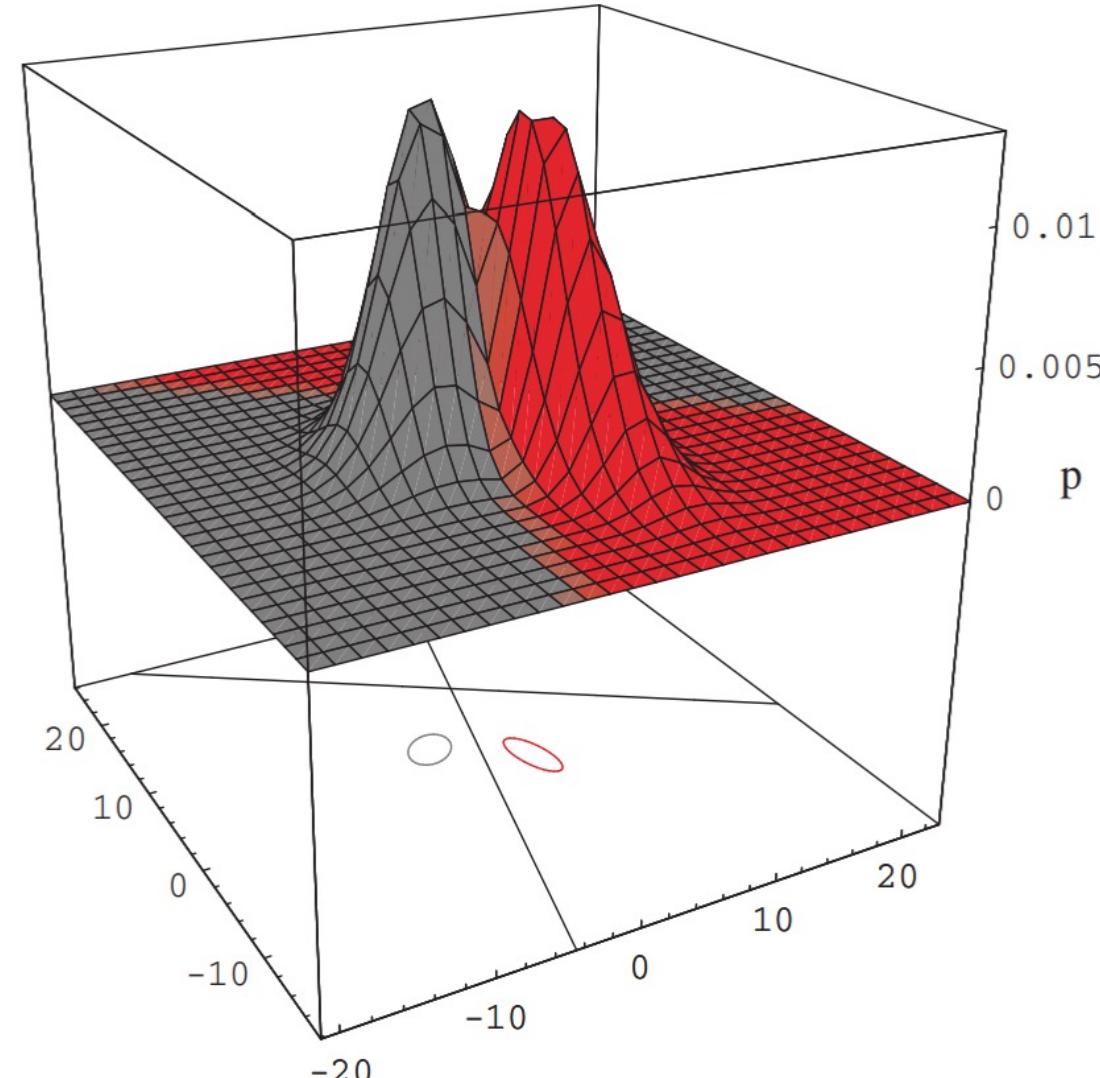
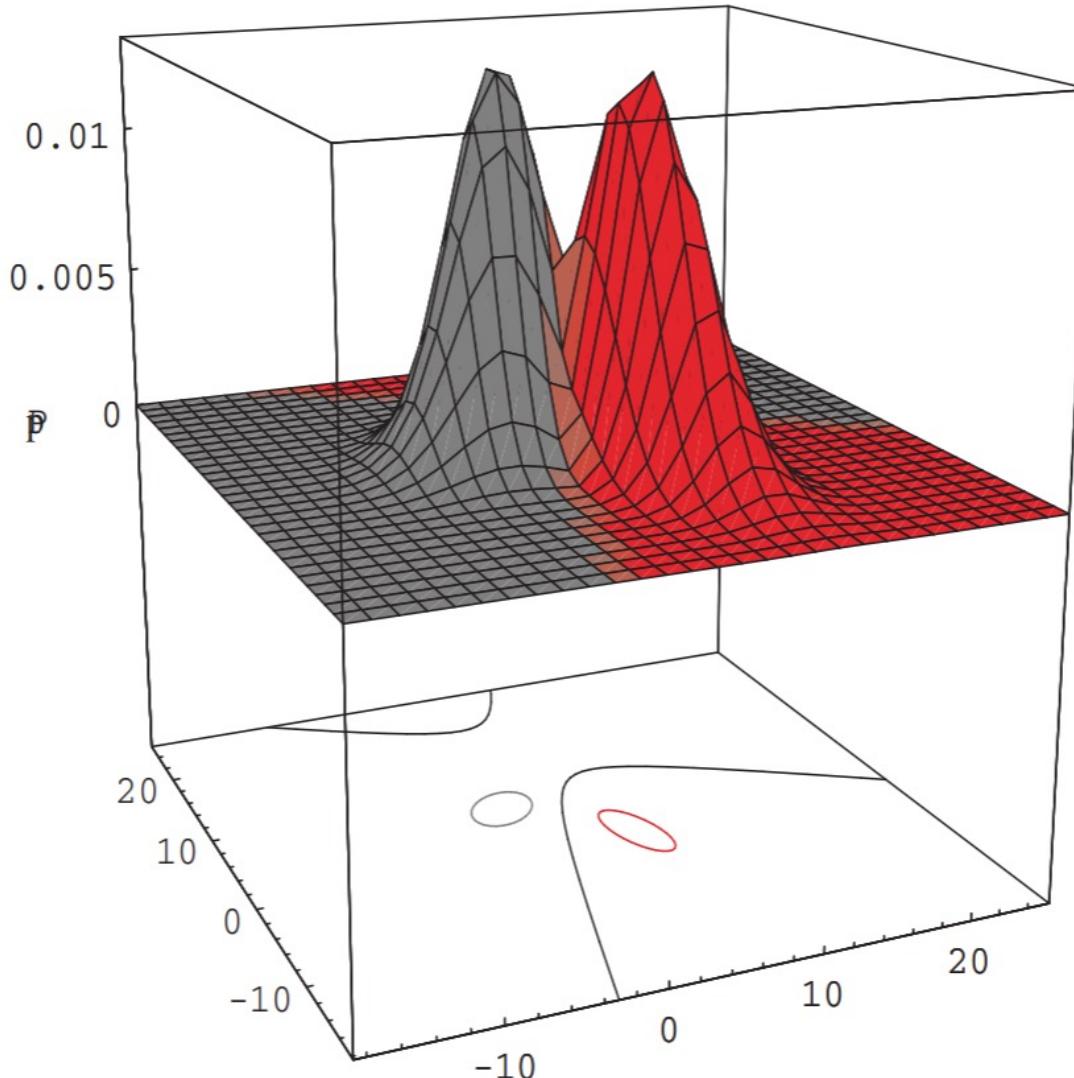
Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood classification
 - Decision boundaries



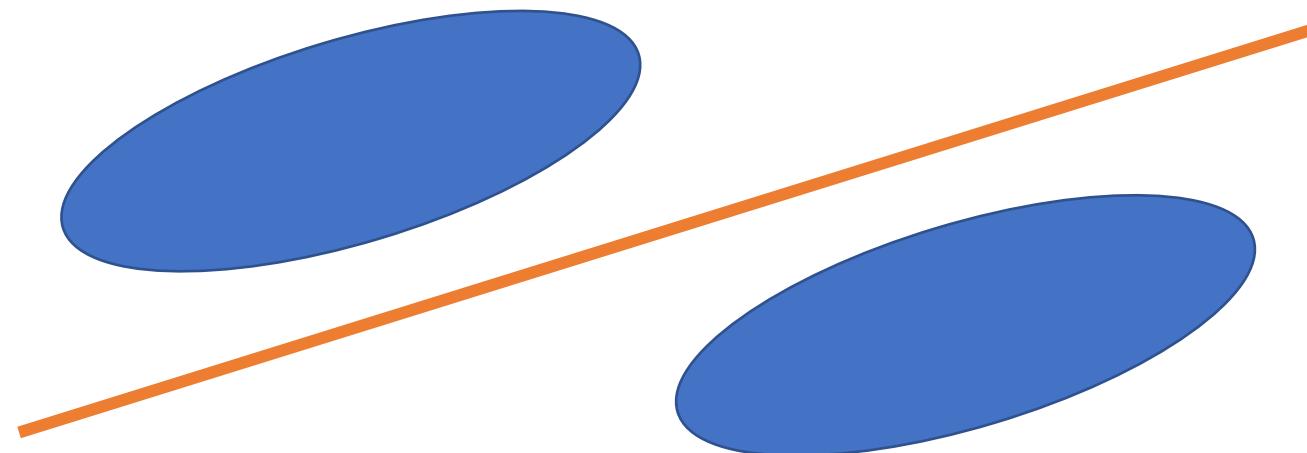
Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood classification
 - Decision boundaries



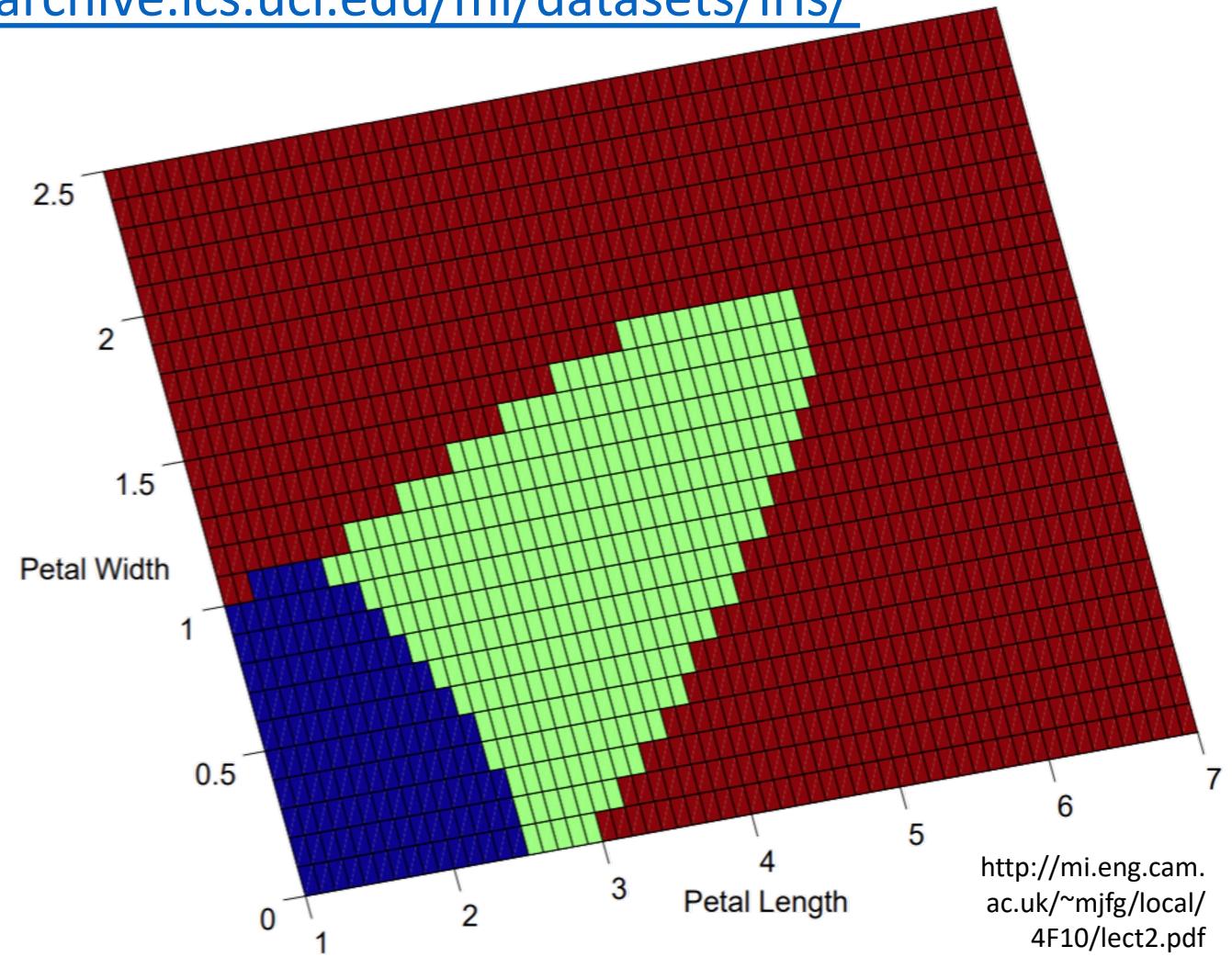
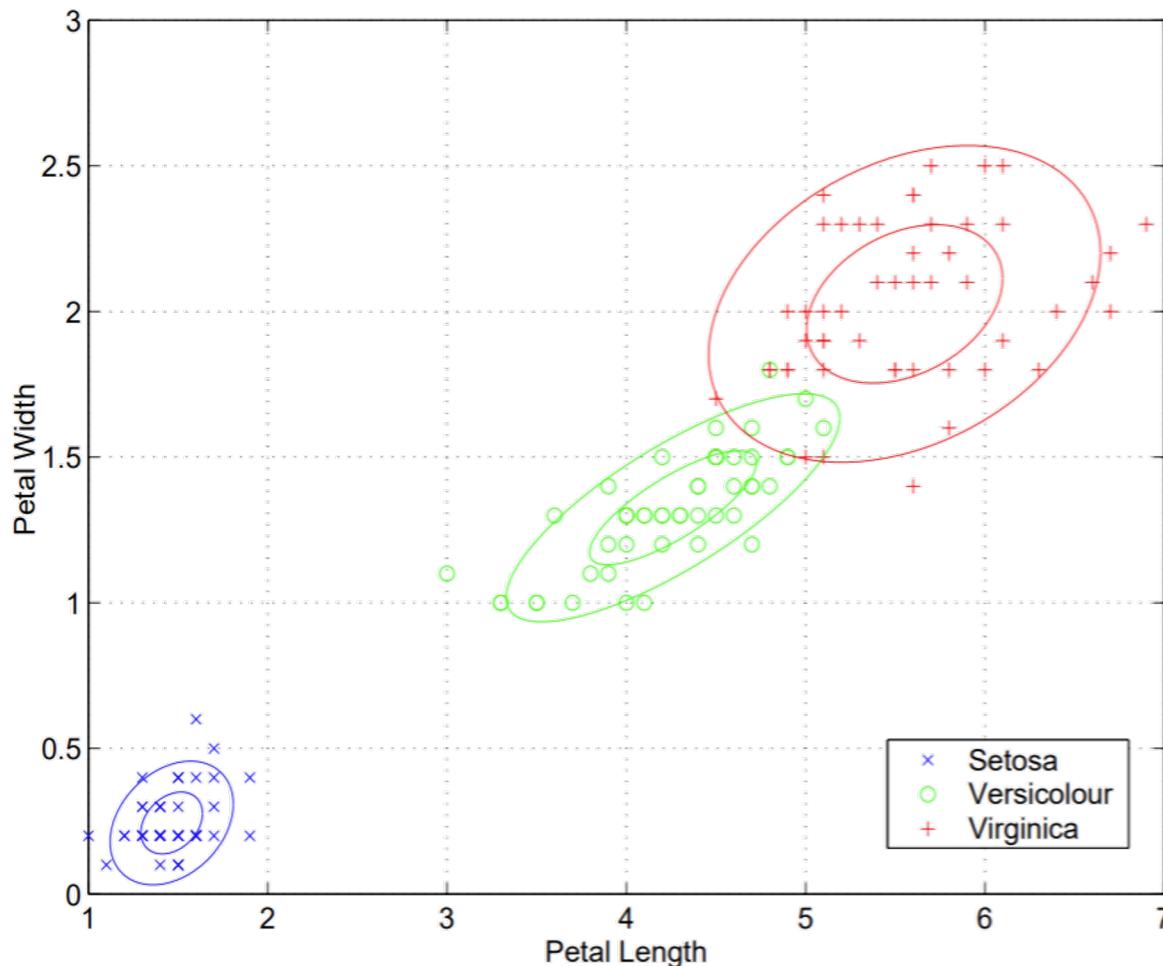
Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood classification
 - Decision boundaries
 - When $C_1 = C_2 = C$, then decision boundary is:
 - $0 = \log (P(x|Class_1) / P(x|Class_2))$
= $-0.5 (x-m_1)^T C^{-1} (x-m_1)$
 $+ 0.5 (x-m_2)^T C^{-1} (x-m_2)$
 - 0
= $+ (m_2-m_1)^T C^{-1} x$
 $+ 0.5 m_1^T C^{-1} m_1$
 $- 0.5 m_2^T C^{-1} m_2$
 - Decision surface is a hyper-plane



Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood classification
 - Example (Data taken from R. A. Fisher's classic [1936 paper](#))
 - UCI ML repository Iris dataset <http://archive.ics.uci.edu/ml/datasets/Iris/>



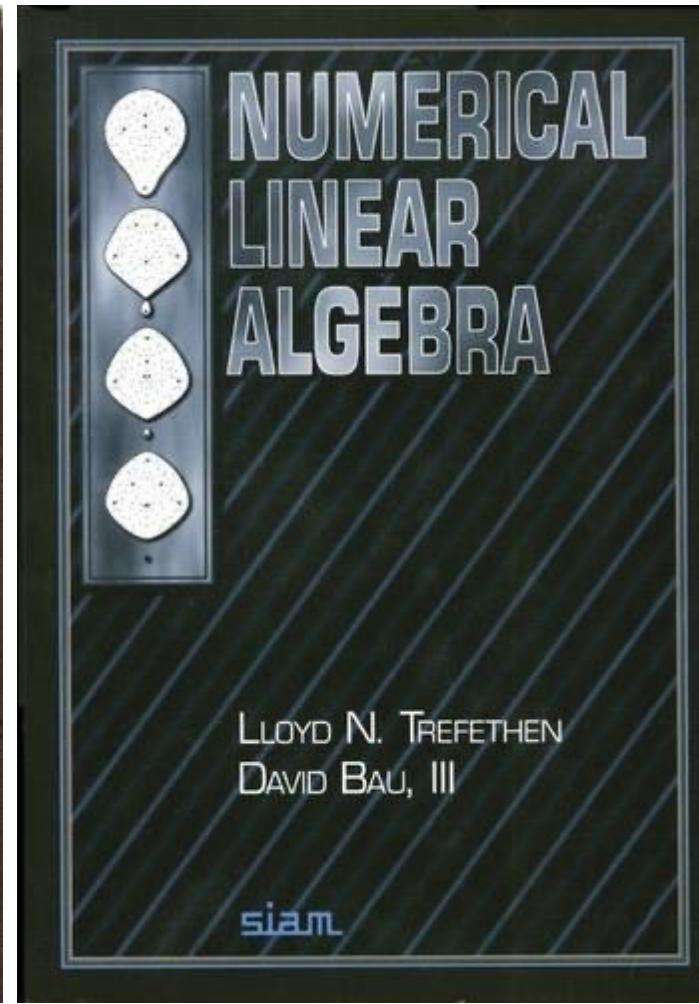
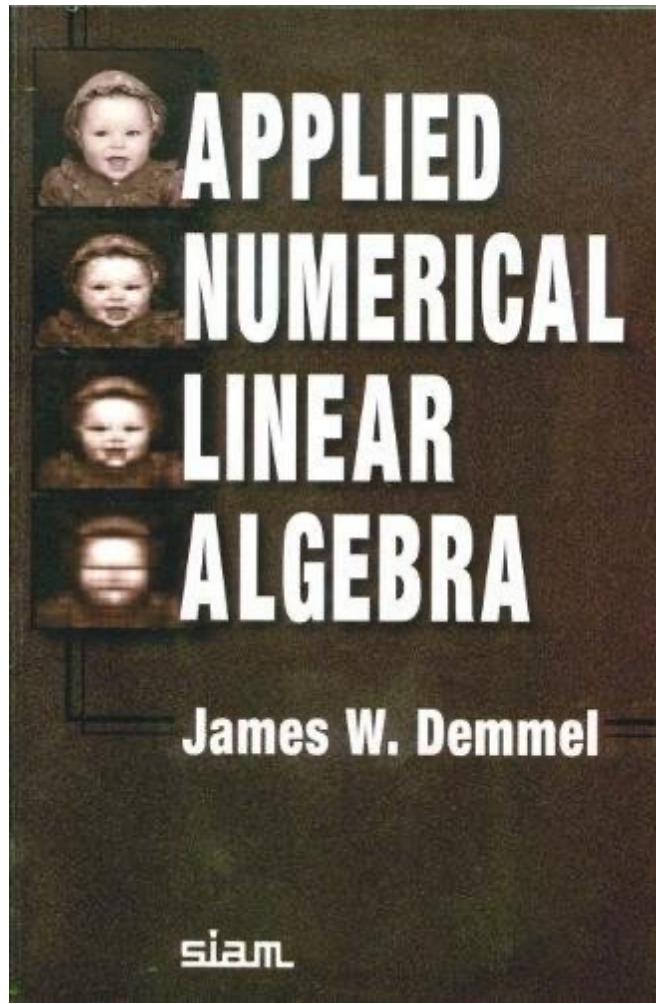
Datasets

- UCI Machine Learning Repository
 - <https://archive.ics.uci.edu/ml/>

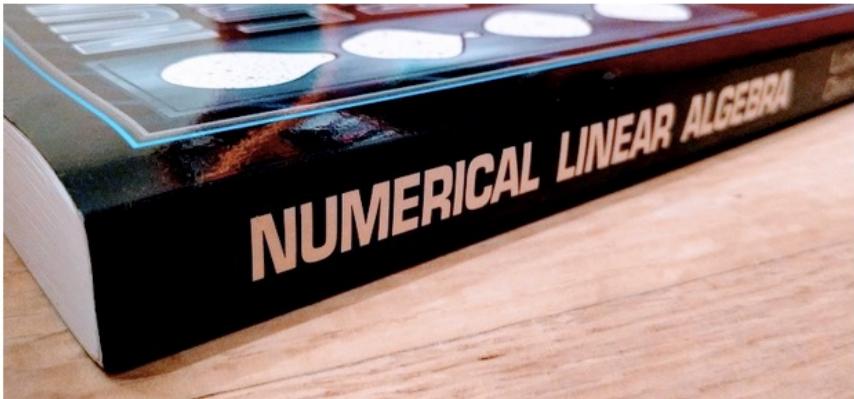


Singular Value Decomposition (SVD)

- Singular Value Decomposition (SVD)
 - What is it about ?
 - What can we say about existence ?
 - What can we say about uniqueness ?
 - How does it help us understand the multivariate Gaussian ?



Singular Value Decomposition (SVD)



Numerical Linear Algebra is the graduate textbook on numerical linear algebra I wrote with my advisor [Nick Trefethen](#) while earning a Masters at Cornell. The book began as a detailed set of notes that I took while attending Nick's course. The writing is intended to capture the spirit of his teaching: succinct and insightful. The hope is to reveal the elegance of this family of fundamental algorithms and dispel the myth that finite-precision arithmetic means imprecise thinking.

[L N Trefethen, D Bau. Numerical linear algebra. Vol. 50. Siam, 1997.](#)



David Bau

PhD Student at [MIT](#)

Verified email at mit.edu - [Homepage](#)

Computer Vision Machine Learning Software Engineering HCI

2015-PRESENT

Massachusetts Institute of Technology, Cambridge, MA

Ph.D. Candidate in Electrical Engineering and Computer Science

Thesis topic: The Representation of Visual Concepts in Deep Networks for Vision

Advisor: Antonio Torralba

Anticipated graduation: June 2021

1992-1994

Cornell University, Ithaca, NY

M.S. in Computer Science

Book coauthored: *Numerical Linear Algebra*

Advisor: Lloyd N. Trefethen

1988-1992

Harvard College, Cambridge, MA

A.B. in Mathematics



Photo by Sarah Bird

Professor L N Trefethen FRS

Professor of Numerical Analysis, University of Oxford

Fellow of Balliol College

Head of Oxford's Numerical Analysis Group

Singular Value Decomposition (SVD)

- Matrix factorization
- Let matrix A be size MxN
- When A is real valued, then SVD of $A = U S V^T$, where:
 - V is orthogonal of size NxN
 - When A is complex: V is unitary
 - U is orthogonal of size MxM
 - When A is complex: U is unitary
 - S is (rectangular) diagonal with size MxN
 - Values on diagonal = singular values
 - Singular values are non-negative real (even when A, U, V are complex-valued)
 - If the m-th columns of U and V are u_m and v_m , respectively, then:

When $M \leq N$, $A = \sum_{m=1}^M s_m u_m v_m^T$

Singular Value Decomposition (SVD)

- $A = U S V^T$
- An example, in pictures:

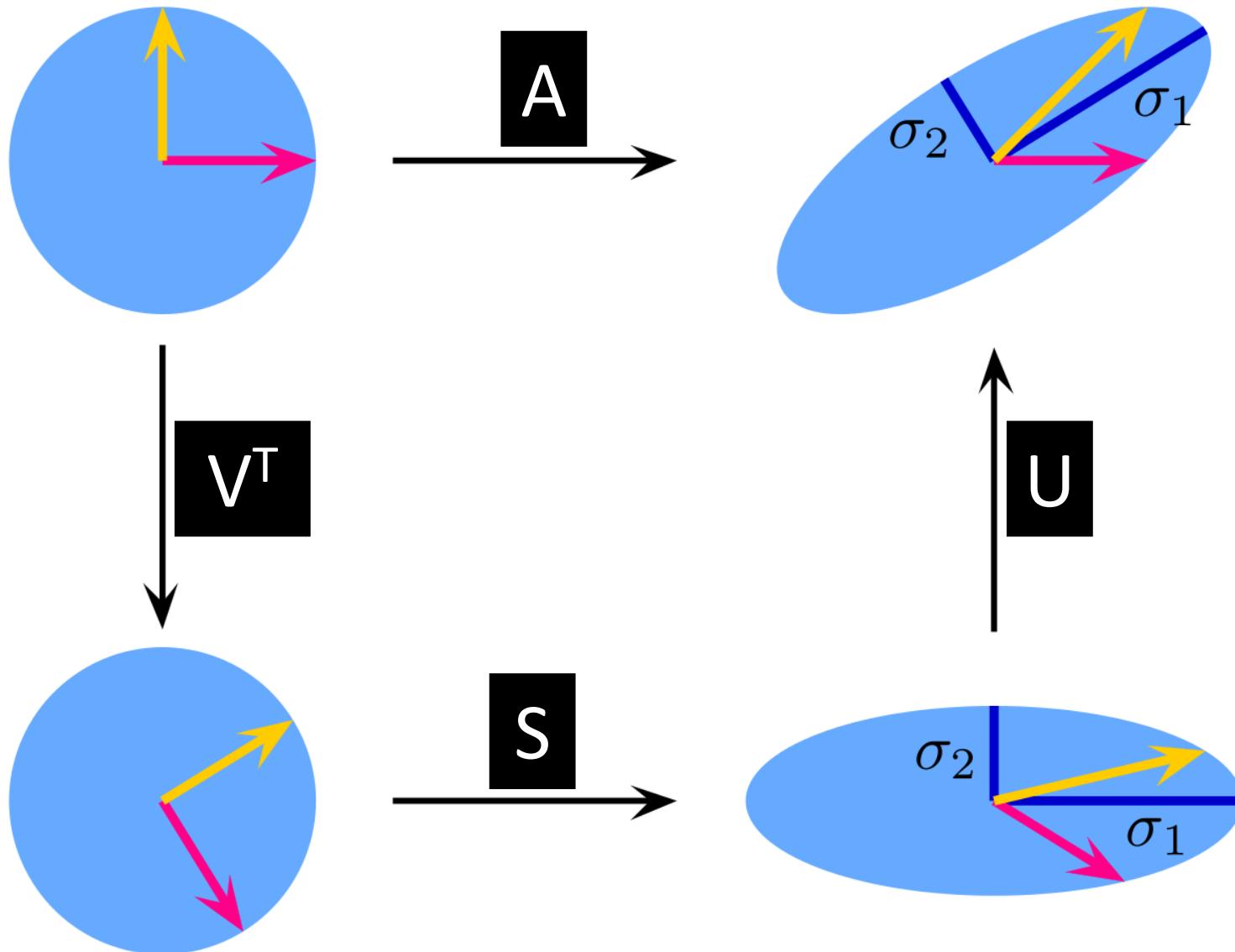
$$A = U S V^T$$

The diagram illustrates the Singular Value Decomposition (SVD) of a 4x4 matrix A. The matrices are represented as grids of colored squares.

- A:** A 4x4 gray grid.
- U:** A 4x4 matrix with three distinct vertical color bands: pink (leftmost), purple (middle), and light blue (rightmost).
- S:** A 4x4 diagonal matrix with colored entries:
 - Top-left entry: orange.
 - Diagonal entries: yellow.
 - Off-diagonal entries: gray circles.
- V^T :** A 4x4 matrix with four distinct horizontal color bands: teal (topmost), green, blue, and light green (bottommost).

Singular Value Decomposition (SVD)

- Geometric interpretation of the action of a matrix A on a vector
 - In this example,
 A is square



Singular Value Decomposition (SVD)



WIKIPEDIA
The Free Encyclopedia

Article [Talk](#)

Read [Edit](#)



Dragunov sniper rifle

From Wikipedia, the free encyclopedia

Not to be confused with [Degtyarev sniper rifle](#).

The **Dragunov sniper rifle** (formal Russian: Снайперская Винтовка систéмы Драгунóва образцá 1963 года, *Snáyperskaya Vintóvka sistém'y Dragunóva obraz'tsá 1963 goda* (**SVD-63**), officially "Sniper Rifle, System of Dragunov, Model of the Year 1963") (**GRAU** index **6V1** (ГРАУ Индекс 6В1)) is a semi-automatic designated marksman rifle chambered in **7.62×54mmR** and developed in the Soviet Union.



Singular Value Decomposition (SVD)

- Matrix norm
 - Induced by a vector norm

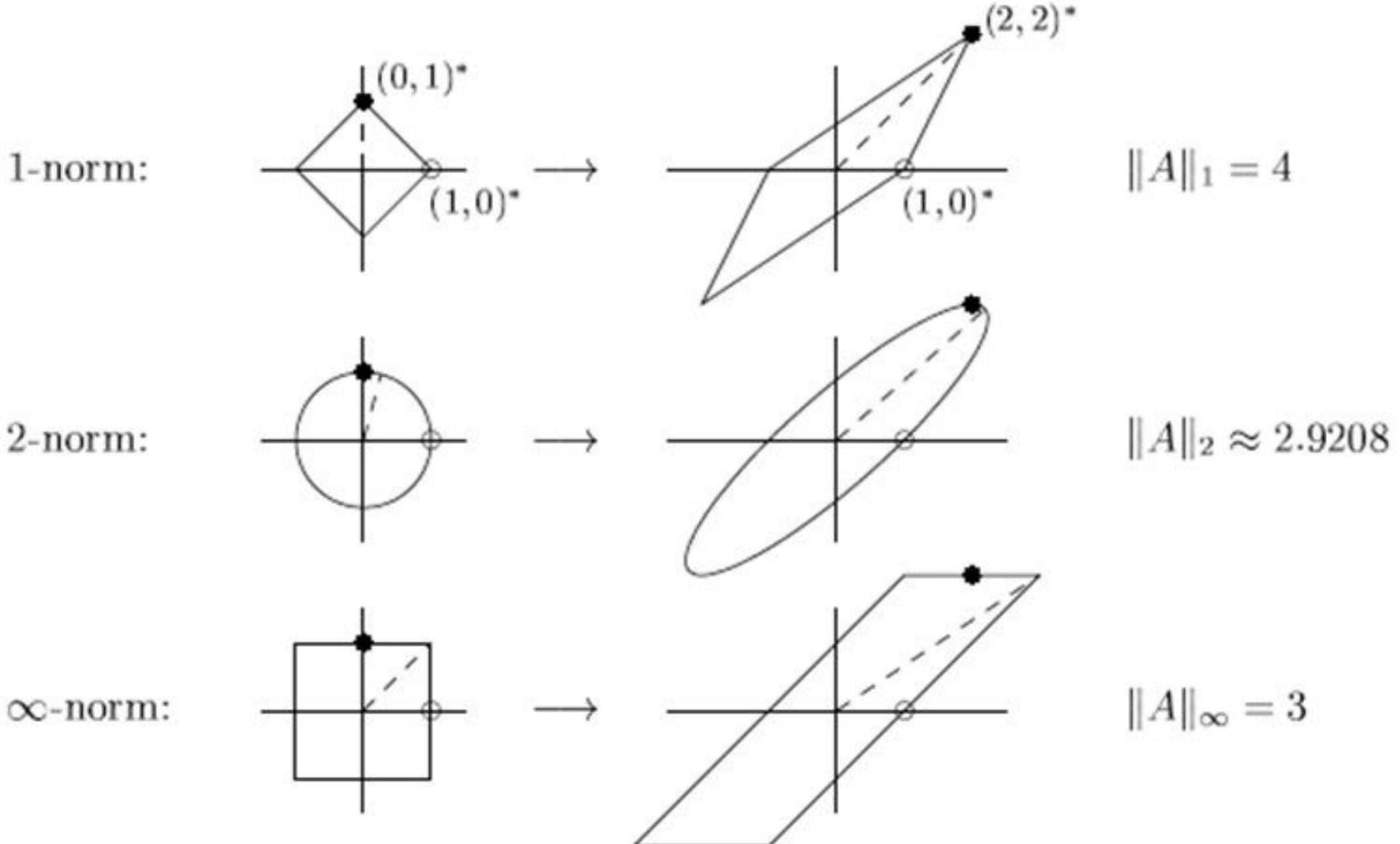
For a vector $x \in \mathbb{R}^N$, consider the vector 2-norm as $\|x\|_2$

For matrix A of size $M \times N$, the induced norm is defined as $\|A\|_2 := \max_{x \neq 0} (\|Ax\|_2 / \|x\|_2) \geq 0$

- Geometric interpretation related to 2-norm
 - Apply “linear operator” A to all unit-norm vectors x (starting at origin)
 - Let $y := Ax$, for all such x
 - Then, among all vectors y , pick the norm of the vector y' that has the largest norm

Singular Value Decomposition (SVD)

- Matrix norm
 - Induced by a vector norm

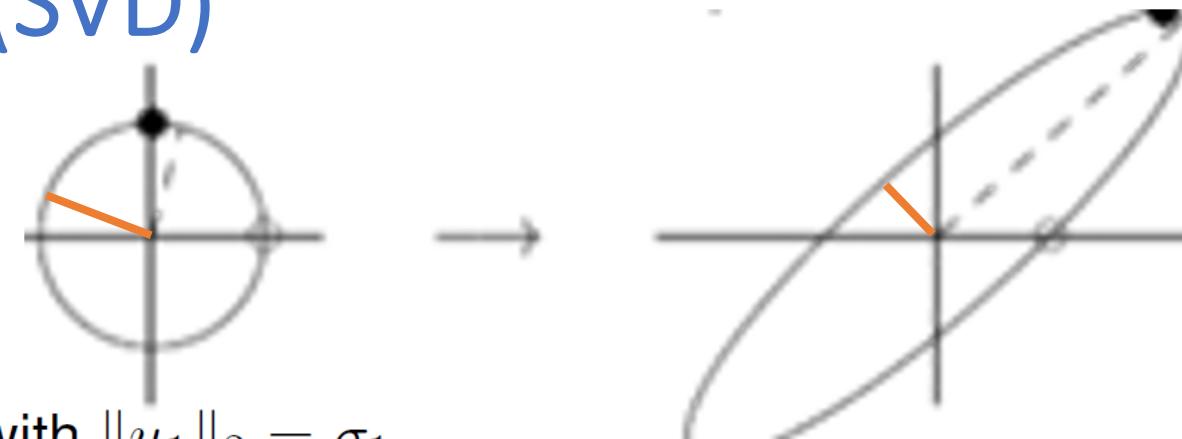


Singular Value Decomposition (SVD)

- **Existence**, for any real matrix A

Let $\sigma_1 := \|A\|_2 \geq 0$

Thus, $\exists v_1 \in \mathbb{R}^N$ with $\|v_1\|_2 = 1$ and $u_1 := Av_1$ with $\|u_1\|_2 = \sigma_1$



Consider orthogonal matrix U as a basis for \mathbb{R}^M , with columns u_j , and first column $u_1/\|u_1\|_2$

Consider orthogonal matrix V as a basis for \mathbb{R}^N , with columns v_j , and first column v_1

Then, $U^\top AV = S = \begin{bmatrix} \sigma_1 & \mathbf{w}^\top \\ \mathbf{0} & B \end{bmatrix}$ where sub-matrix B has size $(M - 1) \times (N - 1)$

We will now show that row-vector $\mathbf{w}^\top = \mathbf{0}^\top$

Singular Value Decomposition (SVD)

- Existence

Then, $U^\top A V = S = \begin{bmatrix} \sigma_1 & \mathbf{w}^\top \\ \mathbf{0} & B \end{bmatrix}$ where sub-matrix B has size $(M - 1) \times (N - 1)$

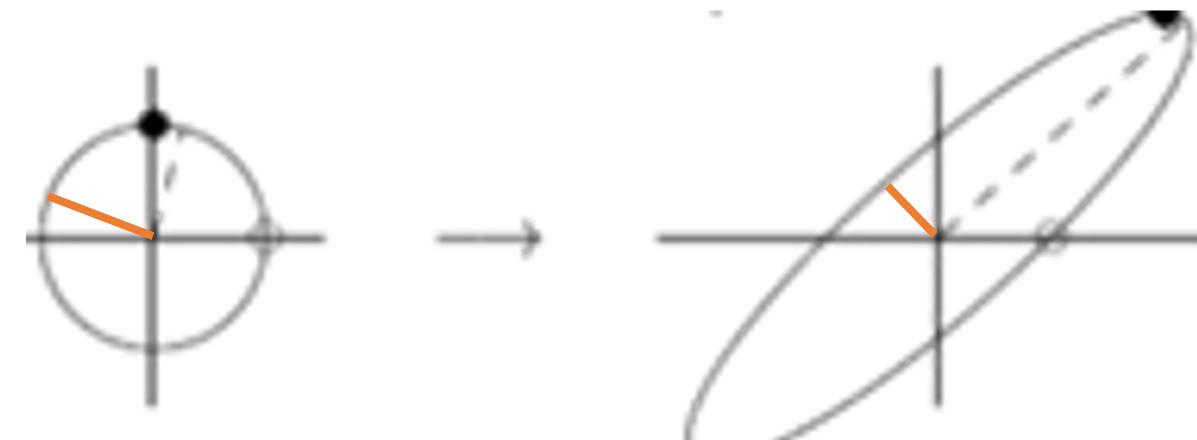
We will now show that row-vector $\mathbf{w}^\top = \mathbf{0}^\top$

Because U and V are orthogonal matrices, we get $\|S\|_2 = \|A\|_2 = \sigma_1$

$$\text{Now, } \left\| S \begin{bmatrix} \sigma_1 \\ \mathbf{w} \end{bmatrix} \right\|_2 \geq \sigma_1^2 + \mathbf{w}^\top \mathbf{w}$$

$$= \sqrt{\sigma_1^2 + \mathbf{w}^\top \mathbf{w}} \left\| \begin{bmatrix} \sigma_1 \\ \mathbf{w} \end{bmatrix} \right\|_2$$

which implies that $\|S\|_2 \geq \sqrt{\sigma_1^2 + \mathbf{w}^\top \mathbf{w}}$



Q.E.D.

Singular Value Decomposition (SVD)

- Existence

Then, $U^\top AV = S = \begin{bmatrix} \sigma_1 & \textcircled{\text{W}}^\top \\ \mathbf{0} & B \end{bmatrix}$ where sub-matrix B has size $(M - 1) \times (N - 1)$

Thus, $U^\top AV = S$ that has all zeros in the first row and first column, except at the top left position that has σ_1

Thus, A has a factorization of the form $A = USV^\top$, where U and V are orthogonal

Singular Value Decomposition (SVD)

- Existence

- How to analyze S further ? Induction on size of A, i.e., MxN

If $N = 1$ or $M = 1$, then:

S is a vector of size $M \times 1$ or $1 \times N$,

B doesn't exist (0×0 matrix), and

$$S = I_{M \times M} [\sigma_1, \mathbf{0}^\top]_{M \times 1}^\top I_{1 \times 1} \text{ or } S = I_{1 \times 1} [\sigma_1, \mathbf{0}^\top]_{1 \times N} I_{N \times N}$$

For our original matrix A , by the induction hypothesis,

the remaining submatrix B has a factorization of the form EDF^\top

Then,

$$A = USV^\top$$

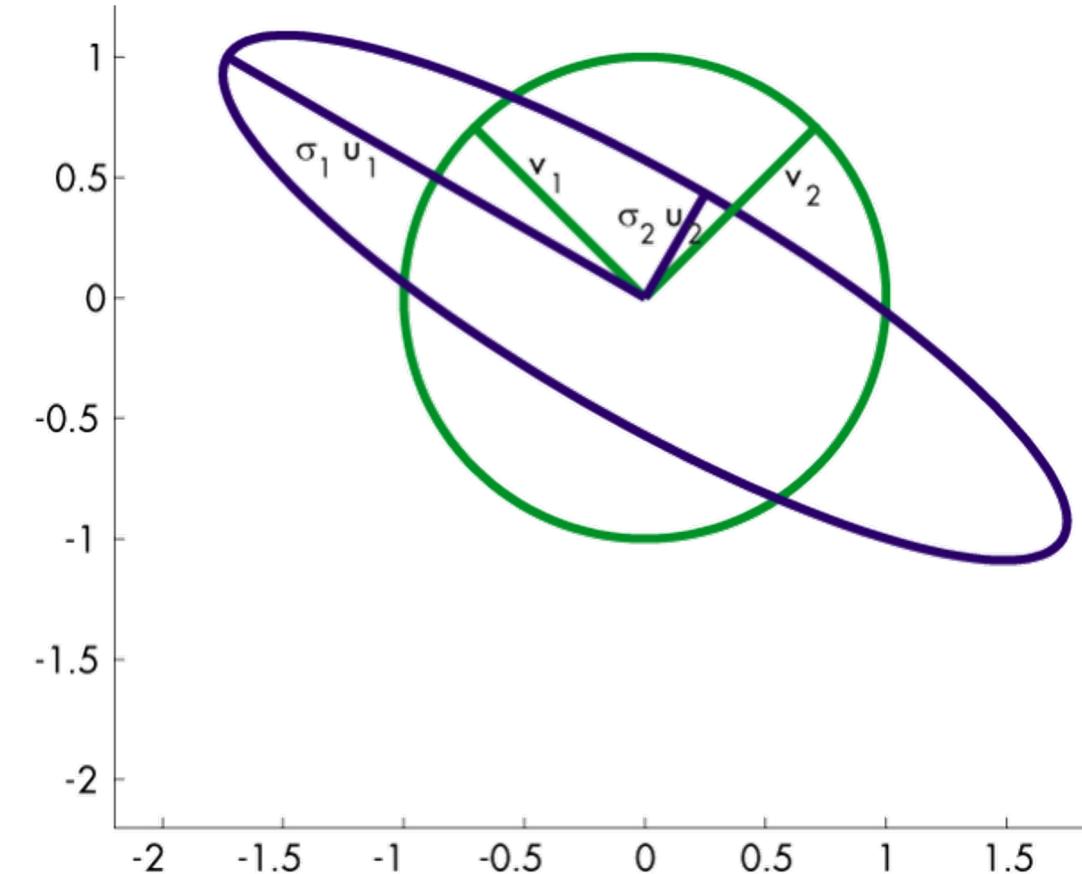
$$= U \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & E \end{bmatrix} \begin{bmatrix} \sigma_1 & \mathbf{0}^\top \\ \mathbf{0} & D \end{bmatrix} \begin{bmatrix} 1 & \mathbf{0}^\top \\ \mathbf{0} & F \end{bmatrix}^\top V^\top \quad (\text{where the larger matrices, containing sub-matrices } E \text{ and } F, \text{ are also orthogonal})$$

$$= U' \begin{bmatrix} \sigma_1 & \mathbf{0}^\top \\ \mathbf{0} & D \end{bmatrix} (V')^\top \quad (\text{with } D \text{ diagonal and } U', V' \text{ orthogonal})$$

$$S = \begin{bmatrix} \sigma_1 & \cancel{\mathbf{0}^\top} \\ \mathbf{0} & B \end{bmatrix}$$

Singular Value Decomposition (SVD)

- Properties of singular values, vectors
 - What does $A = U S V^T$ imply ?
 - Some insights via algebra and geometry
 - Let i-th column of V be v_i
 - Let j-th column of U be u_j
 - What is Av_i ? For example, take $i = 2$.
(assume S is at least of size 2x2)
 - $$\begin{aligned} Av_2 &= USV^T v_2 \\ &= US [0 \ 1 \ 0 \ \dots \ 0]^T \\ &= U [0 \ S_{22} \ 0 \ \dots \ 0]^T \\ &= S_{22} u_2 \end{aligned}$$
 - Thus, Av_1 is along u_1 , and, hence, orthogonal to all other columns of U
 - Also, Av_2 is along u_2 , and, hence, orthogonal to all other columns of U, ...
 - Also, if any vector v is orthogonal to v_1 , then Av is orthogonal to Av_1 or u_1



Singular Value Decomposition (SVD)

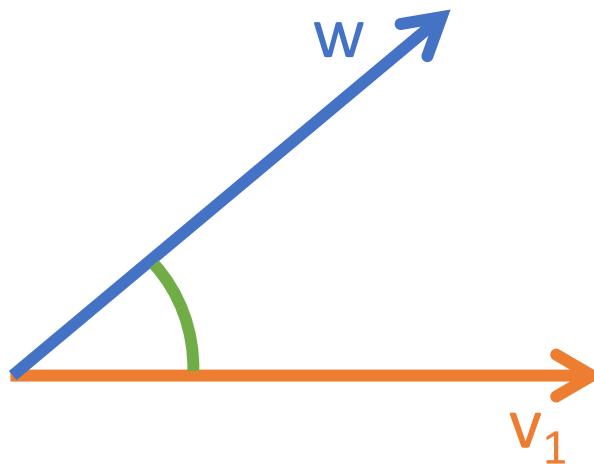
- **Uniqueness analysis** of singular values, singular vectors

The matrix 2-norm $\|A\|_2$ is unique (by definition). Let the norm be represented by $\sigma_1 = \|A\|_2$

Is the right-singular vector v_1 (of unit norm) unique (upto sign) ?

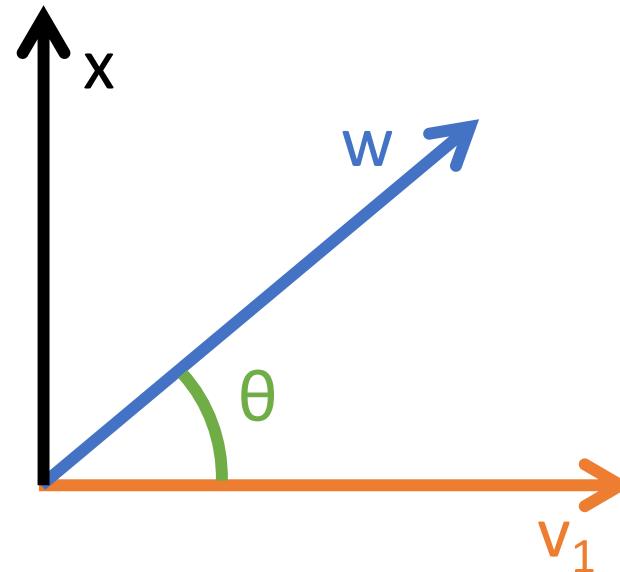
Suppose \exists a unit-norm vector w , linearly independent with v_1 , for which also $\|Aw\|_2 = \sigma_1$

What happens then ?



Singular Value Decomposition (SVD)

- Uniqueness analysis



Suppose \exists a unit-norm vector w , linearly independent with v_1 , for which also $\|Aw\|_2 = \sigma_1$

What happens then ?

Define a unit vector x that is in the same direction as the component of w orthogonal to v_1

We know that $\|A\|_2 = \sigma_1$. Thus, by definition of the matrix norm, $\|Ax\|_2 \leq \sigma_1$

But we can show that $\|Ax\|_2 = \sigma_1$

Singular Value Decomposition (SVD)

- Uniqueness analysis

Let angle between w and v_1 equal θ .

Then, (see the picture) we can rewrite

$$w = (\|w\|_2 \cos \theta)v_1 + (\|w\|_2 \sin \theta)x$$

$$= cv_1 + sx$$

$$\text{where } c^2 + s^2 = 1.$$

where $(Ax)^\top (Av_1) = 0$

because $A = USV^\top$ (existence), we get

$$x^\top (A^\top A)v_1$$

$$= x^\top (VS^\top SV^\top)v_1$$

$$= [0, y^\top] S^\top S [1, 0, \dots, 0]^\top \text{ (where we define } [0, y^\top] \equiv x^\top V)$$

$$= [0, z^\top] [S_{11}^2, 0, \dots, 0]^\top \text{ (where } z \text{ scales each element of } y \text{ by the corresponding diagonal element of } S^\top)$$

$$= 0$$

Then,

$$\sigma_1^2$$

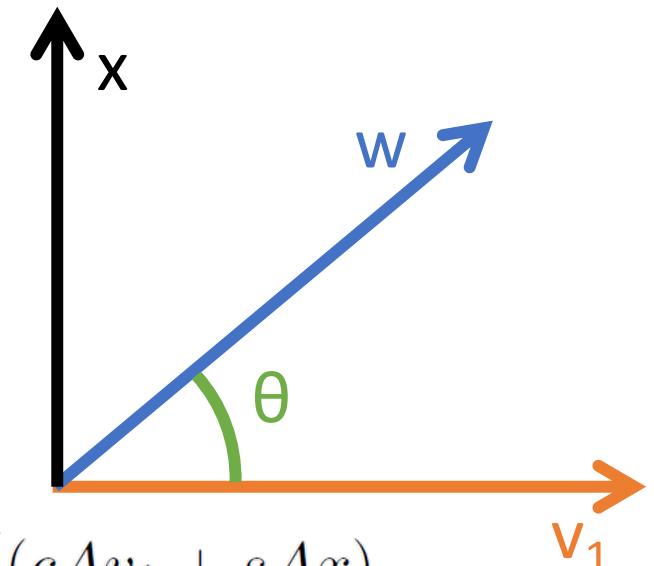
$$= \|Aw\|_2^2$$

$$= (Aw)^\top (Aw)$$

$$= (cAv_1 + sAx)^\top (cAv_1 + sAx)$$

$$= c^2\sigma_1^2 + s^2\|Ax\|_2^2 + 2cs(Ax)^\top (Av_1)$$

$$= c^2\sigma_1^2 + s^2\|Ax\|_2^2$$

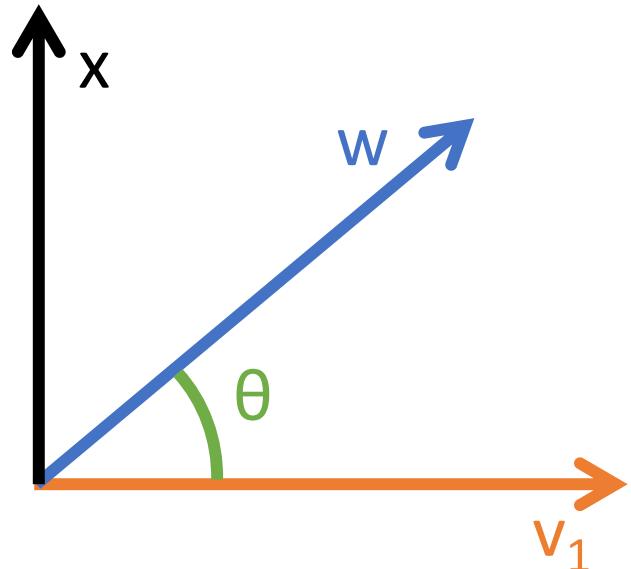


$$\|Ax\|_2 = \sigma_1$$

Singular Value Decomposition (SVD)

- Uniqueness analysis

Because x is constructed to be orthogonal to v_1 , if $\|Ax\|_2 = \sigma_1$, then x must be the second right-singular vector of A , also associated with the same singular value σ_1



We can see this as follows:

$$\begin{aligned}\sigma_1 &= \|Ax\|_2 \\ &= \|(USV^\top)x\|_2 \text{ (because } A = USV^\top \text{, where } U \text{ and } V \text{ are orthogonal)} \\ &= \|S[0, y^\top]^\top\|_2 \\ &= \|By\|_2 \text{ (because } S \text{ has that special structure for first row and column)}\end{aligned}$$

$$S = \begin{bmatrix} \sigma_1 & \del{w}^\top \\ 0 & B \end{bmatrix}$$

Because $\|B\|_2 \leq \|A\|_2$ (can be shown by contradiction), y must be the principal singular vector of B

Singular Value Decomposition (SVD)

- Uniqueness analysis
 - Why is $\text{norm}(B) \leq \text{norm}(A)$?
 - We know that $A = USV^T$, where U and V are orthogonal, and S is as shown above
 - Let $\beta := \text{norm}(B)$
 - By definition of $\text{norm}(B)$,
there exists a unit-norm column-vector y such that $\text{norm}(By) = \beta$
 - Use that y to construct a *longer (but still) unit-norm* column vector $x := V [0, y^T]^T$
 - $\text{norm}(Ax)$
 $= \text{norm} (USV^T V [0, y^T]^T)$
 $= \text{norm} (S [0, y^T]^T)$
 $= \text{norm} ([0, (By)^T]^T)$
 $= \text{norm}(By)$
 $= \beta$
 - Thus, there exists a vector x such that $\text{norm}(Ax) = \beta$,
which implies that $\text{norm}(A)$ cannot be less than β , i.e., $\text{norm}(A) \geq \beta = \text{norm}(B)$

$$U^T A V = S = \begin{bmatrix} \sigma_1 & \cancel{\mathbf{w}}^T \\ \mathbf{0} & B \end{bmatrix}$$

Singular Value Decomposition (SVD)

- Uniqueness analysis

This concludes that, for the unique matrix-norm $\sigma_1 = \|A\|_2$,

- (i) if principal singular vector v_1 isn't unique (upto sign, i.e., $\pm v_1$), then singular value σ_1 is repeated (isn't “simple”),
or, equivalently,
- (ii) if singular value σ_1 is simple (no multiplicity), then principal singular vector v_1 is unique (upto sign, i.e., $\pm v_1$).

Once σ_1, v_1, u_1 are determined,

the remainder of the SVD is determined by the action of A on the space orthogonal to $\pm v_1$ (that space is uniquely defined)

- Properties of other singular values and singular vectors follows by induction
- Thus, if all singular values are distinct, then all singular vectors are unique (upto sign)

Singular Value Decomposition (SVD)

- How does SVD help us in understanding the **multivariate Gaussian** ?
 - Consider $X := AW$, where:
 - Components of W are independent standard-normal. A is of size $M \times N$, where $M < N$.
 - We use $A := USV^T$, where:
 - S is $M \times N$ (rectangular) diagonal. U is $M \times M$ orthogonal. V is $N \times N$ orthogonal.
 - AW
 - = $USV^T W$
 - = $U S W'$ (where components of W' are also independent standard-normal)
 - = $U S' W''$ (where S' is square with columns as the first M columns of S ,
 W'' is first M components of W')
 - = $A' W''$ (where $A' = US'$ is $M \times M$, and W'' is $M \times 1$)
 - $\text{Covariance}(X) = C = AA^T = U SS^T U^T = A'A'^T$, where:
 - SS^T is square diagonal of size $M \times M$
 - For matrix C to be SPD, the rank of S needs to be M (M non-zero singular values)