CS 215
Data Analysis and Interpretation

Multivariate Statistics: Multivariate Gaussian

Suyash P. Awate

Multivariate Gaussian – Definition

- Consider a vector random variable $X := [X_1; X_2; \ldots; X_D]$
  - 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 $W_1, \cdots, W_N$ (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; …; 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 $W_1, \cdots, W_N$ (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 $W_1, \cdots, W_D$ with $A := I_{D \times D}$ and $\mu := 0$, i.e. $X = W$

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

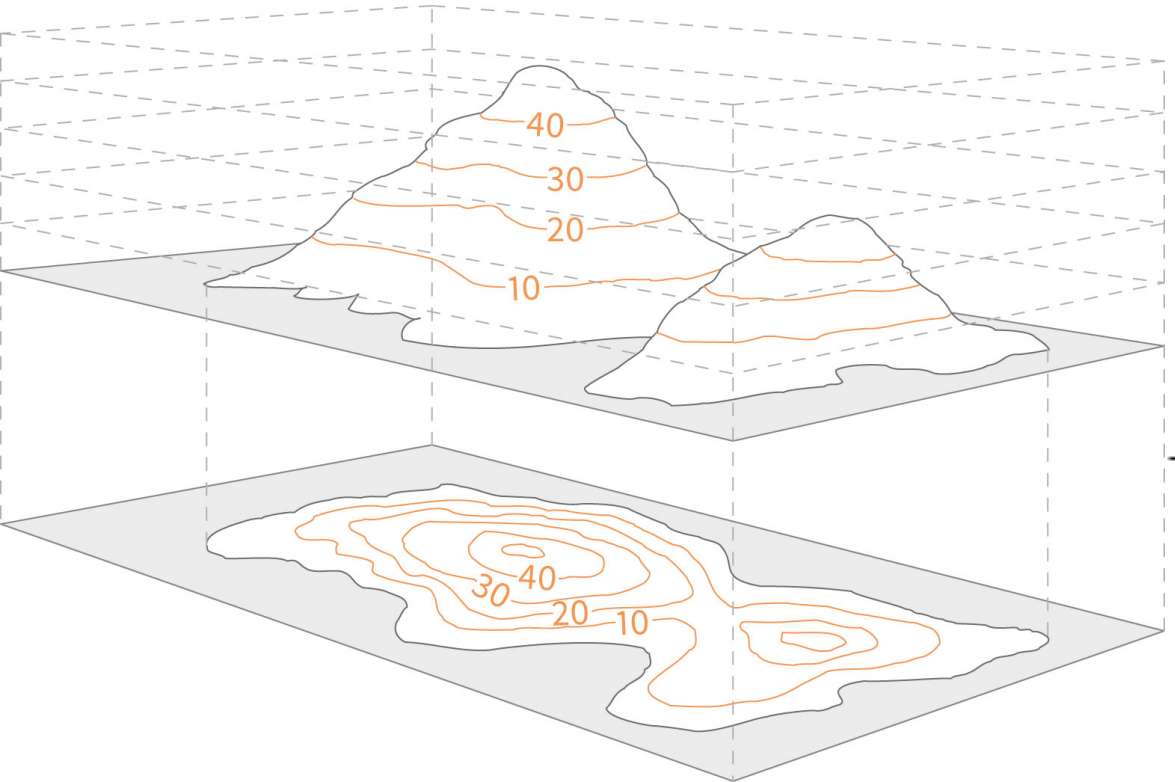
Multivariate Gaussian – Identity A

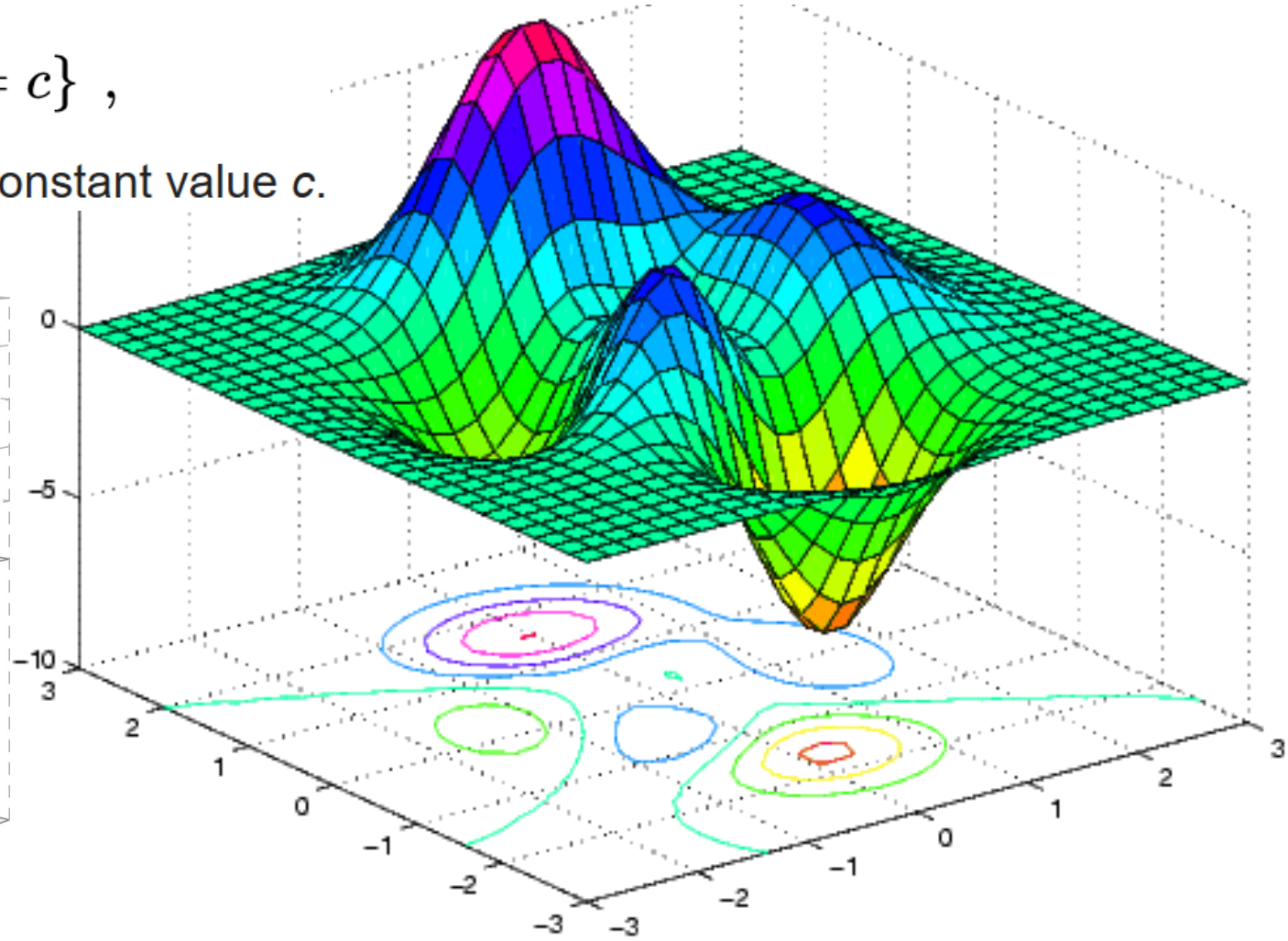
- What are the level sets of the PDF ?

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

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

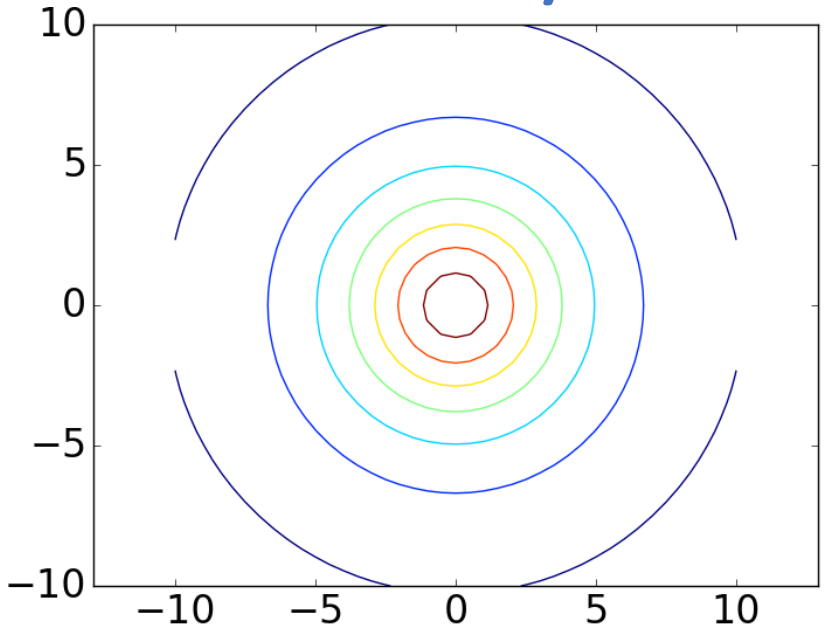
that is, a set where the function takes on a given constant value *c*.

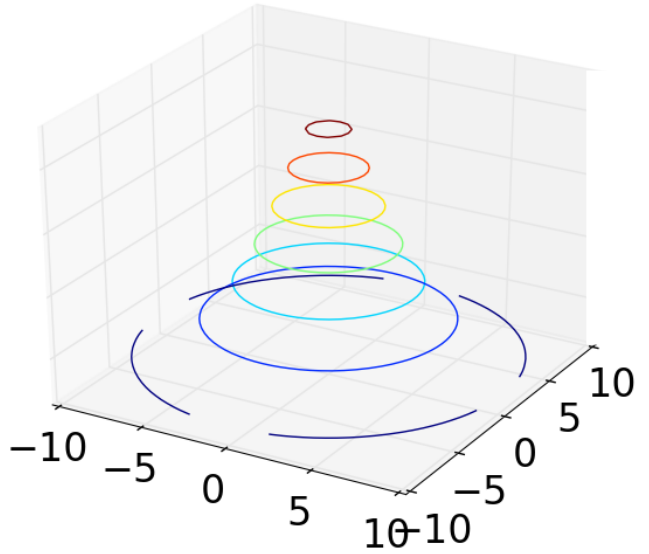




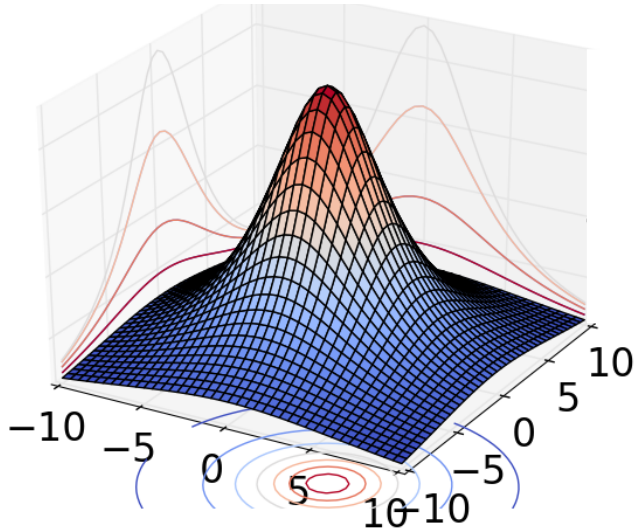
Multivariate Gaussian – Identity A

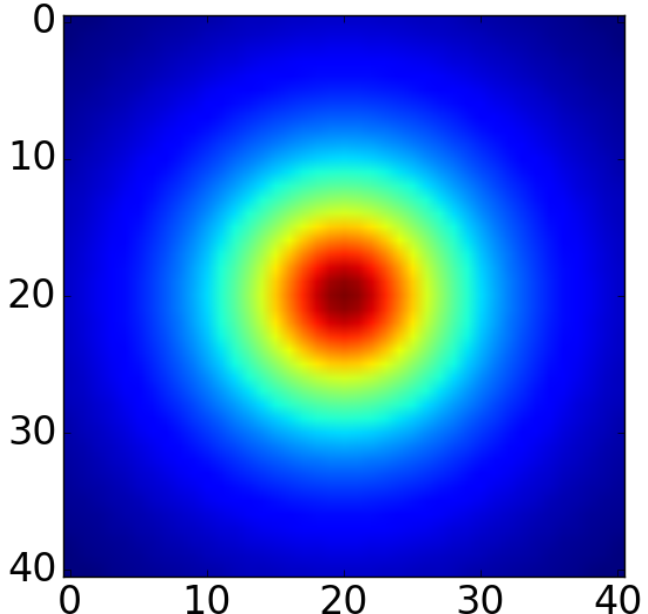
- Isotropic / spherical multivariate Gaussian
  - Level sets





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

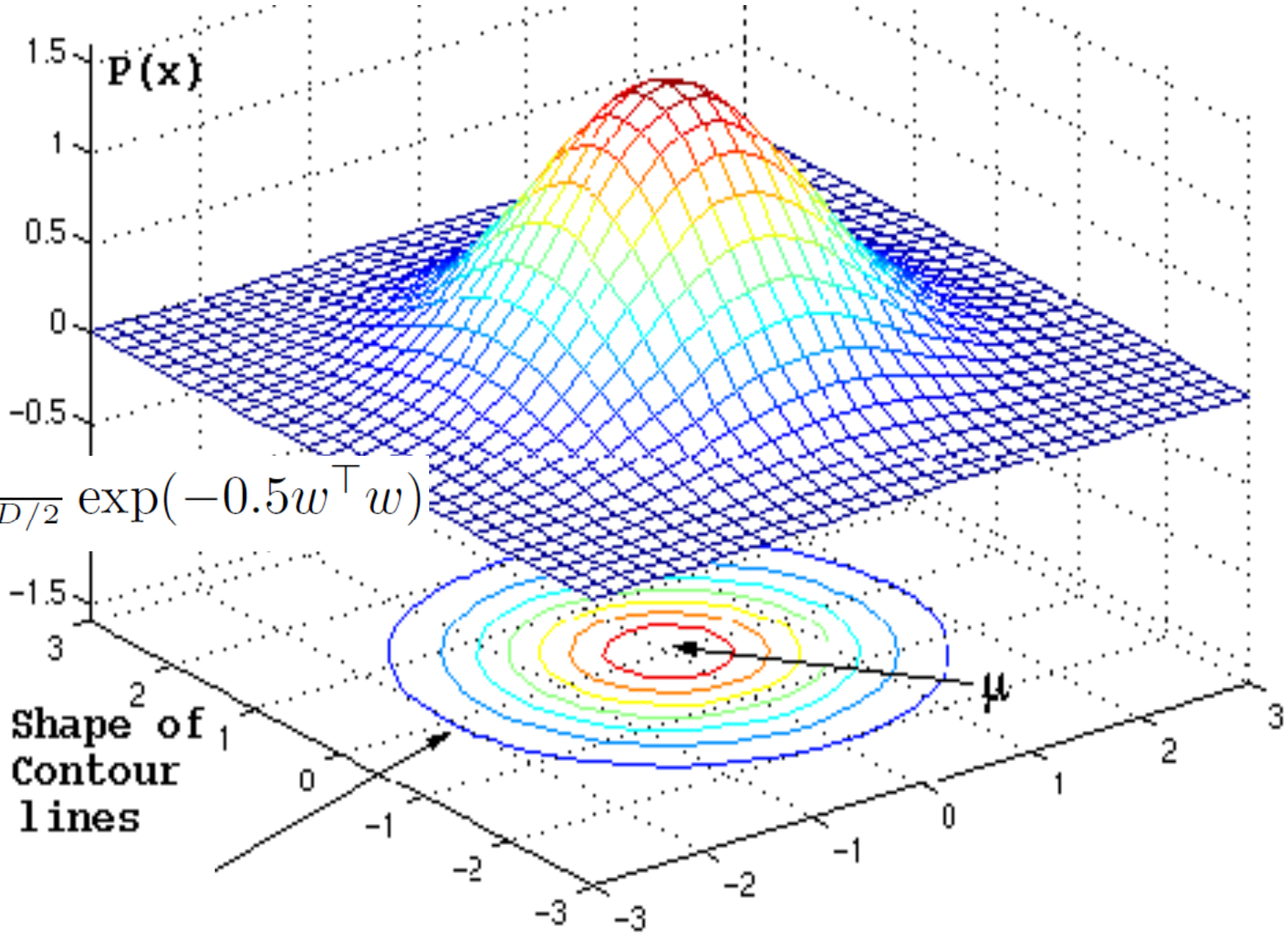




Multivariate Gaussian – Identity A

- Isotropic / spherical multivariate Gaussian
  - Level sets

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

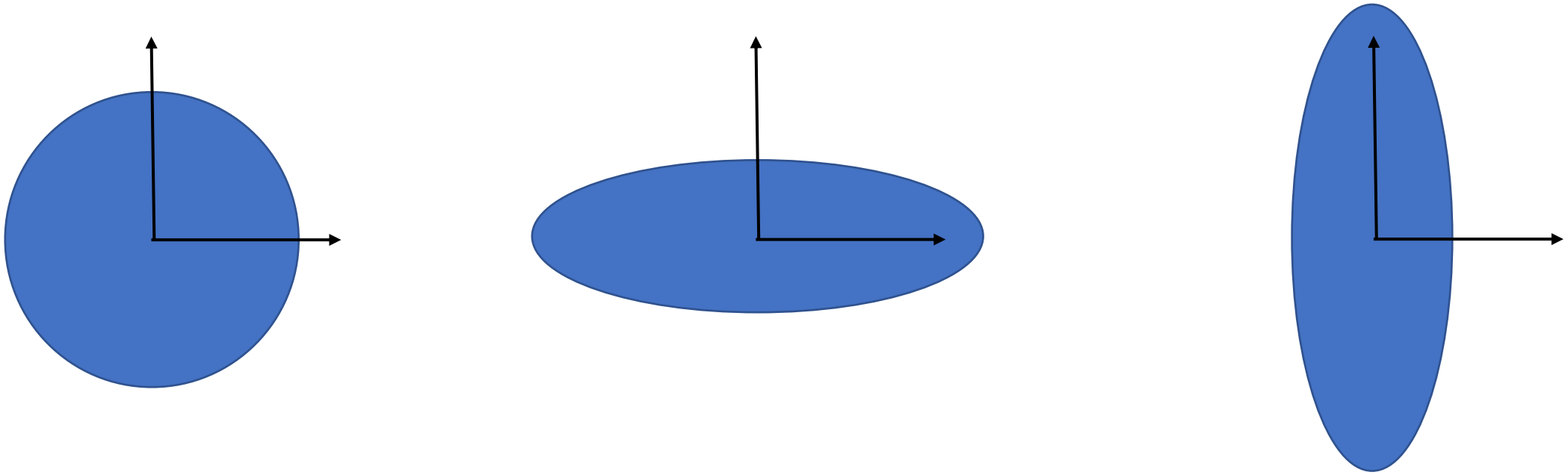


Multivariate Gaussian – Diagonal A

- $X = A\,W + \mu$
- What is PDF q(X) for **non-singular** square **diagonal** matrix A, 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 $\sigma_D = |A_{DD}|$
  - $P(X) = P(X_1, X_2, \ldots, X_D) = G(X_1; \mu_1, \sigma_1^2)\; G(X_2; \mu_2, \sigma_2^2)\; \ldots\; G(X_D; \mu_D, \sigma_D^2)$
  - Any level set of PDF q(X) is a hyper-ellipsoid with:
    - Mean at μ
    - Axes aligned with cardinal axes

Multivariate Gaussian – Diagonal A

- $X = A\,W + \mu$
- What is PDF q(X) for **non-singular** square **diagonal** matrix A, some μ ?
  - $P(X) = P(X_1, X_2, \ldots, X_D) = G(X_1; \mu_1, \sigma_1^2)\, G(X_2; \mu_2, \sigma_2^2) \ldots G(X_D; \mu_D, \sigma_D^2)$
  - Example 1-3 (left to right):
    both means (μ_1, μ_2) are zero,
    both variances are (σ_1^2, σ_2^2): (4,4), (9,1),(1,9)

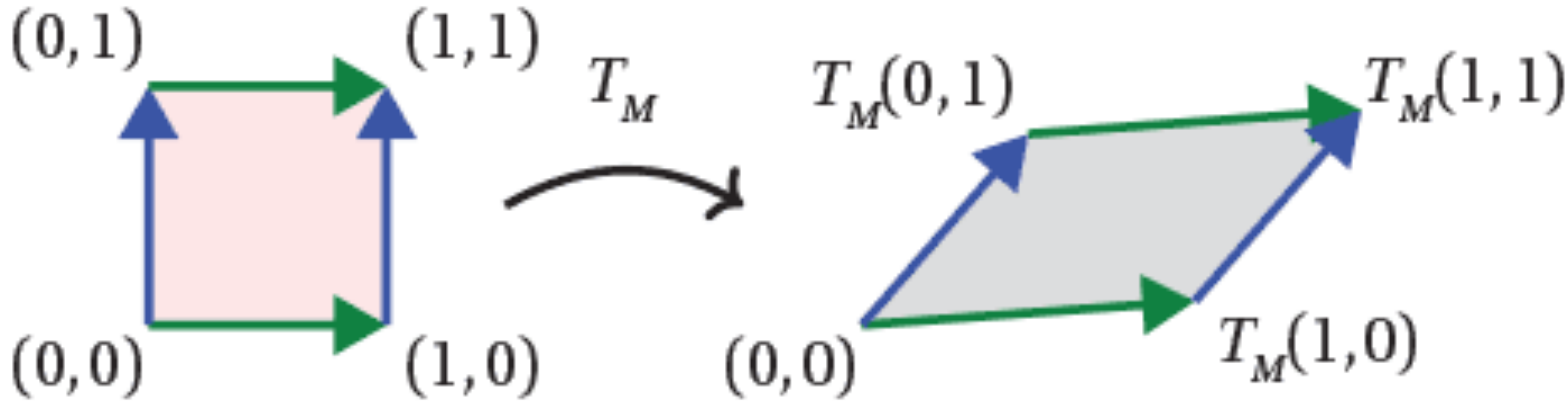


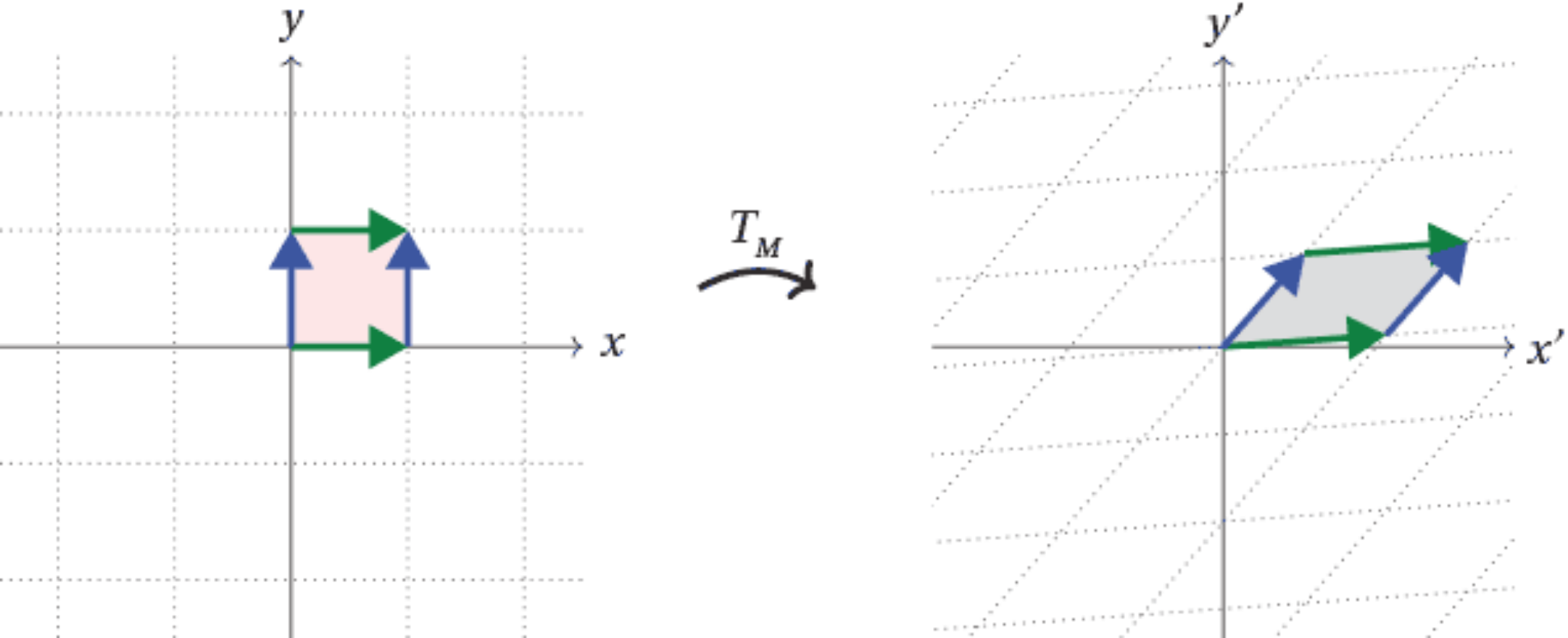
Multivariate Gaussian – Non-Singular A

- $X = A\,W + \mu$
- 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 $g^{-1}(.)$

Multivariate Gaussian – Non-Singular A

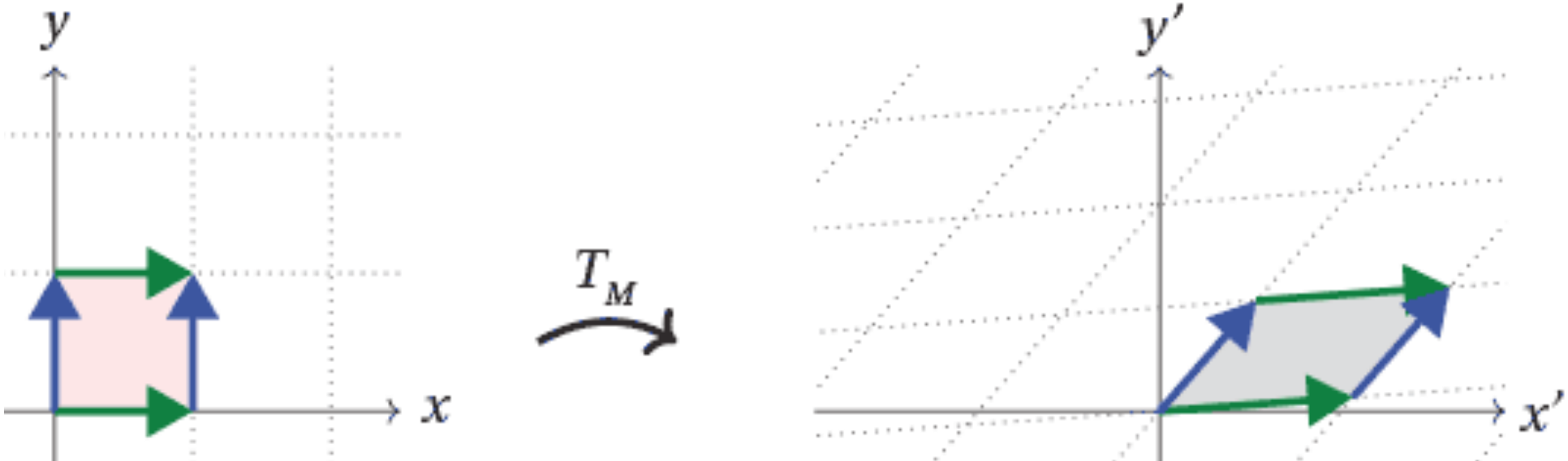
- Linear transformation
  $W := A^{-1} X$





Multivariate Gaussian – Non-Singular A

- Linear transformation $W := A^{-1} X$
  - Transformation A^{-1} maps
    an infinitesimal hyper-cube (dX) δ x δ x … x δ (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

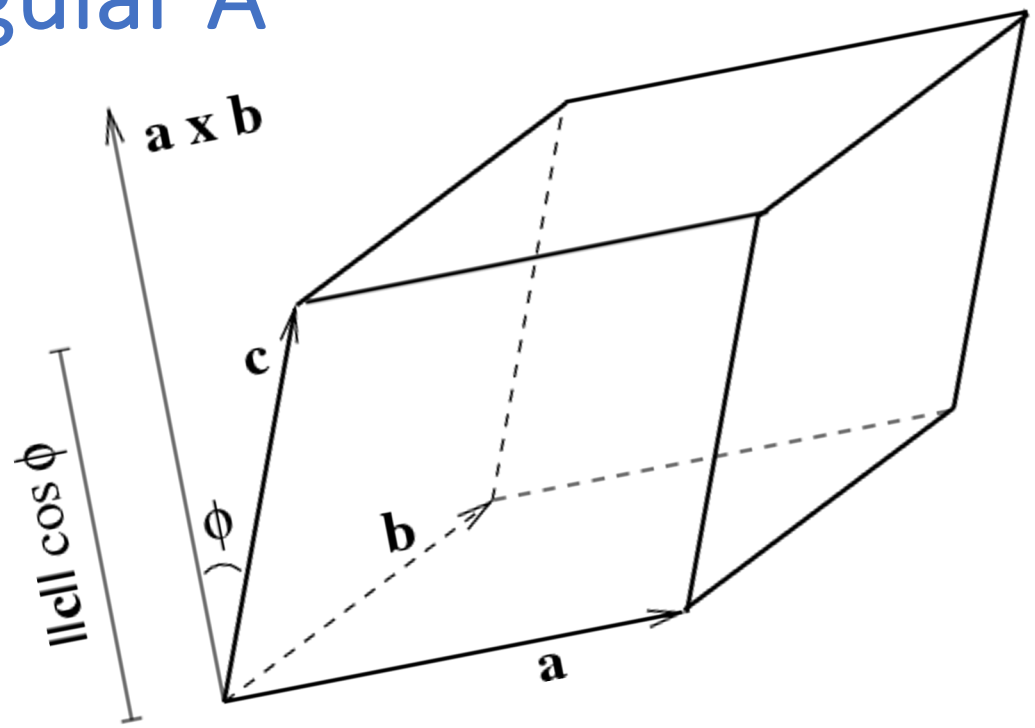
- Volume of a parallelepiped (in 3D)
  - Scalar triple product

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

$$= \mathbf{b} \cdot (\mathbf{c} \times \mathbf{a}) = \mathbf{c} \cdot (\mathbf{a} \times \mathbf{b})$$

$$= -\mathbf{a} \cdot (\mathbf{c} \times \mathbf{b}) = -\mathbf{c} \cdot (\mathbf{b} \times \mathbf{a}) = -\mathbf{b} \cdot (\mathbf{a} \times \mathbf{c})$$

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



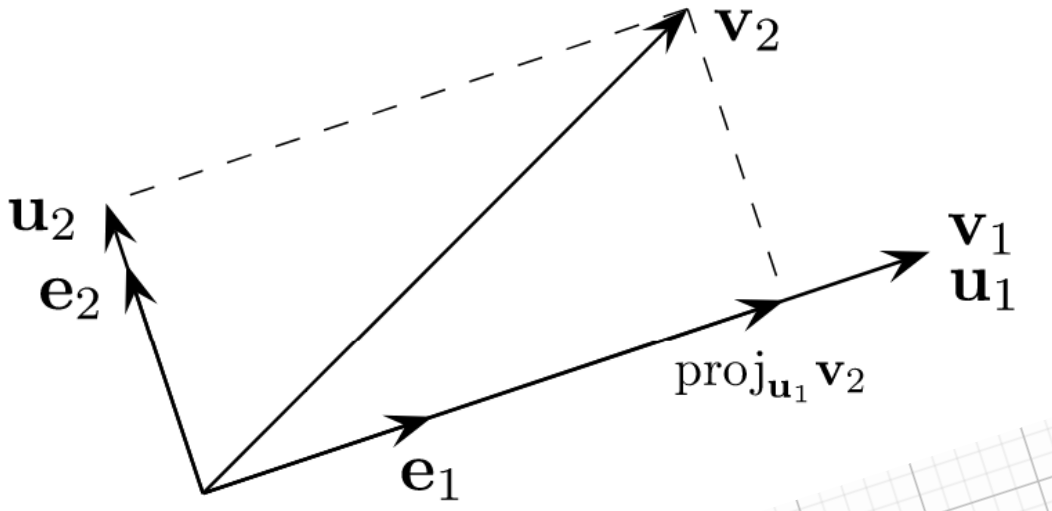
$$\text{Volume} = \text{area of base} \cdot \text{height}$$
$$= \|\mathbf{a} \times \mathbf{b}\| \, \|\mathbf{c}\| \, |\cos\phi| = |(\mathbf{a} \times \mathbf{b}) \cdot \mathbf{c}|$$

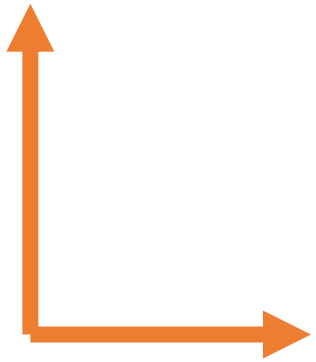
Multivariate Gaussian – Non-Singular A

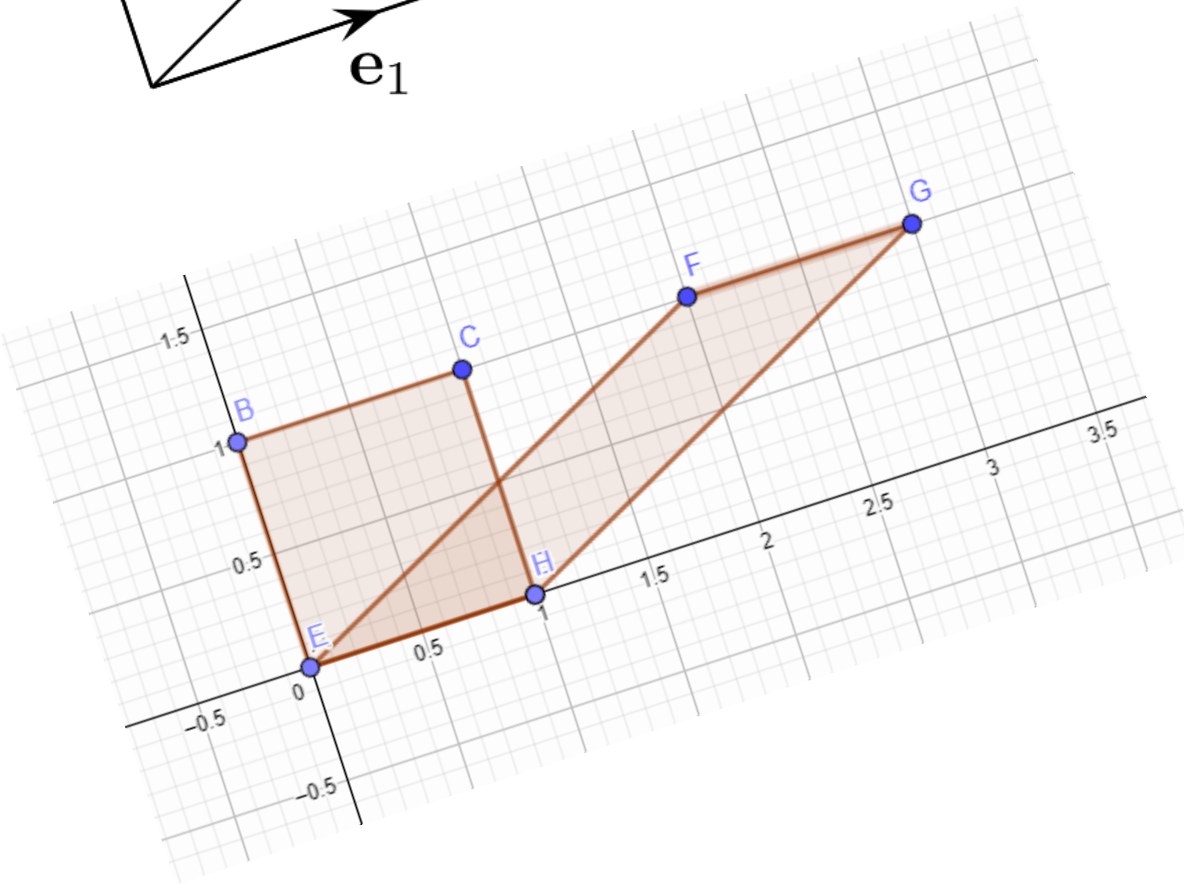
- Why is volume of hyper-parallelepiped given by determinant of matrix with columns as sides of hyper-parallelepiped ?
  - The following is an argument (not a proof; a separate inductive proof exists):
  - 2 important properties from linear algebra:
    Adding multiples of one column/side to another:
    - 1) doesn't change determinant, because determinant function is multi-linear
    - 2) doesn't change volume, because it causes a skew translation of hyper-parallelepiped
  - Using Gram-Schmidt orthogonalization,
    transform matrix A^{-1} to a matrix, say, A^{-1}_{ortho} with orthogonal columns
    (NOT orthonormal columns; that would have determinant 1)
    - This doesn't change determinant or volume

Multivariate Gaussian – Non-Singular A

- Gram–Schmidt orthogonalization
  - $\{v_1, v_2\}$ to $\{u_1, u_2\}$





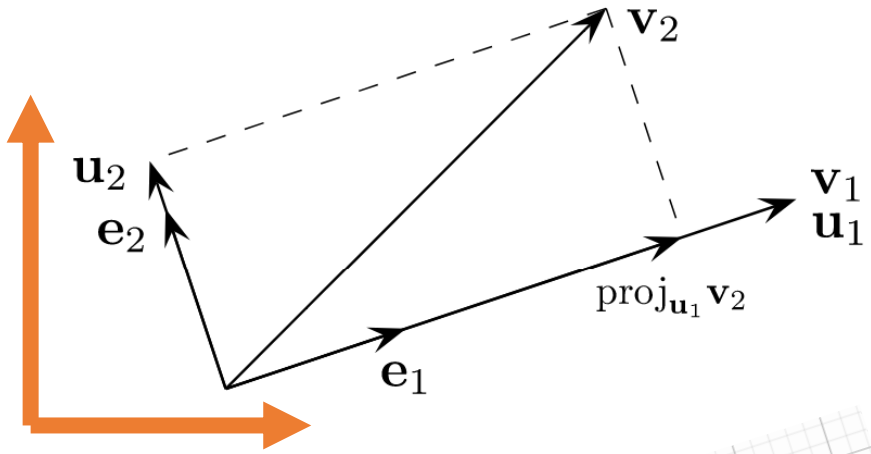


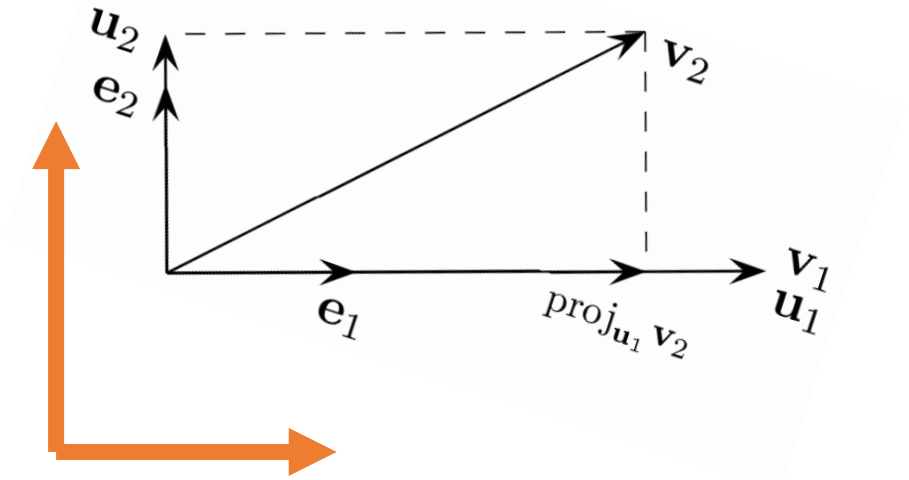
Multivariate Gaussian – Non-Singular A

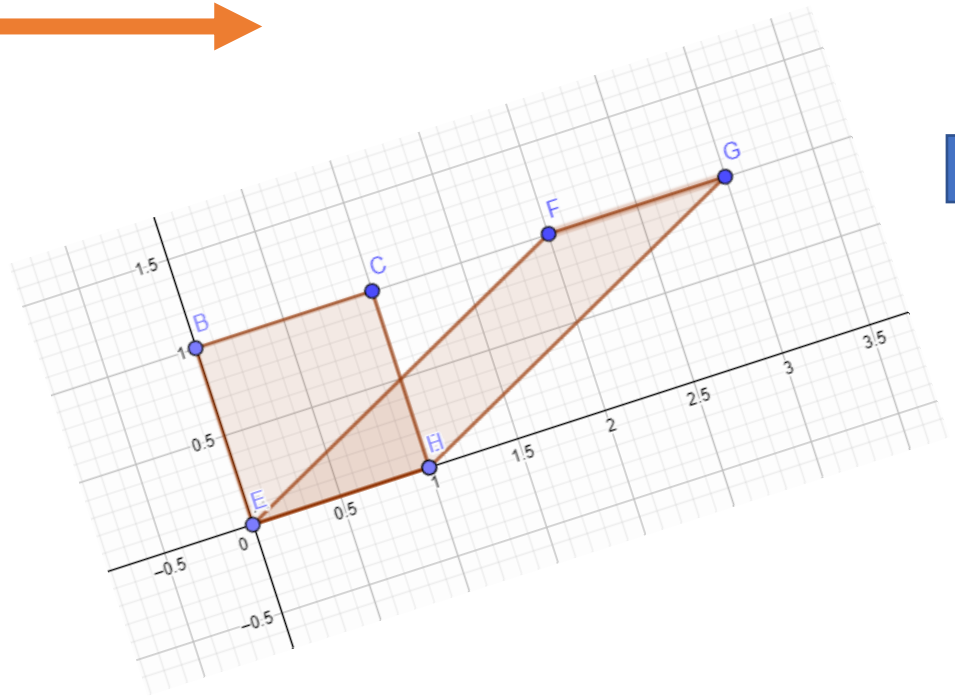
- Why is volume of hyper-parallelepiped given by determinant of matrix with columns as sides of hyper-parallelepiped ?
  - The following is an argument (not a proof; a separate inductive proof exists):
  - 2 important properties from linear algebra:
    Adding multiples of one column/side to another:
    - 1) doesn't change determinant, because determinant function is multi-linear
    - 2) doesn't change volume, because it causes a skew translation of hyper-parallelepiped
  - Using Gram-Schmidt orthogonalization,
    transform matrix A^{-1} to a matrix, say, A^{-1}_{ortho} with orthogonal columns
    (NOT orthonormal columns; that would have determinant 1)
  - Rotate A^{-1}_{ortho} to make it to diagonal form (align columns to cardinal axes)
    - This doesn't change determinant or volume

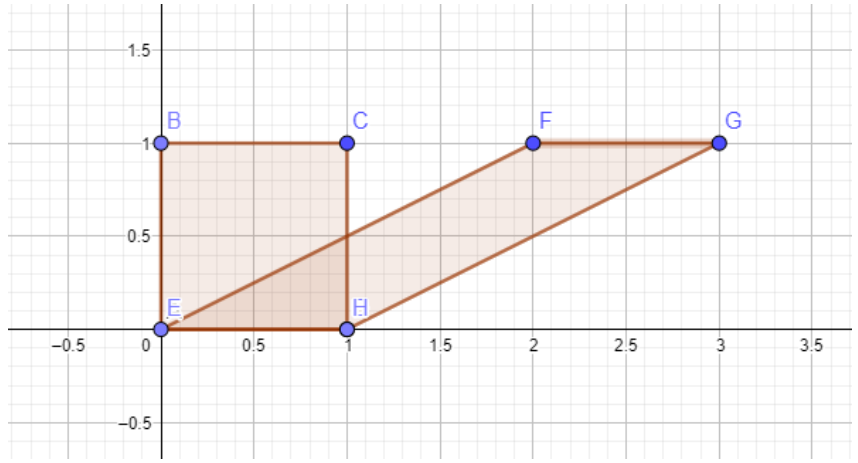
Multivariate Gaussian – Non-Singular A

- Rotation / alignment 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 ?
  - An intuitive argument (not a proof; a separate inductive proof exists):
  - Adding multiples of one column/side to another:
    - 1) doesn't change determinant, because determinant function is multi-linear
    - 2) doesn't change volume, because it causes a skew translation of hyper-parallelepiped
  - Using Gram-Schmidt orthogonalization, transform matrix A^{-1} to a matrix, say, A^{-1}_{ortho} with orthogonal columns (NOT orthonormal columns; that would have determinant 1)
  - Rotate A^{-1}_{ortho} to make it to diagonal form (align columns to cardinal axes)
  - For this diagonal matrix (aligned hyper-rectangle), determinant (= product of diagonal entries) = volume of a hyper-rectangle (= product of side lengths)
  - 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 $g^{-1}(.)$
    - We want the magnitude determinant of Jacobian of $g^{-1}(.)$

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

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 + μ 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 - μ := $g^{-1}(Y)$

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 + μ, where A is invertible
  - Thus,
    Z
    = B (AW + μ) + c
    = (BA)W + (Bμ + c), where matrix BA is invertible

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 $E_{P(X)}[X_i] = E_{P(Xi)}[X_i]$

Multivariate Statistics – Covariance

- Covariance matrix for a general random (column) vector Y:
  $C := E_{P(Y)} [(Y - E[Y]) (Y - E[Y])^T]$
- So,
  C_{ij}
  $= E_{P(Y)} [(Y_i – E[Y_i]) (Y_j – E[Y_j])]$
  $= E_{P(Yi,Yj)} [(Y_i – E[Y_i]) (Y_j – E[Y_j])]$
  $= Cov (Y_i, Y_j)$

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 is symmetric

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

(3) C is positive semi-definite (PSD)

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

Multivariate Gaussian – Mean and Covariance

Multivariate Gaussian – Mean

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

Multivariate Gaussian – Covariance

- The **covariance** matrix of X := AW + μ is AA^T

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

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

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

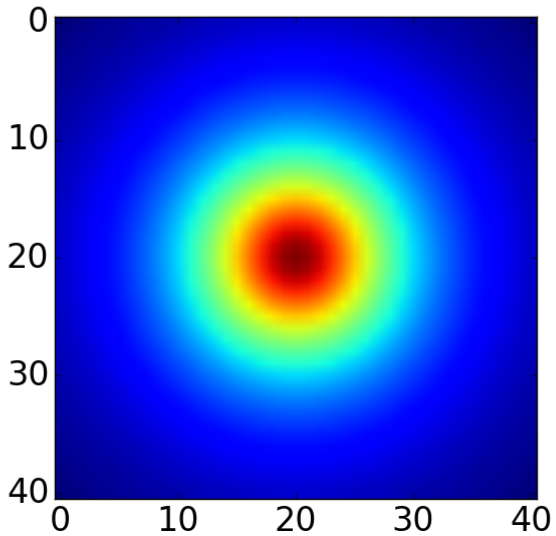
- Diagonal matrix
- Orthogonal matrix
  - Definition: Real square matrix Q whose columns and rows are **orthogonal unit** vectors (i.e., orthonormal vectors) $Q\,Q^T = Q^T\,Q$ = Identity matrix
  - Determinant det(Q) is either +1 or -1
  - "orthogonal" is an over-used term
- Rotation matrix
  - When 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 over-used (sometimes includes improper rotations)
- Reflection matrix
  - An orthogonal matrix that is also symmetric

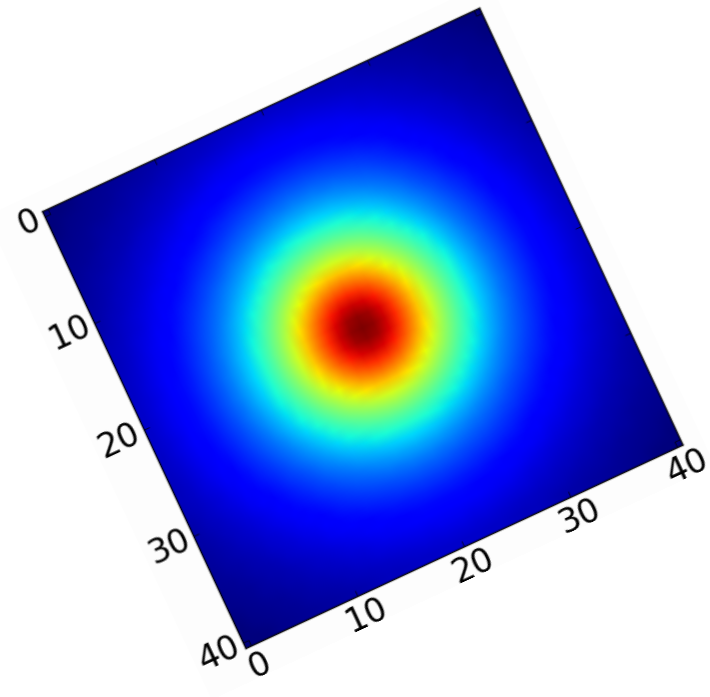
Multivariate Gaussian – Special Cases

- Property (Rotation and/or Reflection):
  If μ = 0; and A = R where R is orthogonal;
  then Y := RW has PDF:

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

- Proof:
  - Transformation of random vectors
    - |det(R)| = 1
    - Inverse transformation is
      W = transpose(R) Y





So, $Q(y) = (1/(2\pi)^{D/2}) \exp(-0.5(R^\top y)^\top R^\top y) = (1/(2\pi)^{D/2}) \exp(-0.5 y^\top y)$
Thus, $Y := RW$ is also a zero-mean isotropic multivariate Gaussian, just like W

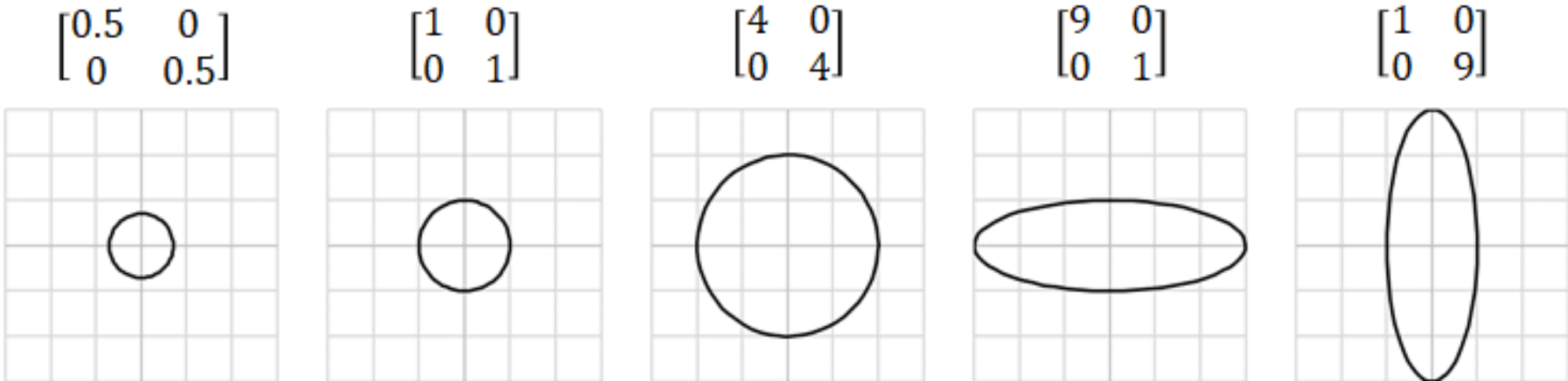
Multivariate Gaussian – Special Cases

- Property (Scaling):
  If μ = 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 $W = S^{-1}Y$

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.



Multivariate Gaussian – Special Cases

- Property (first Scaling, and then Rotation and/or Reflection):
  If μ = 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 $W = (RS)^{-1}Y = S^{-1}R^\top y$

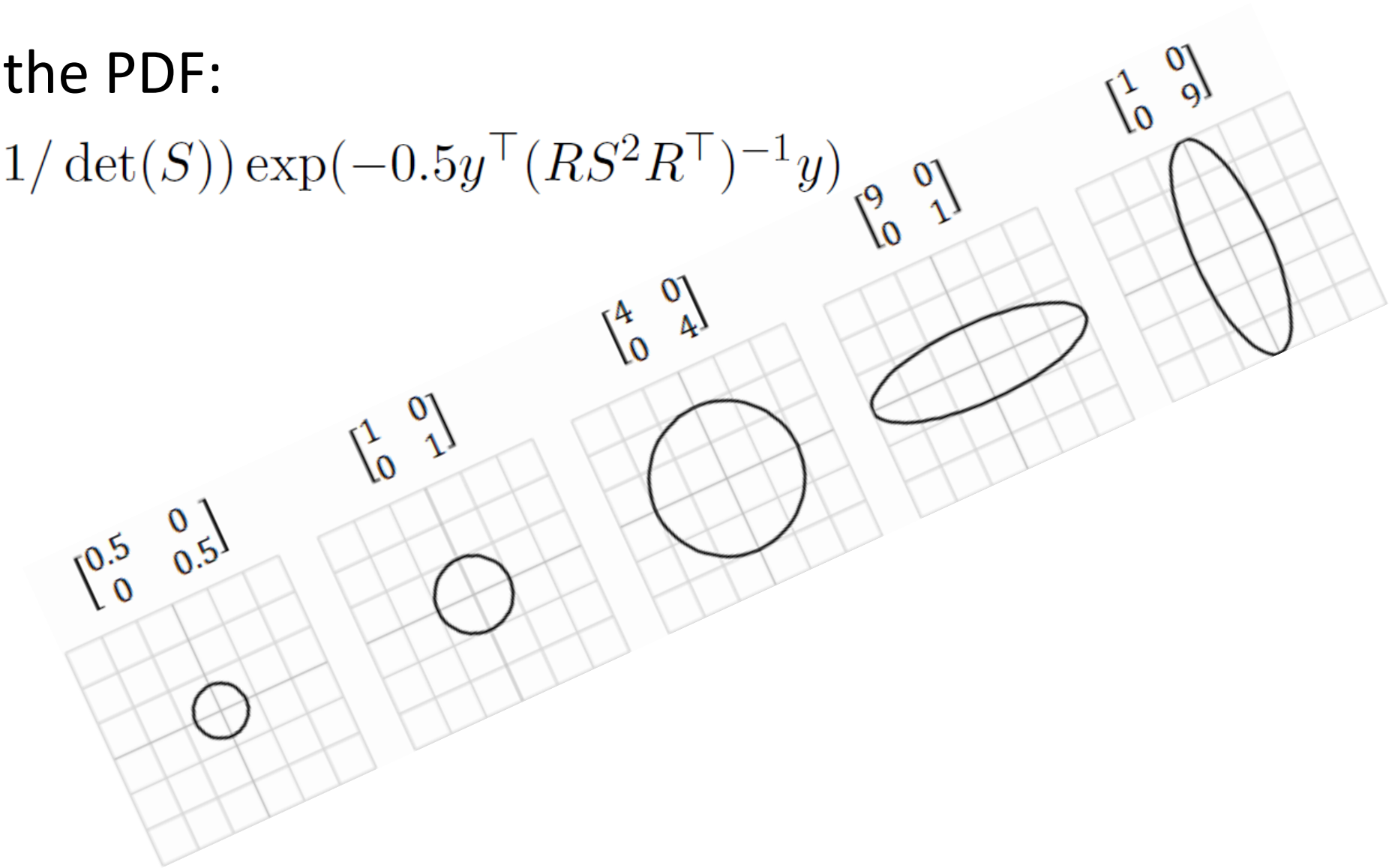
So, $Q(y) = (1/(2\pi)^{D/2})(1/\det(S))\exp(-0.5(S^{-1}R^\top y)^\top S^{-1}R^\top y)$
$= (1/(2\pi)^{D/2}(1/\det(S))\exp(-0.5y^\top(RS^2R^\top)^{-1}y)$

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 μ = 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

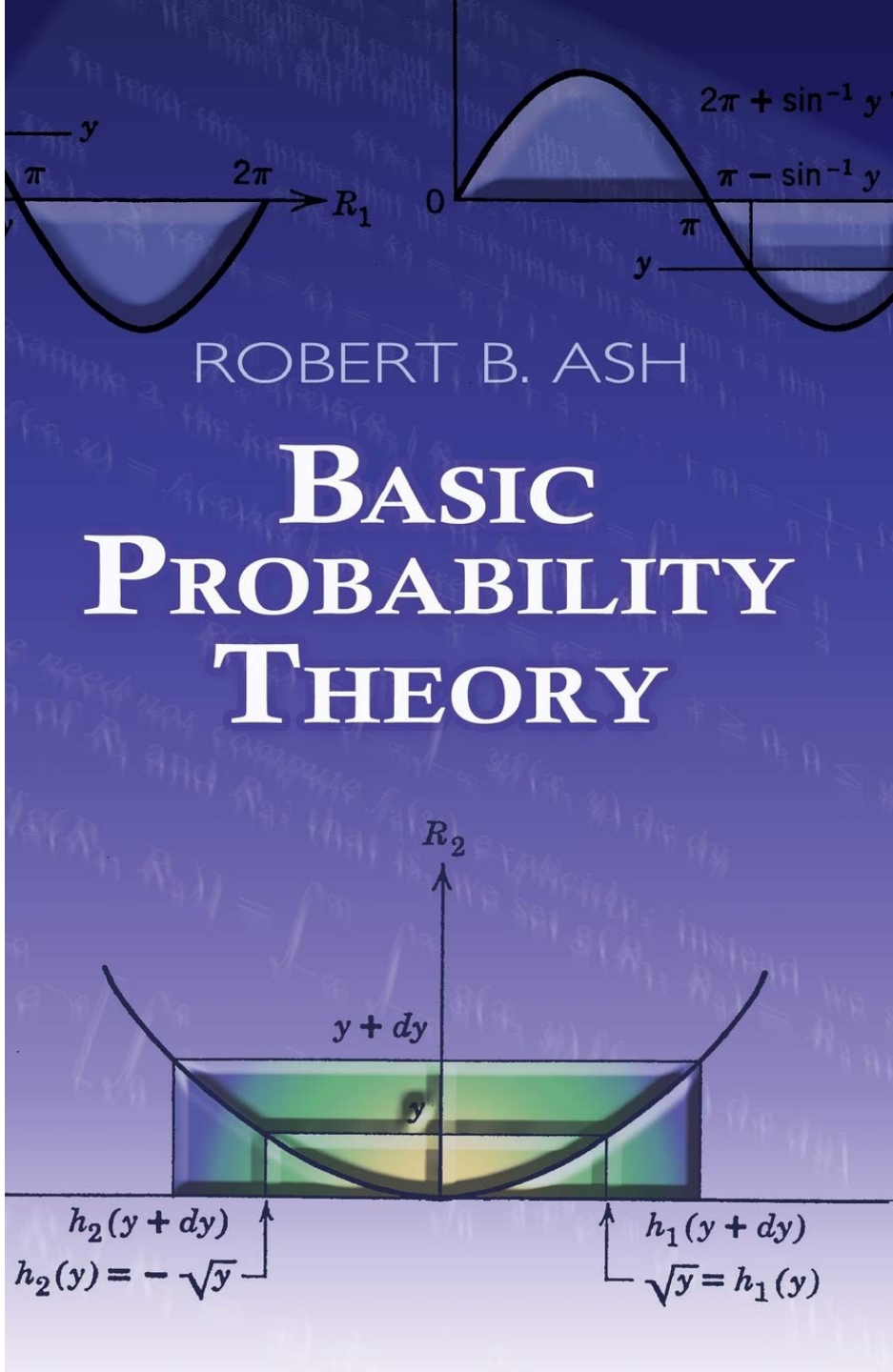
- If X := A W is a multivariate Gaussian, then Y := X + μ is a multivariate Gaussian with

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

  - 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 $C = AA^T$, and A is invertible, then C is positive definite

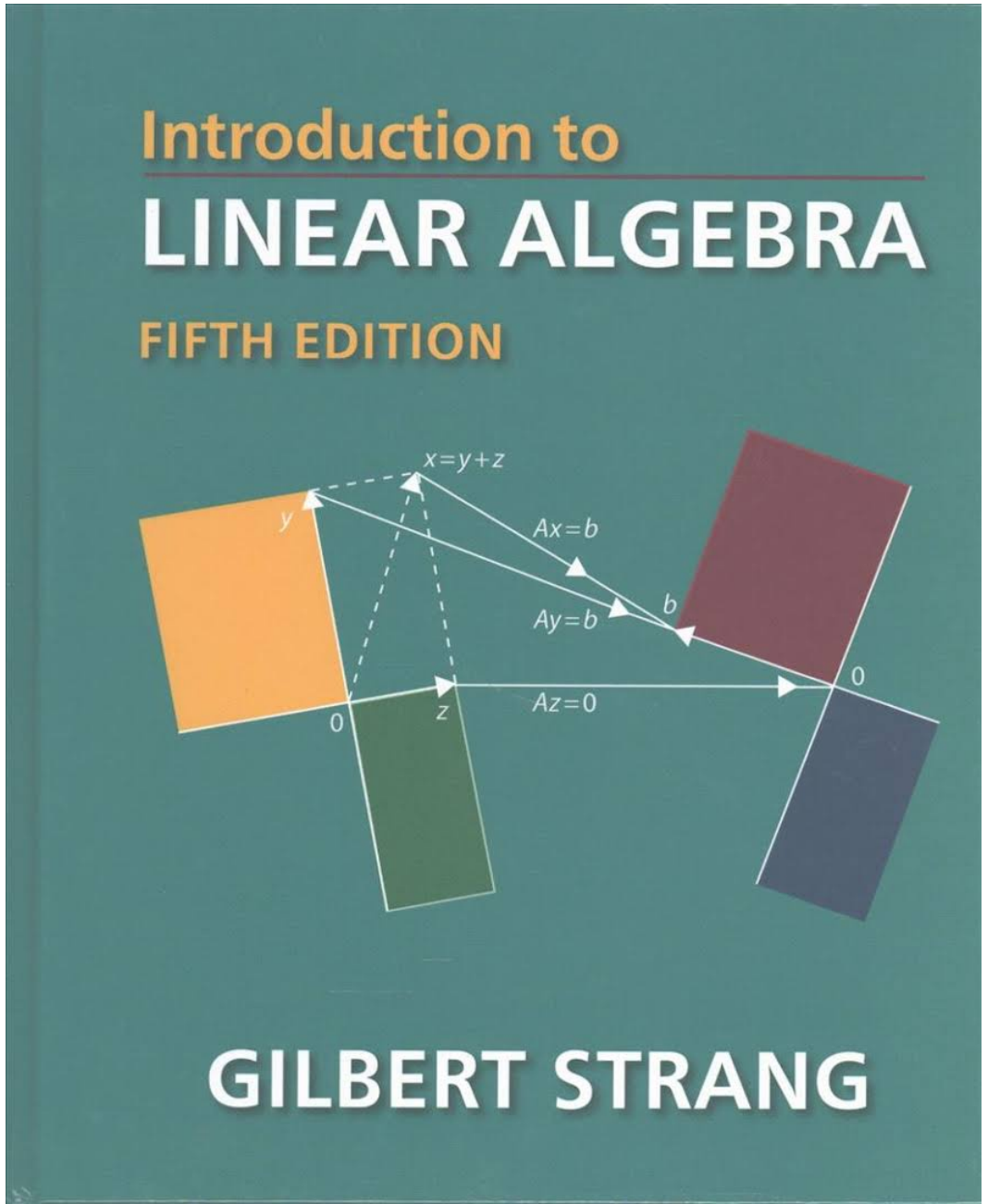
Probability and Statistics

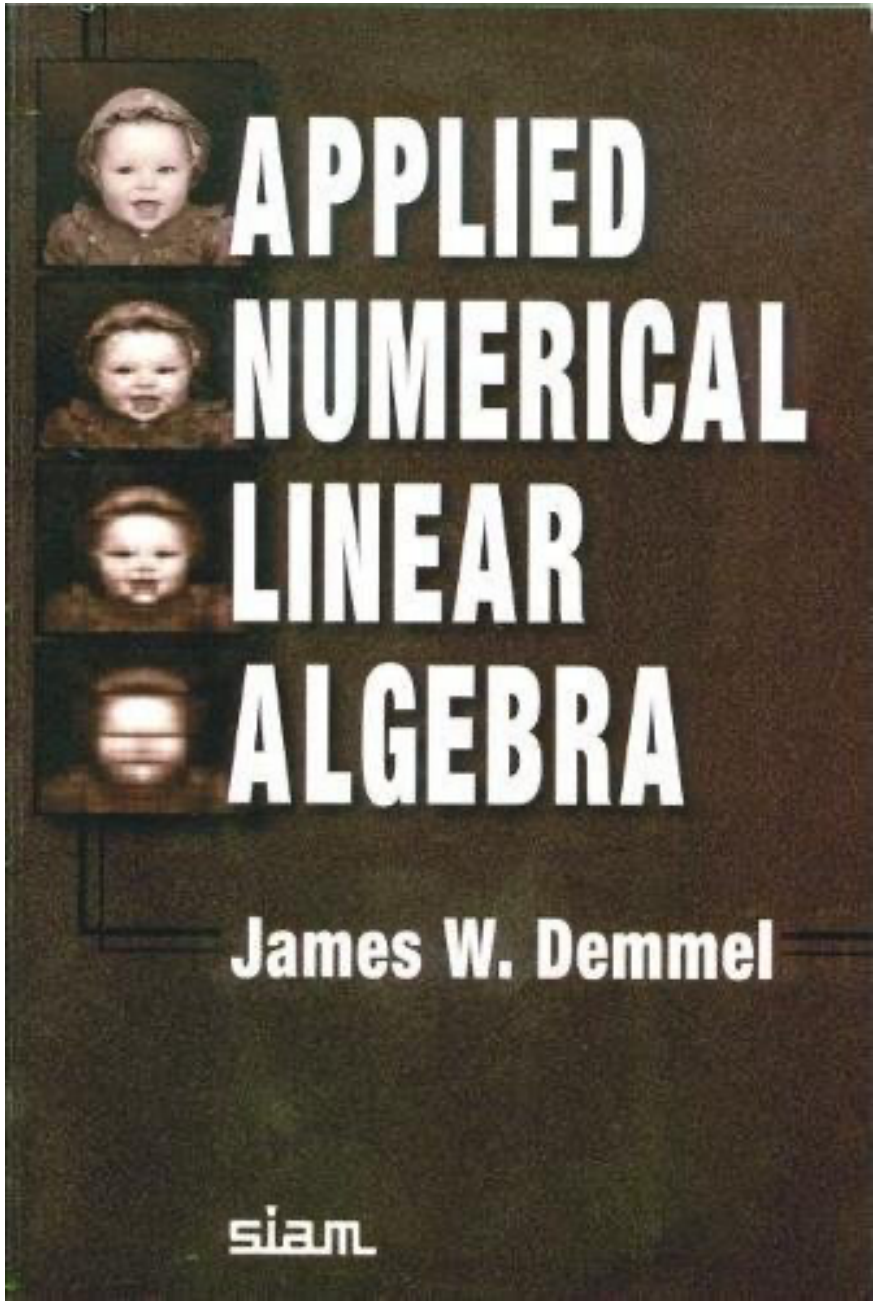
- Reference books specifically for multivariate Gaussian
- Basic Probability Theory, by Robert Ash
  - faculty.math.illinois.edu/~r-ash/BPT.html
  - Link 2



Linear Algebra

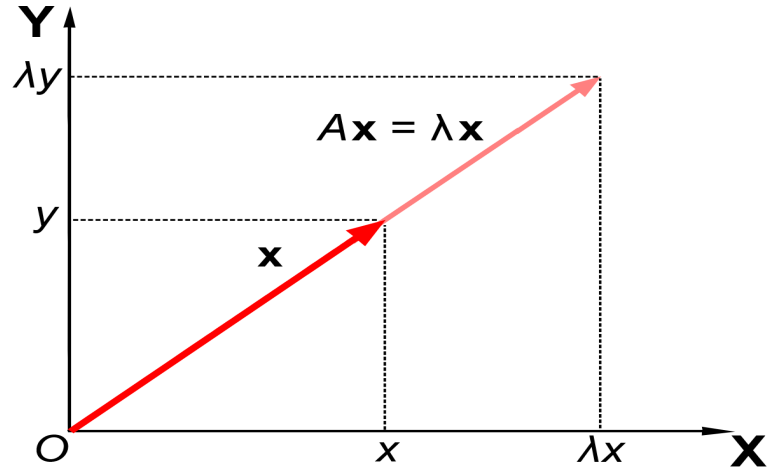
- Reference books





Linear Algebra – Eigen Decomposition

- Eigenvalue and Eigenvector
- For any square NxN matrix A,
  an **eigenvector** is a non-zero vector ‘v’ s.t. Av = λv.
  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 $\vec{\alpha}_i$

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

the above equation can be rewritten as

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

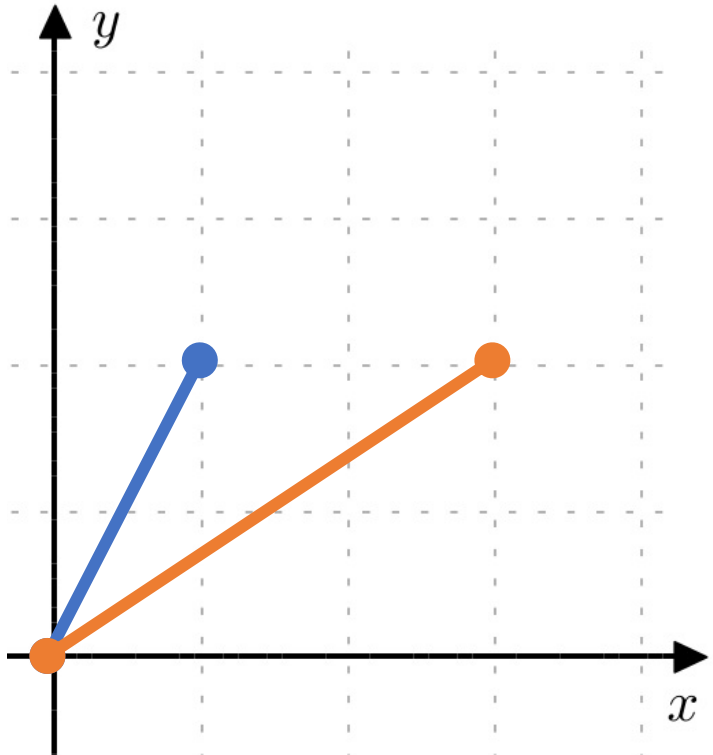
Linear Algebra – Eigen Decomposition

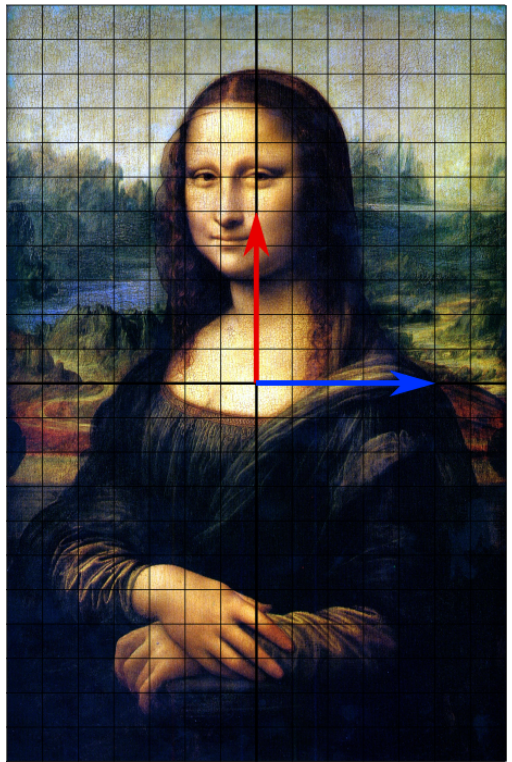
- Invertible doesn't imply diagonalizable
  - A non-diagonalizable matrix is called a defective matrix
  - e.g., 2x2 matrix A as shown. B = inv (A). [V D] = eig (A)
  - Doesn't have a complete basis of eigenvectors
    - Intuition: Action of matrix is to map vector (x,y) to (x+y,y)
      So, any eigenvalue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

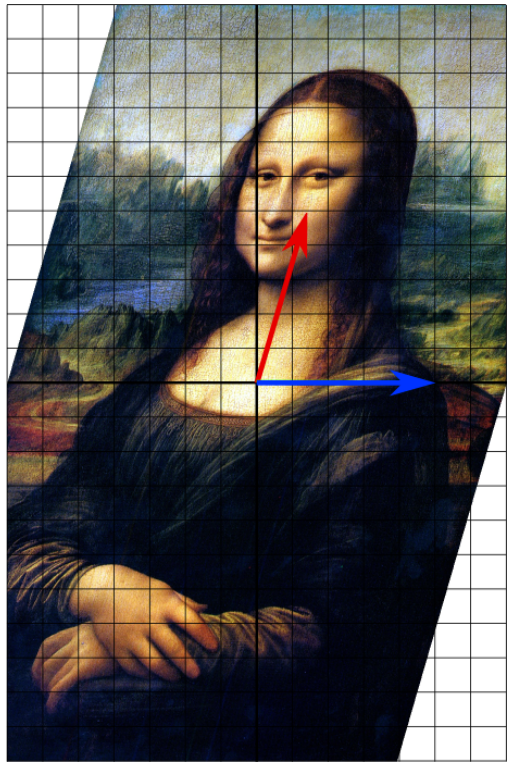
```
A =

     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** NxN matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors
  - If A is a **real symmetric** matrix, then A has all **real** eigenvalues.

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** NxN 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** NxN matrix, then A has N **real** eigenvalues with N **real-valued orthogonal** eigenvectors

• If A is a **real symmetric** matrix, then the eigenvectors of A corresponding to **distinct** eigenvalues are **orthogonal**.

Proof:

Let A have an eigenvector v_i with real eigenvalue λ_i
Let A have an eigenvector v_j with real eigenvalue $\lambda_j \neq \lambda_i$

Then, $\lambda_i v_i^\top v_j = (\lambda_i v_i)^\top v_j = (Av_i)^\top v_j = v_i^\top (Av_j)$ (because A is symmetric)

$= v_i^\top (\lambda_j v_j) = \lambda_j v_i^\top v_j$

Because $\lambda_i \neq \lambda_j$, we get $v_i^\top v_j = 0$

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
  - If A has some repeated values, then:
    - We already have the possibly-non-orthogonal eigenvectors as result of diagonalization
    - For a selected eigenvalue that is repeated,
      we take its corresponding set of eigenvectors and
      orthogonalize that set to get an orthogonal basis for that subspace
    - We do this for every repeated eigenvalue
    - Thus, it is always possible to construct a set of N orthogonal eigenvectors for A

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 + μ 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

Multivariate Gaussian – Level Sets

- If X = A W is a multivariate Gaussian, then Y = X + μ is a multivariate Gaussian with

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

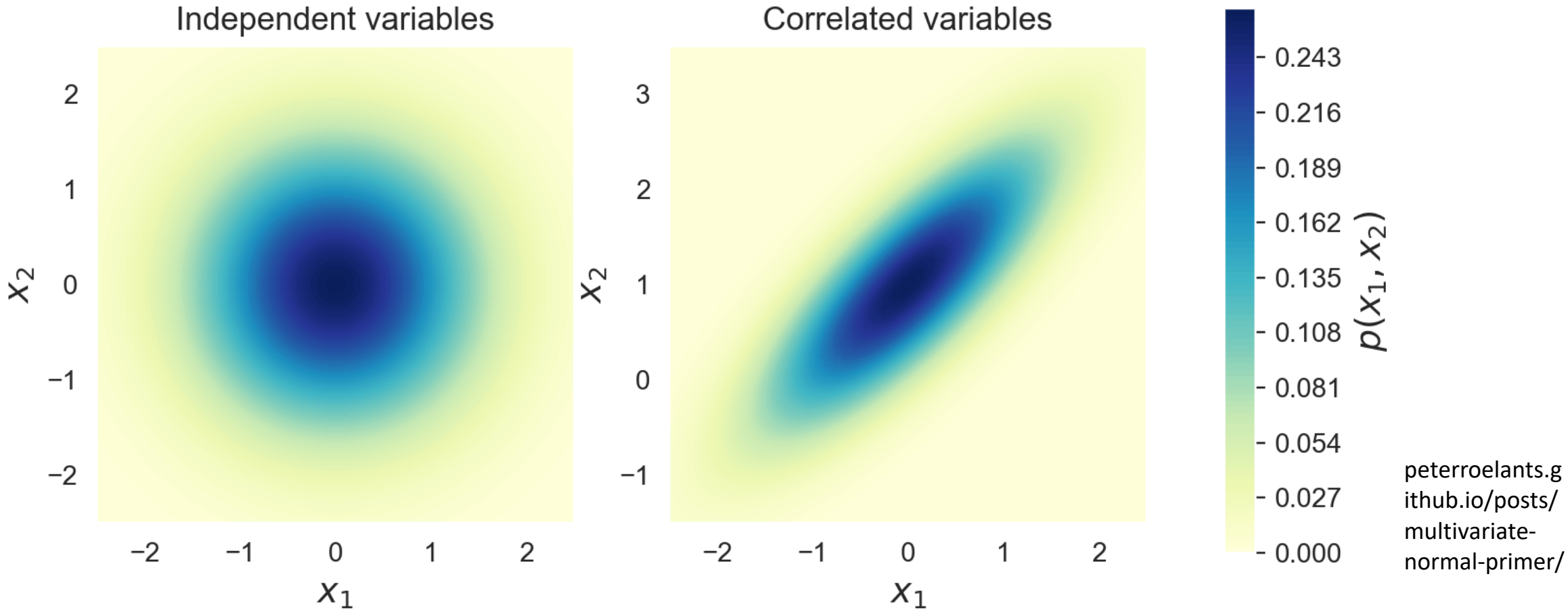
  - What are the **level sets** of this PDF ?
  - Let $C = Q\ D\ Q^T$. Then, $C^{-1} = Q\ D^{-1}\ Q^T$ that is also SPD
  - Each level set satisfies $(y-\mu)^T\ C^{-1}\ (y-\mu) = a$, where $a>=0$
    - Because C^{-1} is SPD; 'a' becomes zero iff $y=\mu$
  - So, $(y-\mu)^T\ Q\ D^{-1}\ Q^T\ (y-\mu) = a$
  - Change to roto-reflected coordinate system represented by orthogonal basis Q
    - Where y maps to $y' = Q^T\ y$, and μ maps to $\mu' = Q^T\ \mu$
  - Then, $(y'-\mu')^T\ D^{-1}\ (y'-\mu') = a$, 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 = A W is a multivariate Gaussian, then Y = X + μ 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 ?

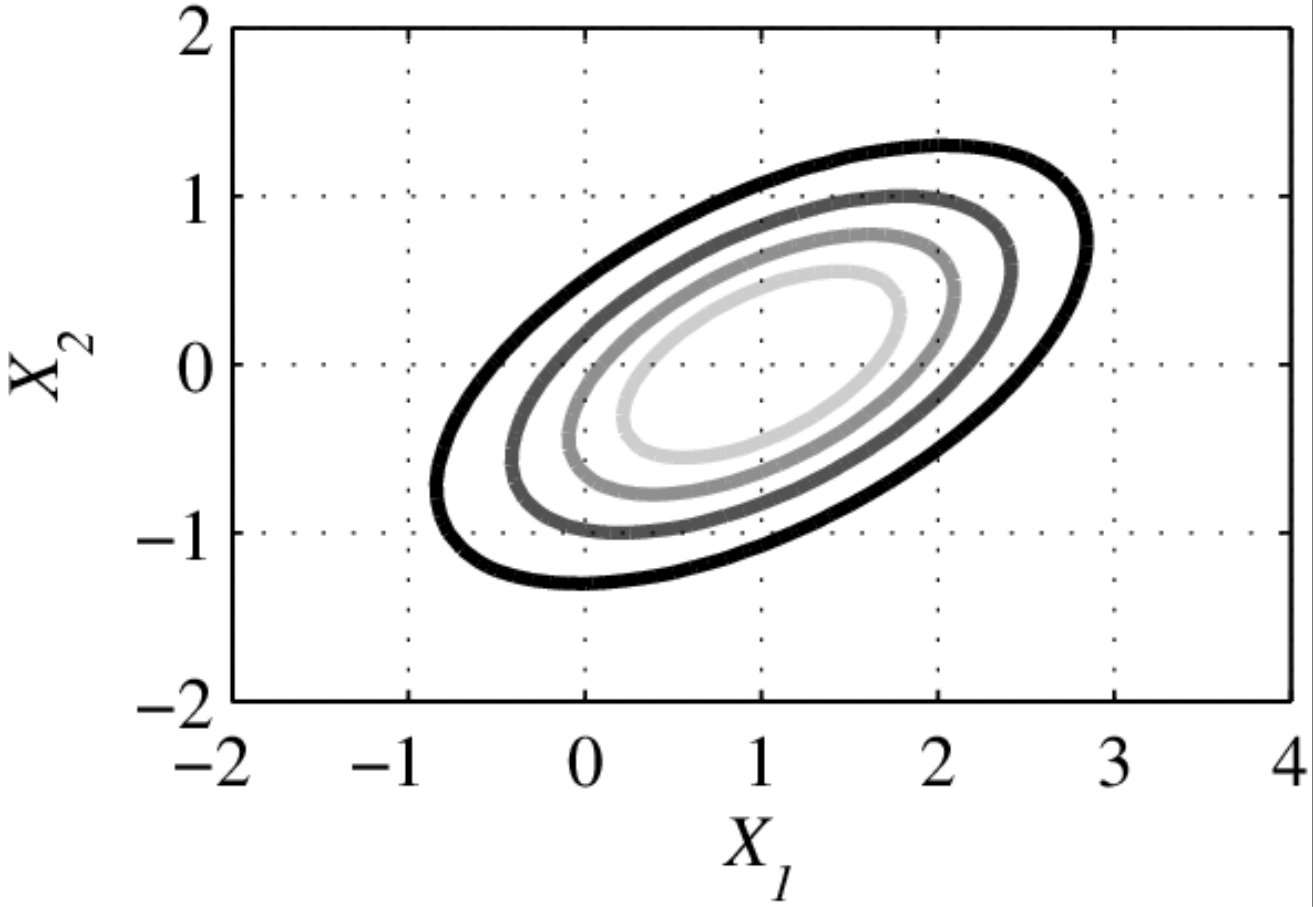


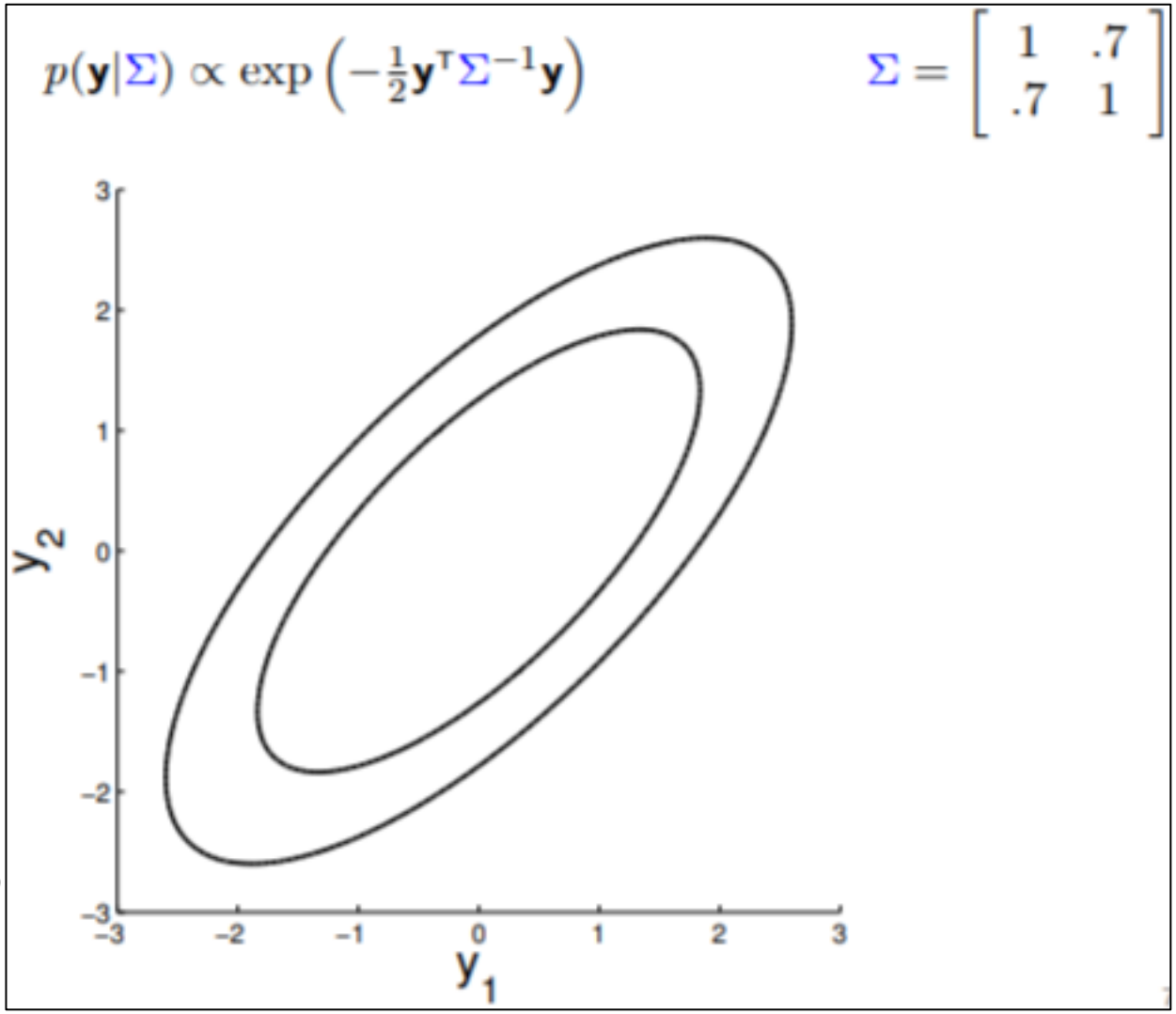
peterroelants.github.io/posts/multivariate-normal-primer/

Multivariate Gaussian – Level Sets

- If X = A W is a multivariate Gaussian, then Y = X + μ 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 ?

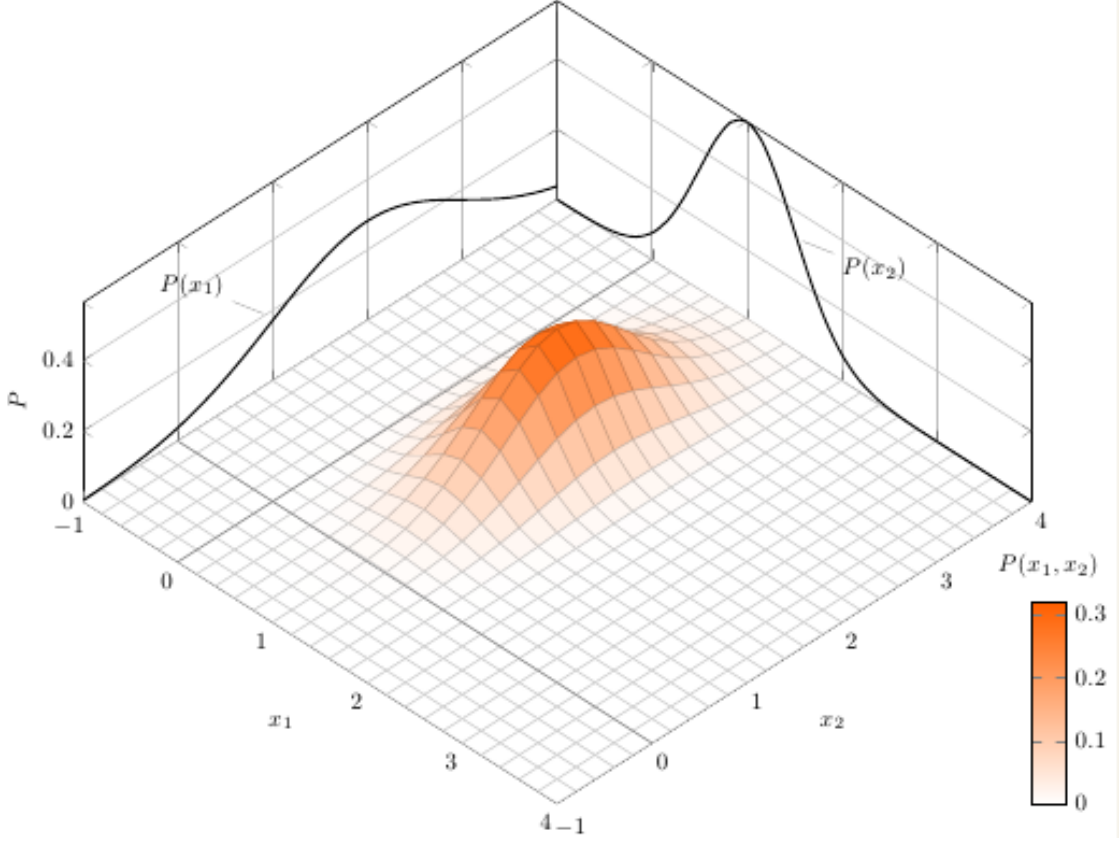




Multivariate Gaussian – Marginals and Conditionals

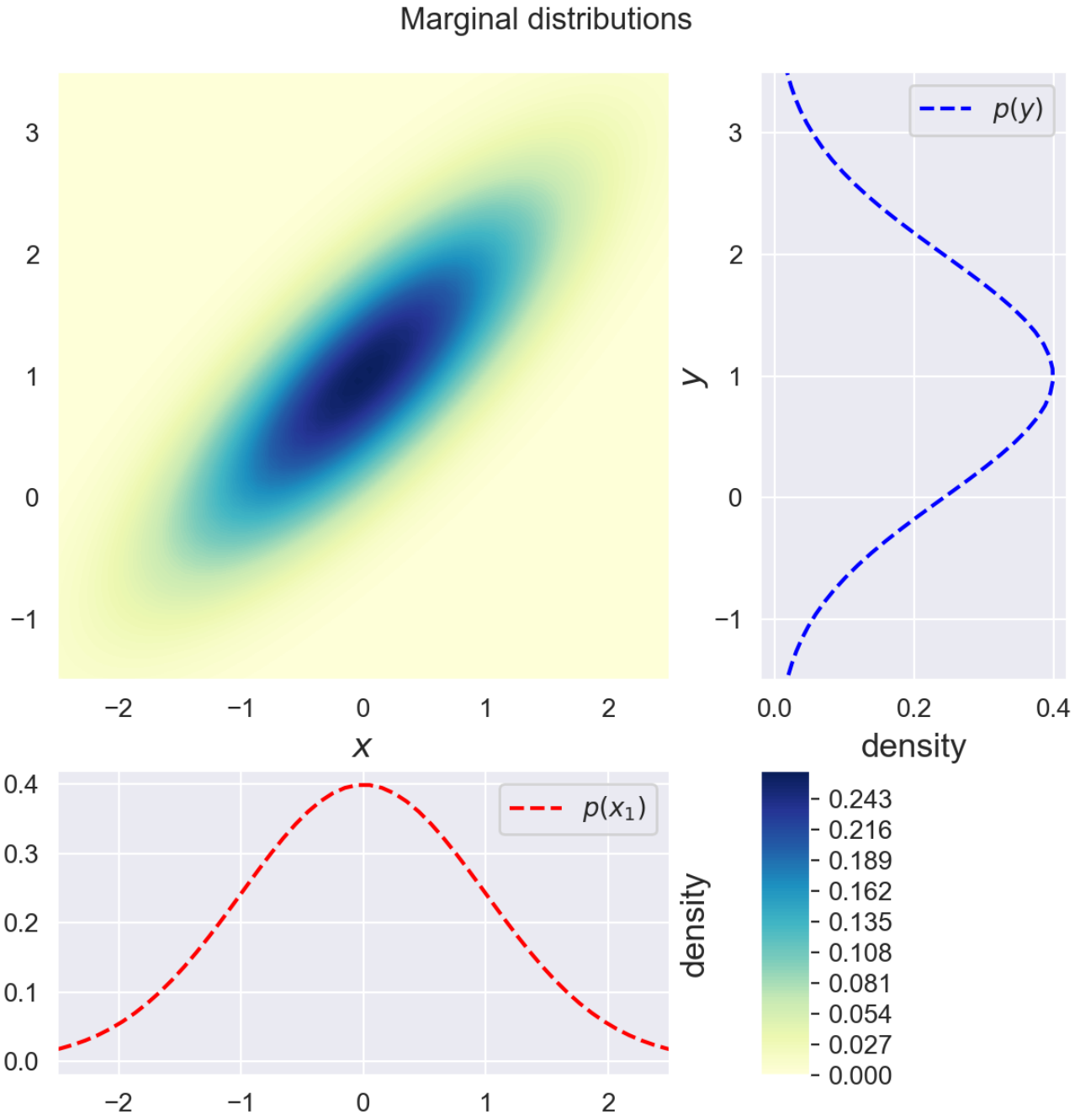
Multivariate Gaussian – Marginals

- **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

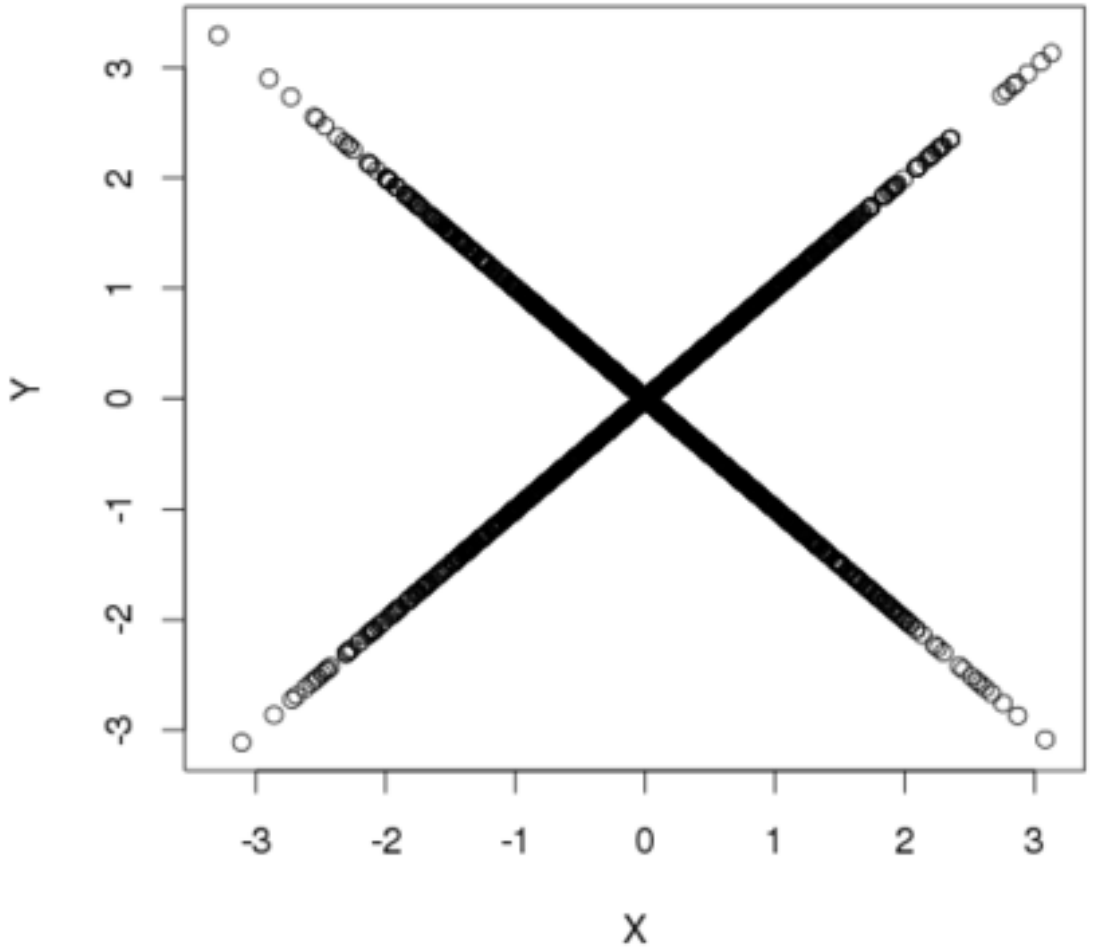


Multivariate Gaussian – Marginals

- **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 MxN, where M < N
    - Each row has all zeros except a 1 at one position
    - e.g., row [1 0 … 0] will select the first component of X
  - If we consider multivariate Gaussian X := AW + μ, where A is invertible, then BX = (BA)W + (Bμ)
    - Note: Because A is invertible (full rank), BA has rank M
    - By definition, BX is also multivariate Gaussian
      - Mean = Bμ, Covariance = $(BA)(BA)^T = BAA^TB^T = BCB^T = C'$, where C' is a square sub-matrix of C corresponding to the chosen M variables

Multivariate Gaussian – Marginals

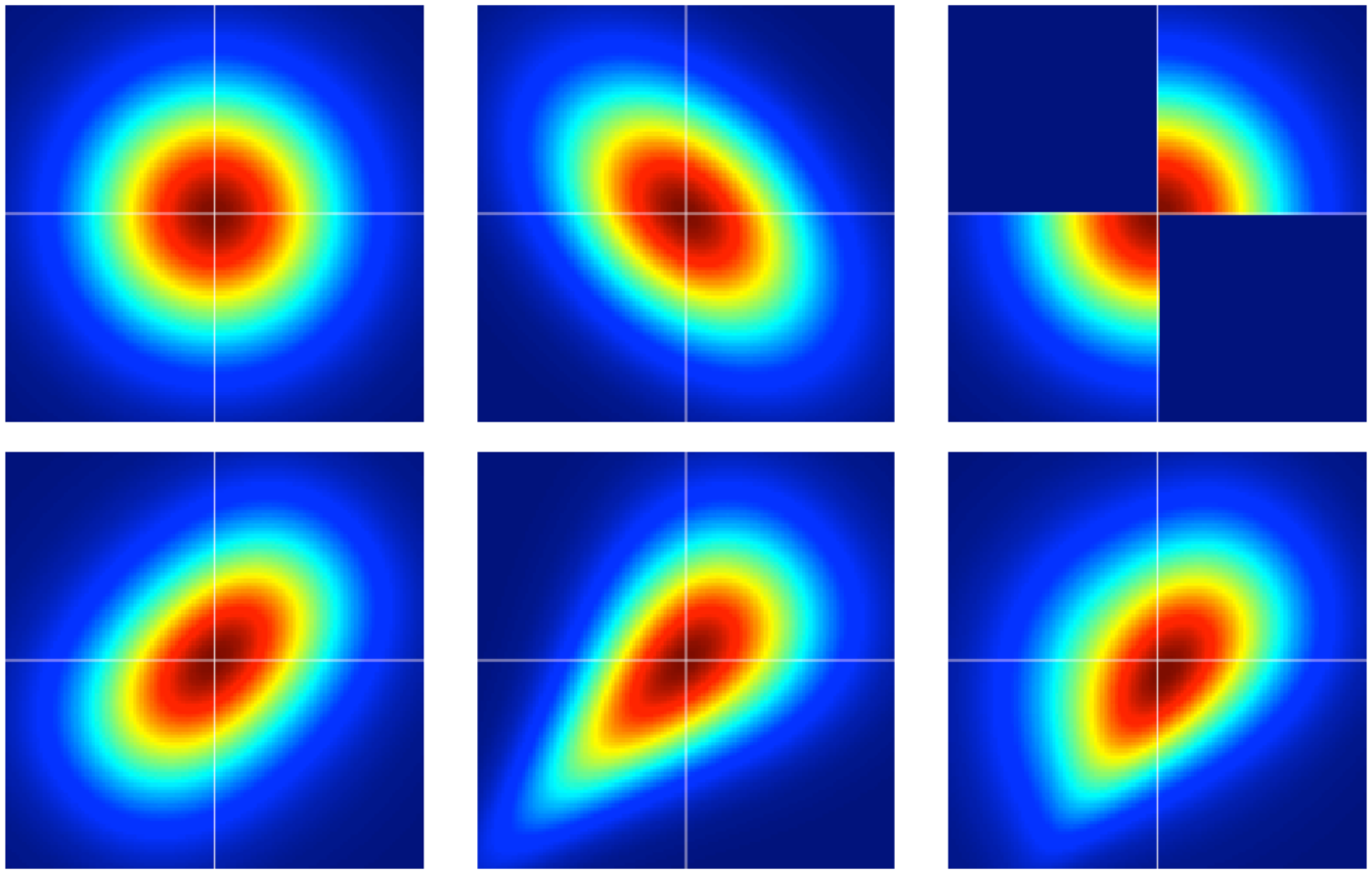
- Marginal PDFs being Gaussian doesn't imply joint PDF is multivariate Gaussian
- Example
  - Let X be a standard Normal
  - Let Y = X (2B – 1)
    where B is Bernoulli with parameter 0.5

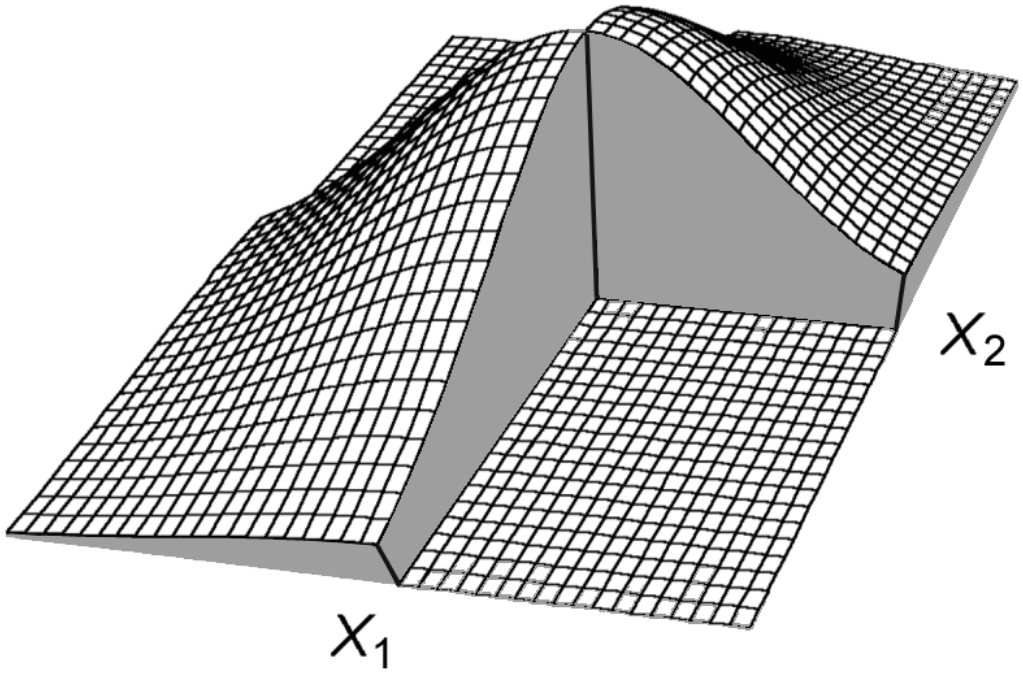


  - More examples
    - https://en.wikipedia.org/wiki/Normally_distributed_and_uncorrelated_does_not_imply_independent

Multivariate Gaussian – Marginals

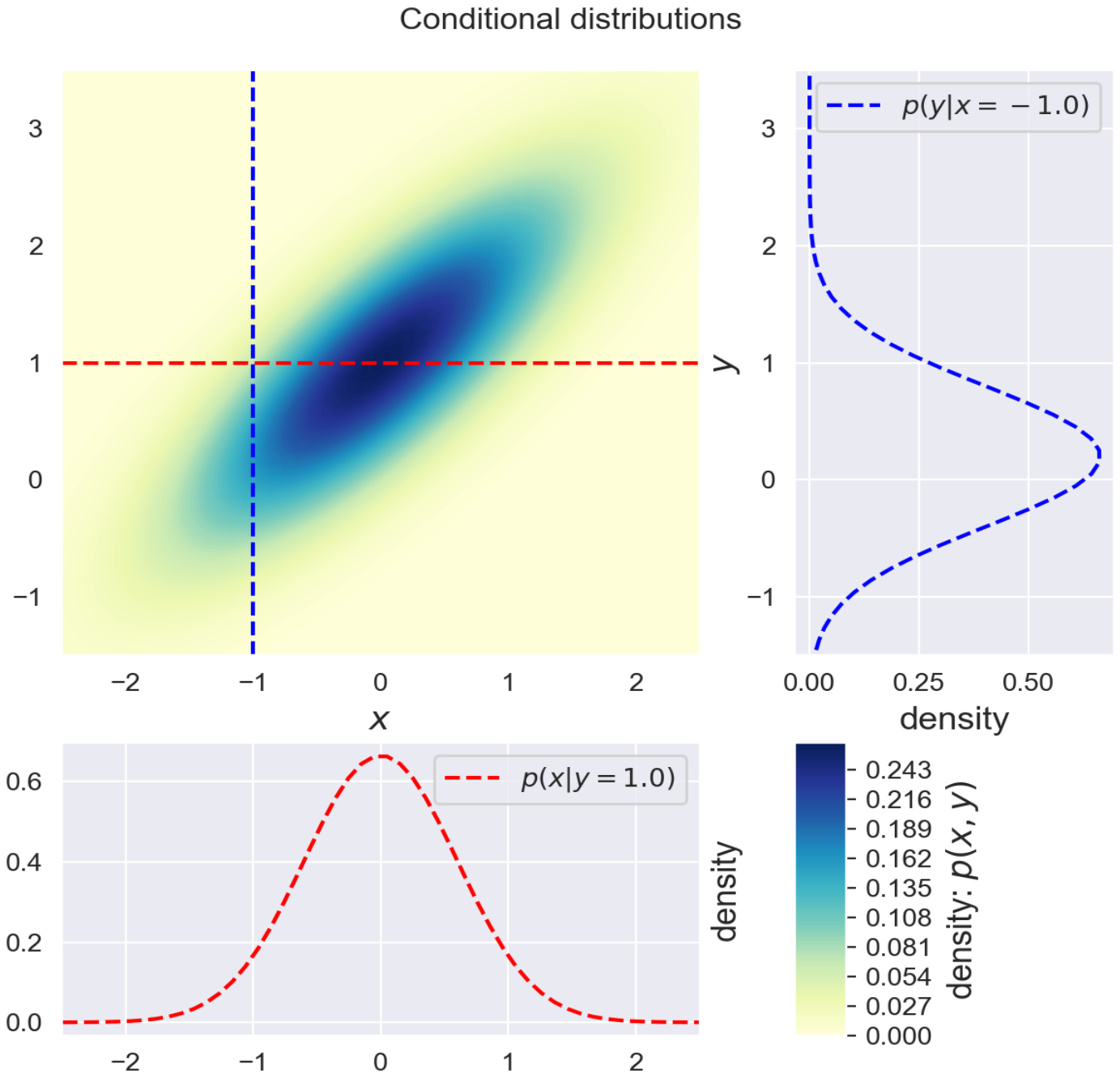
- Marginal PDFs being Gaussian doesn't imply joint PDF is multivariate Gaussian
  - Only top-row left, top-row middle, bottom-row left are bivariate Gaussian
  - All marginals are Gaussian





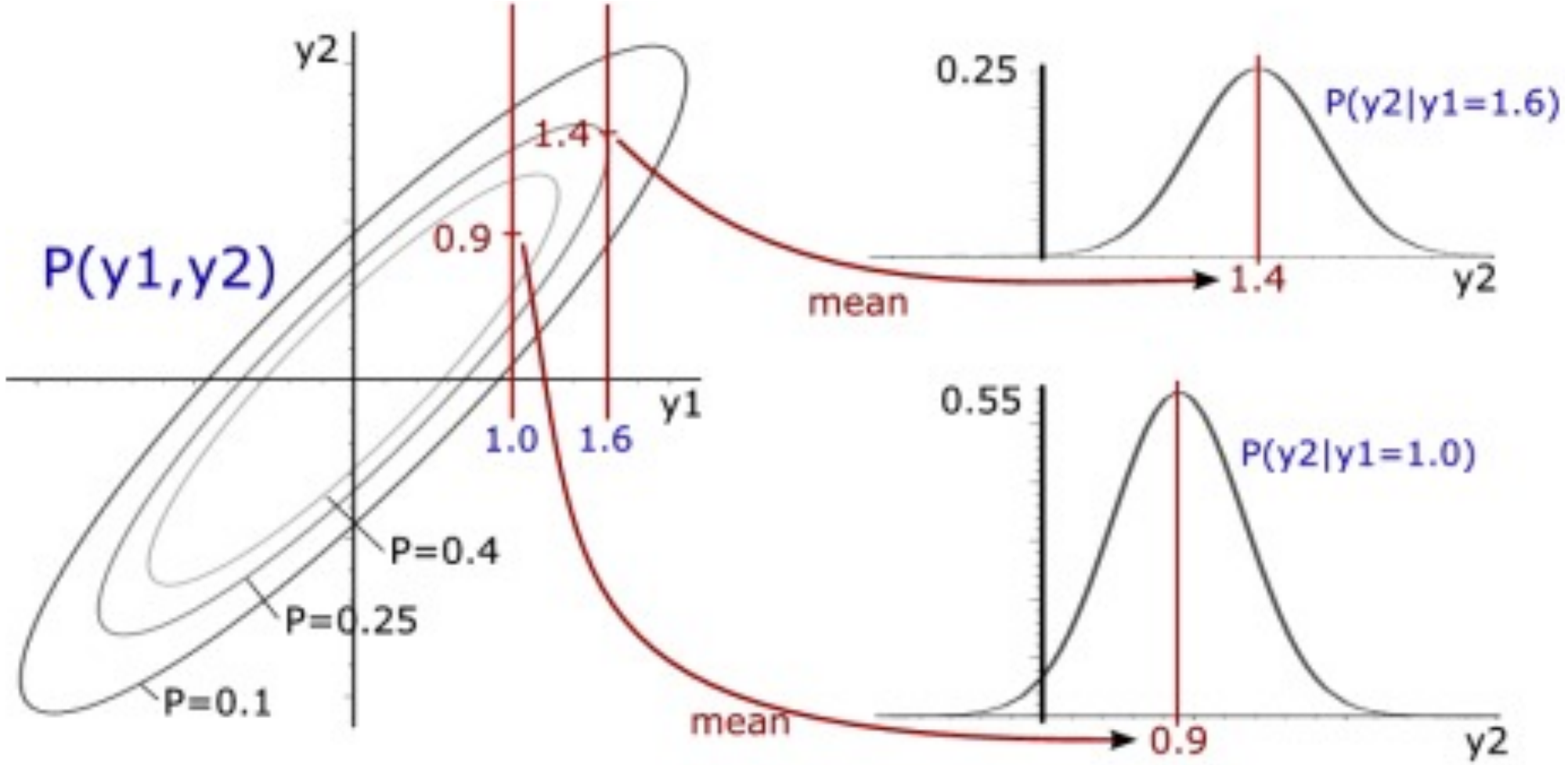
Multivariate Gaussian – Conditionals

- **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
  - $P(X_1 | X_2=x_2)$ = $P(X_1,X_2=x_2) / P(X_2=x_2)$



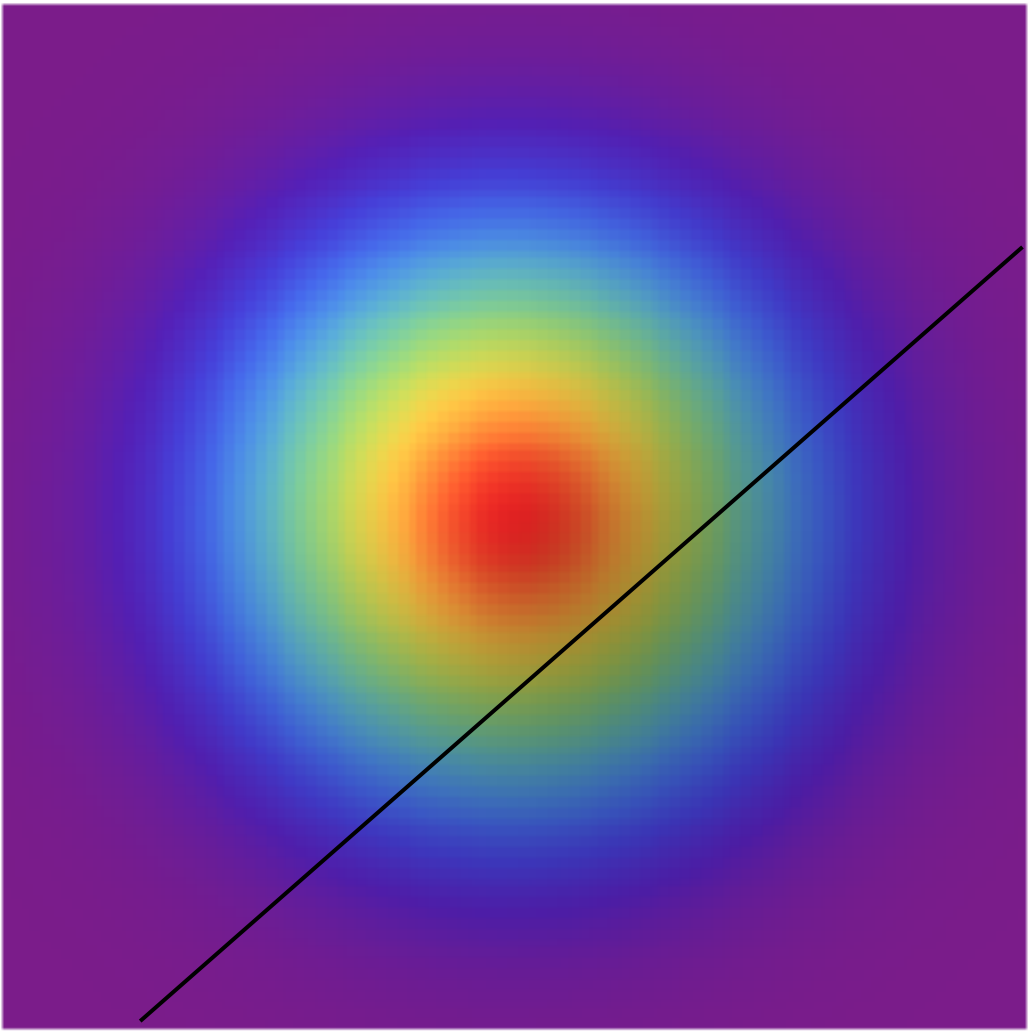
Multivariate Gaussian – Conditionals

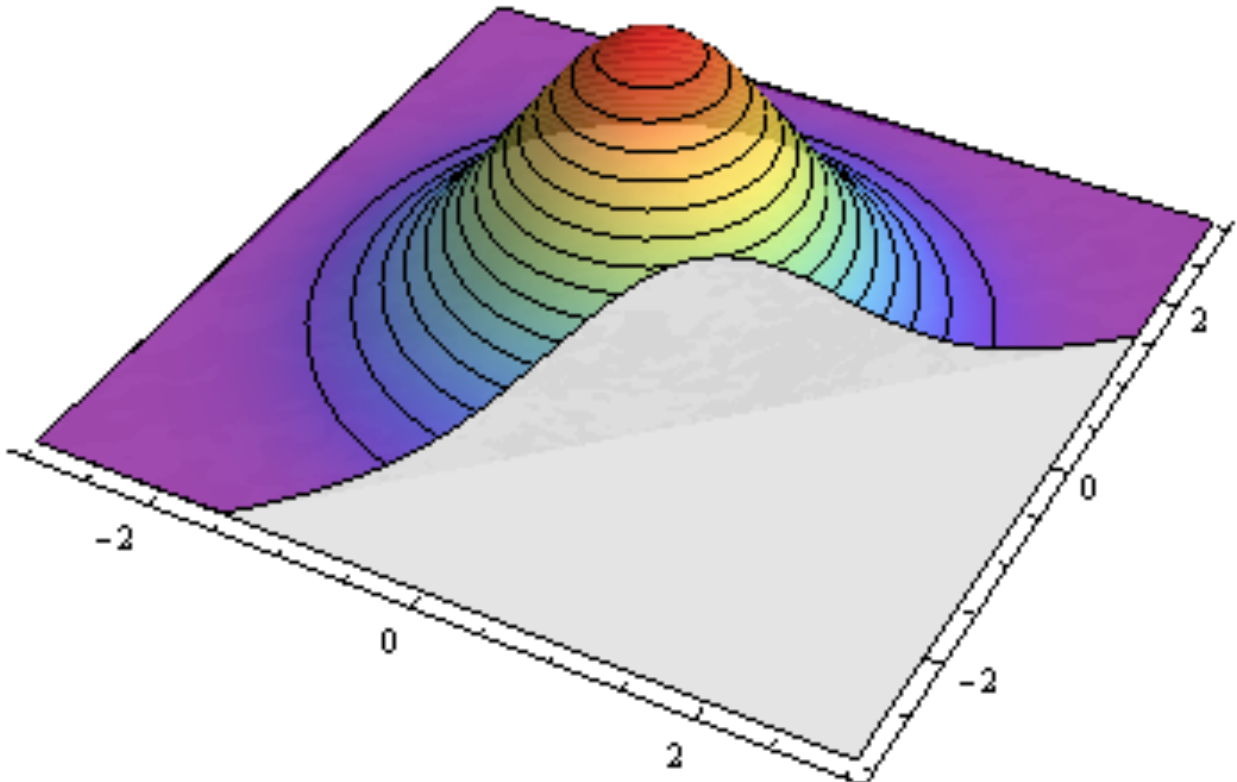
- **Conditional PDFs**
  - If multivariate Gaussian X is partitioned into X_1 and X_2, then the conditional PDF $P(X_1 | X_2=x_2)$ is also a multivariate Gaussian



Multivariate Gaussian – Conditionals

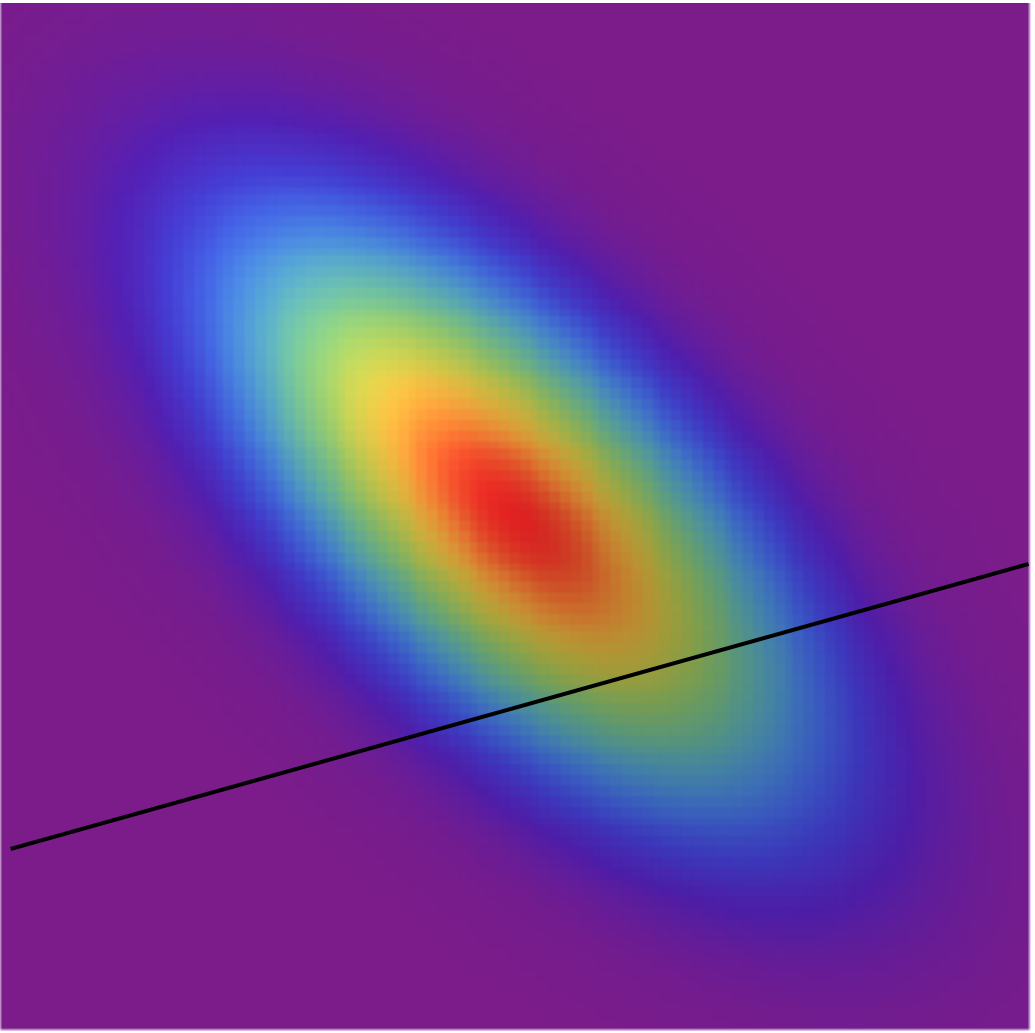
- **“Conditional” PDFs**
  - What about this way of slicing ?





Multivariate Gaussian – Conditionals

- **“Conditional” PDFs**
  - What about this way of slicing ?



Multivariate Gaussian – ML Estimation

Multivariate Gaussian – ML Estimation

- Data: $\{y_1, \ldots, y_N\}$

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

- Take log-likelihood function
- ML estimate for mean vector (= sample mean)
  - Take derivative w.r.t. μ, and assign to zero. Solve.
  - Quadratic form $a^\top Ba$
    $= \Sigma_i \Sigma_j a_i a_j B_{ij}$ (where B is symmetric)

$$\frac{d}{d\mu}(x-\mu)^\top C^{-1}(x-\mu) = 2C^{-1}(x-\mu)$$

  - Partial derivative w.r.t. a_k
    $= \Sigma_j a_j B_{kj} + \Sigma_i a_i B_{ik}$
    $= 2 \Sigma_j B_{kj} a_j$ (because B is symmetric)
    = 2 (k-th row of B * 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: $\{y_1, \ldots, y_N\}$

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

- 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

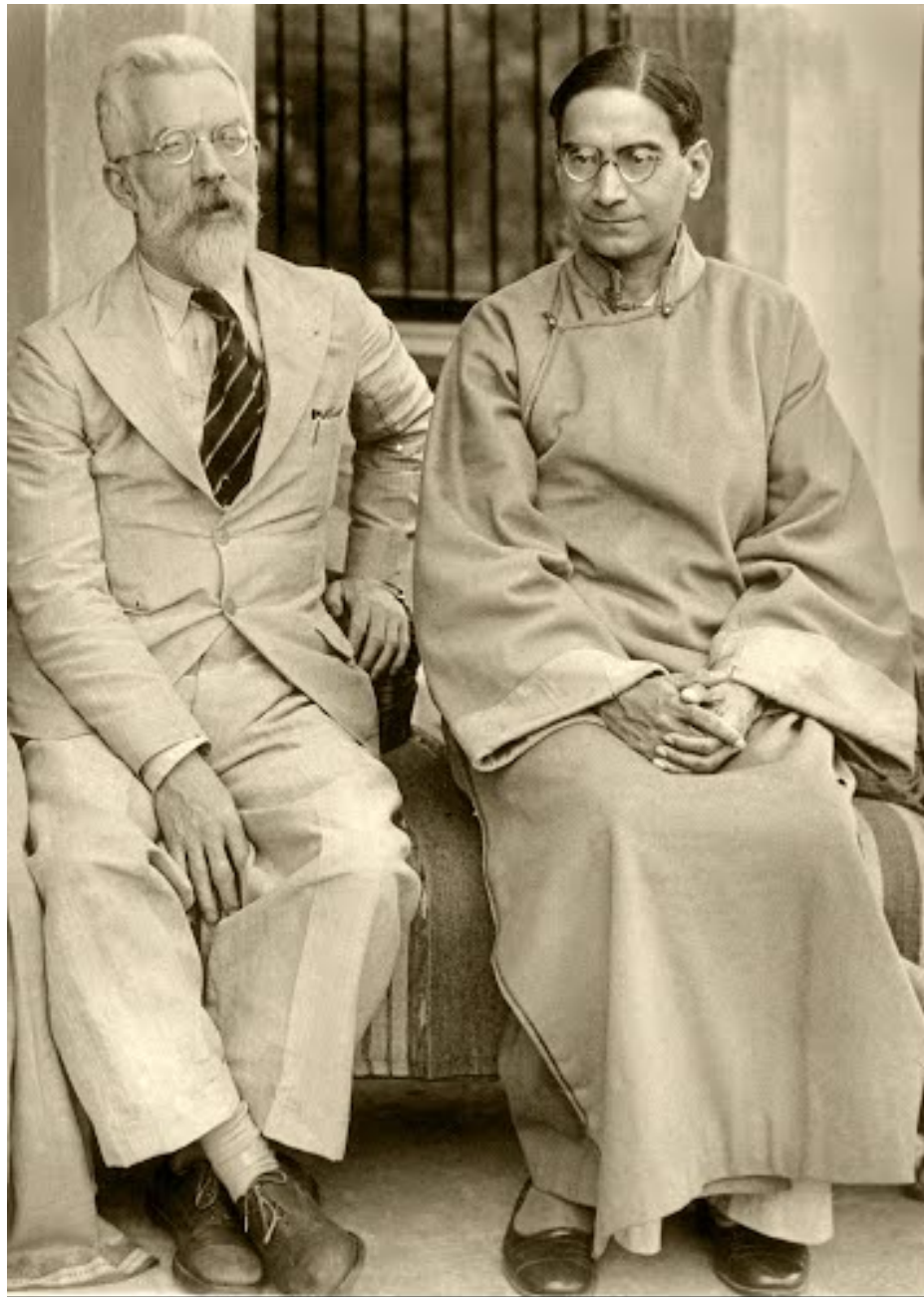
- “Matrix Calculus”
- http://www.ee.ic.ac.uk/hp/staff/dmb/matrix/calculus.html
- https://en.wikipedia.org/wiki/Matrix_calculus
- http://www.matrixcalculus.org/

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



Multivariate Gaussian – Mahalanobis Distance

- $d(y,\mu; C)^2 := (y-\mu)^T C^{-1} (y-\mu)$
- 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**
- When C is any SPD matrix, then ?

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

Multivariate Gaussian – Mahalanobis Distance

- $d(y,\mu; C)^2 := (y-\mu)^T C^{-1} (y-\mu)$
- Property: The Mahalanobis distance is a true distance metric
- Proof:
  - A distance metric is a function d(.,.)->Real 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) <= d(x,z) + d(z,y)
    - These imply non-negativity (i.e., d(x,y) >= 0, for all x,y):
      0 = d(x,x) <= d(x,y) + d(y,x) = 2 d(x,y)
  - In our case of SPD matrix C:
    - C being S**PD** implies: d(x,y; C) >= 0 for all x,y
    - C being S**PD** implies: d(x,y; C) = 0 iff x=y
    - C being **S**PD implies: 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 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 <= RHS is equivalent to showing LHS^2 <= RHS^2

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

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

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

Multivariate Gaussian – Mahalanobis Distance

- Property: The Mahalanobis distance is a true distance metric
- Proof (when covariance matrix 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 $u + v = x - y$

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

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

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

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

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

RHS^2 $= u^\top C^{-1} u + v^\top C^{-1} v + 2\sqrt{u^\top C^{-1} u}\sqrt{v^\top C^{-1} v}$

$= \sum_d u_d^2/\sigma_d^2 + \sum_d v_d^2/\sigma_d^2 + 2\sqrt{\sum_d u_d^2/\sigma_d^2}\sqrt{\sum_d v_d^2/\sigma_d^2}$

Multivariate Gaussian – Mahalanobis Distance

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

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

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

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

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 $= 2 \parallel a \parallel \parallel b \parallel$

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 $|\langle a, b\rangle| \leq \parallel a \parallel \parallel b \parallel$ 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):

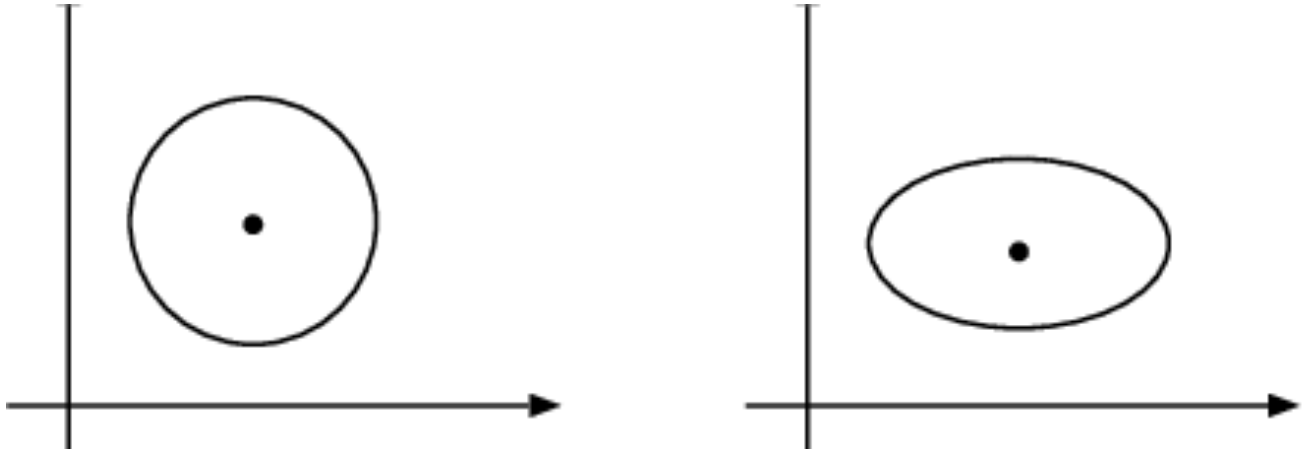
3) Triangular inequality: $d(x, y) \leq d(x, z) + d(z, y), \forall x, y, z$

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 = Identity) ?
  - How do the level sets change ?

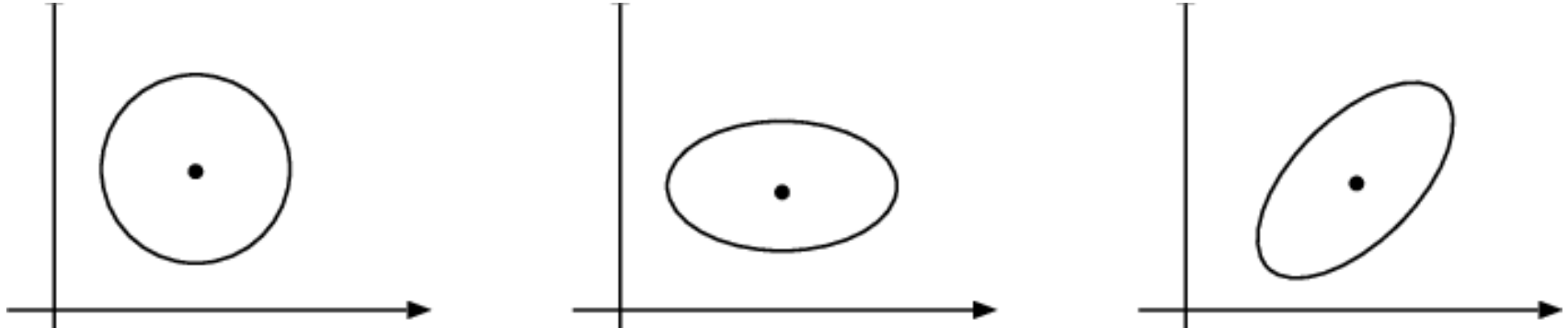
Let $X := SW$, where S is a diagonal matrix that rescales the units along each coordinate axes
Then, what is the covariance matrix ? $A = S$. Thus, $C = AA^\top = SS^\top = S^2$
Then, Mahalanobis distance between x and the mean (origin) is $x^\top C^{-1} x = x^\top S^{-2} x$



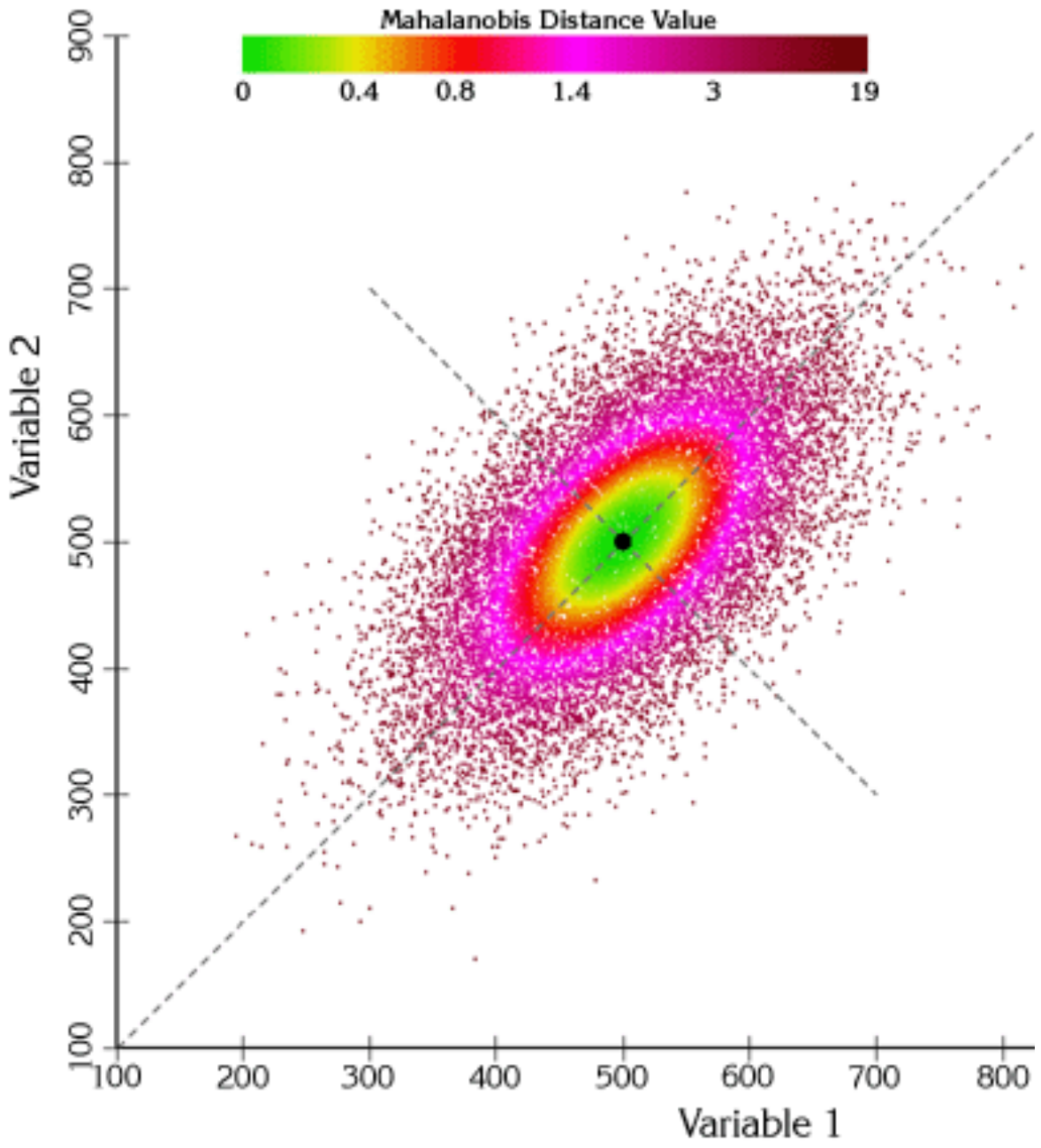
Multivariate Gaussian – Mahalanobis Distance

- 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 = 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

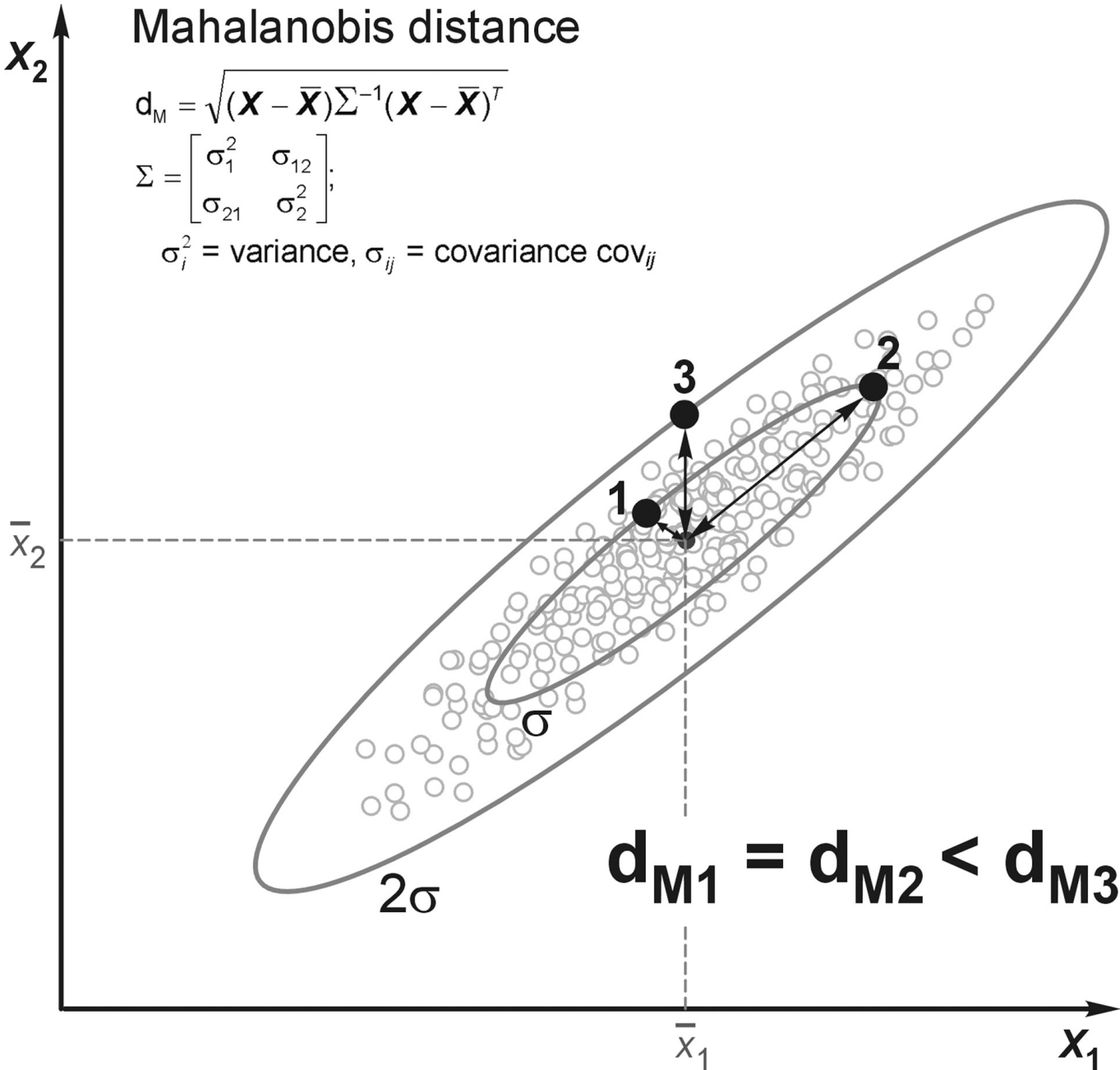


Mahalanobis distance

$$d_M = \sqrt{(\boldsymbol{X} - \bar{\boldsymbol{X}})\Sigma^{-1}(\boldsymbol{X} - \bar{\boldsymbol{X}})^T}$$

$$\Sigma = \begin{bmatrix} \sigma_1^2 & \sigma_{12} \\ \sigma_{21} & \sigma_2^2 \end{bmatrix};$$

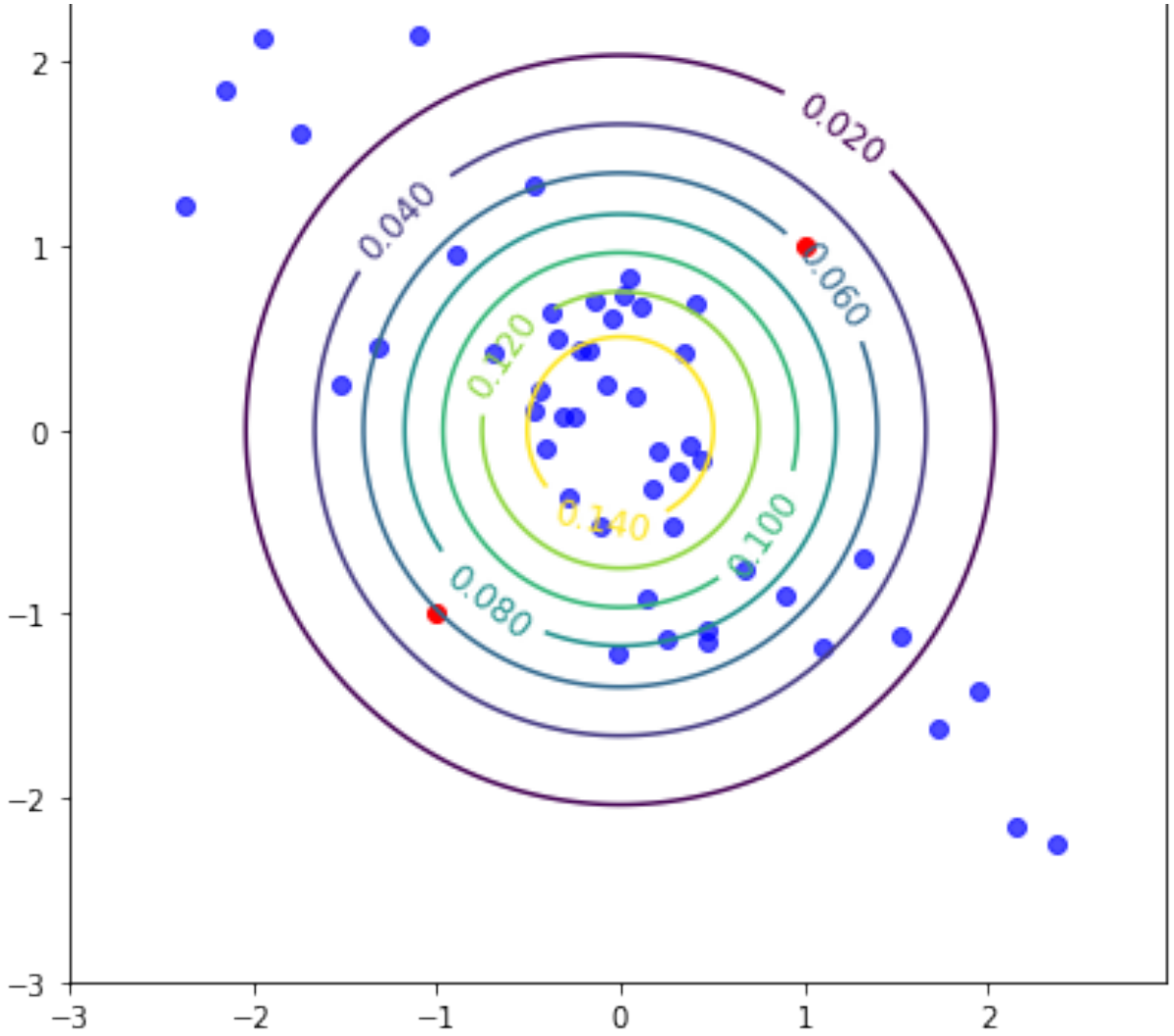
σ_i^2 = variance, σ_{ij} = covariance cov_{ij}

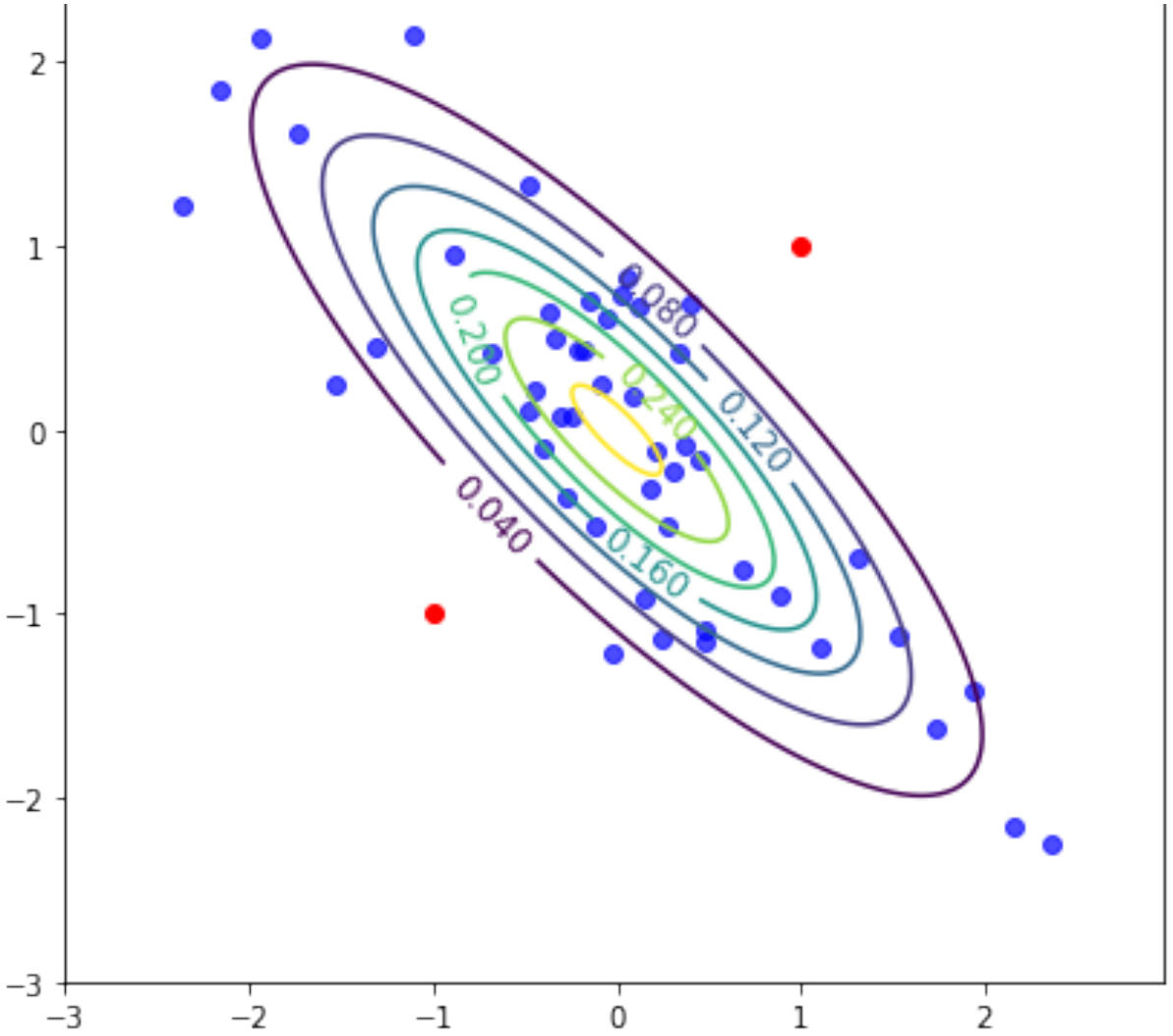


Multivariate Gaussian – Applications

Multivariate Gaussian – Applications

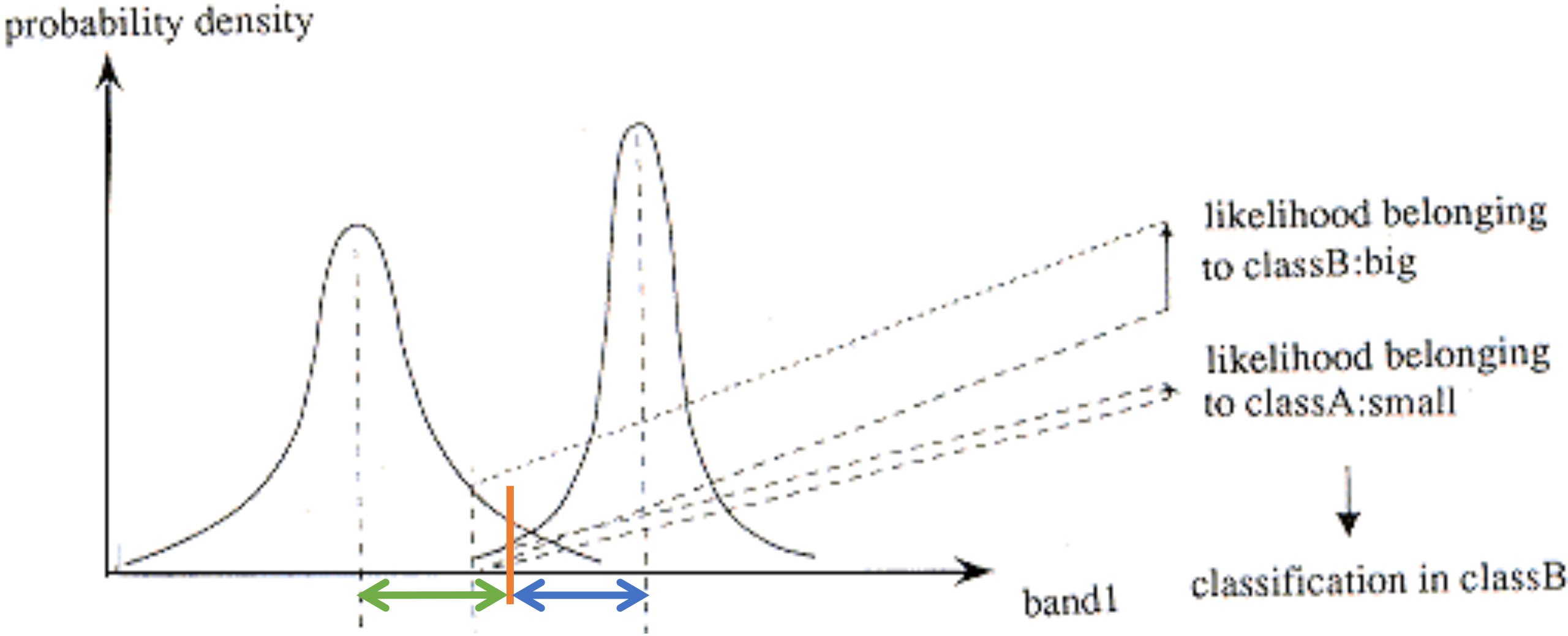
- Multivariate Gaussian (Mahalanobis distance) for **anomaly detection**





Multivariate Gaussian – Applications

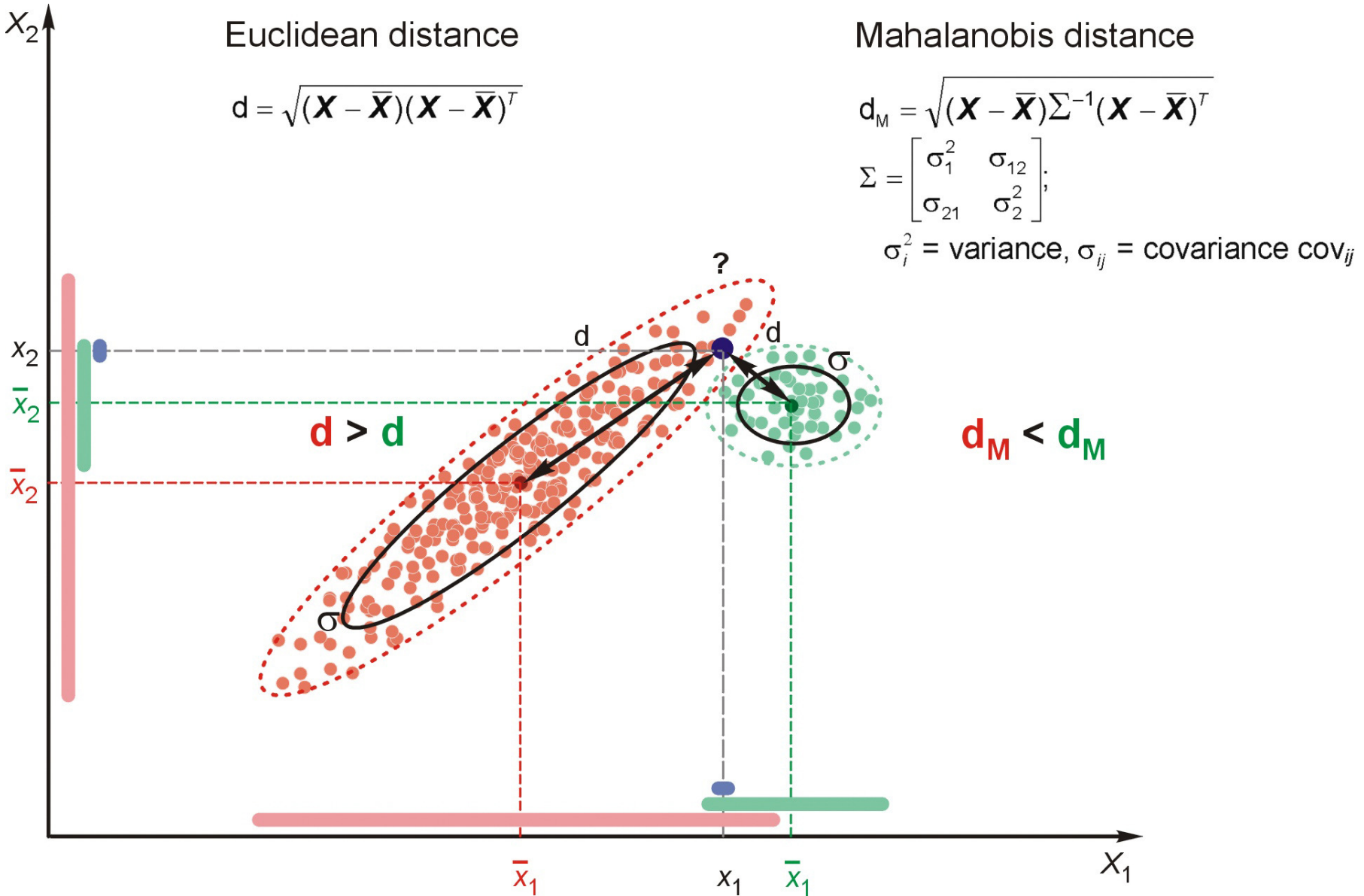
- Multivariate Gaussian for maximum-likelihood **classification**



http://sar.kangwon.ac.kr/etc/rs_note/rsnote/cp11/cp11-7.htm

Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood classification



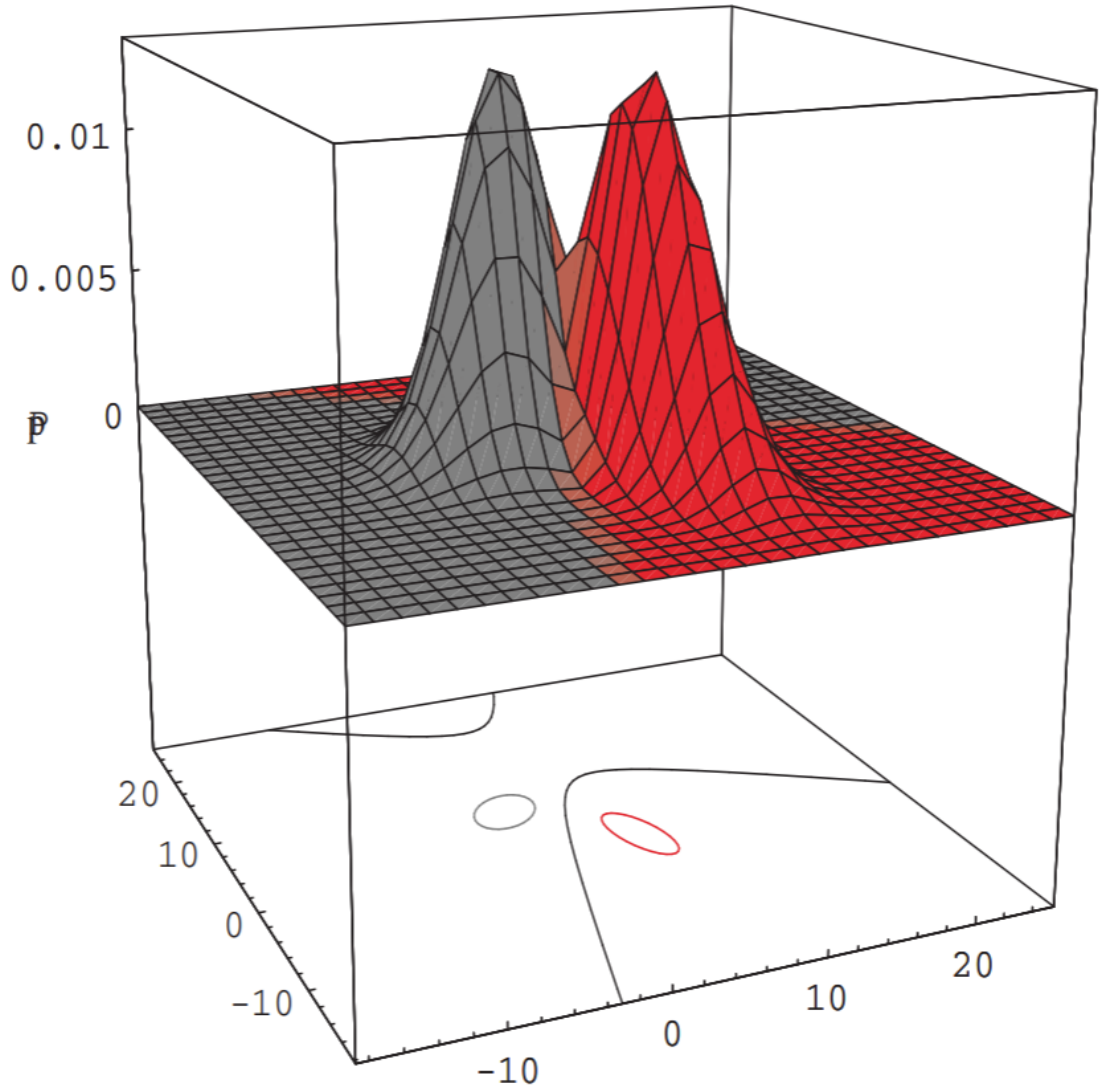
https://onlinelibrary.wiley.com/doi/full/10.1111/maps.13314

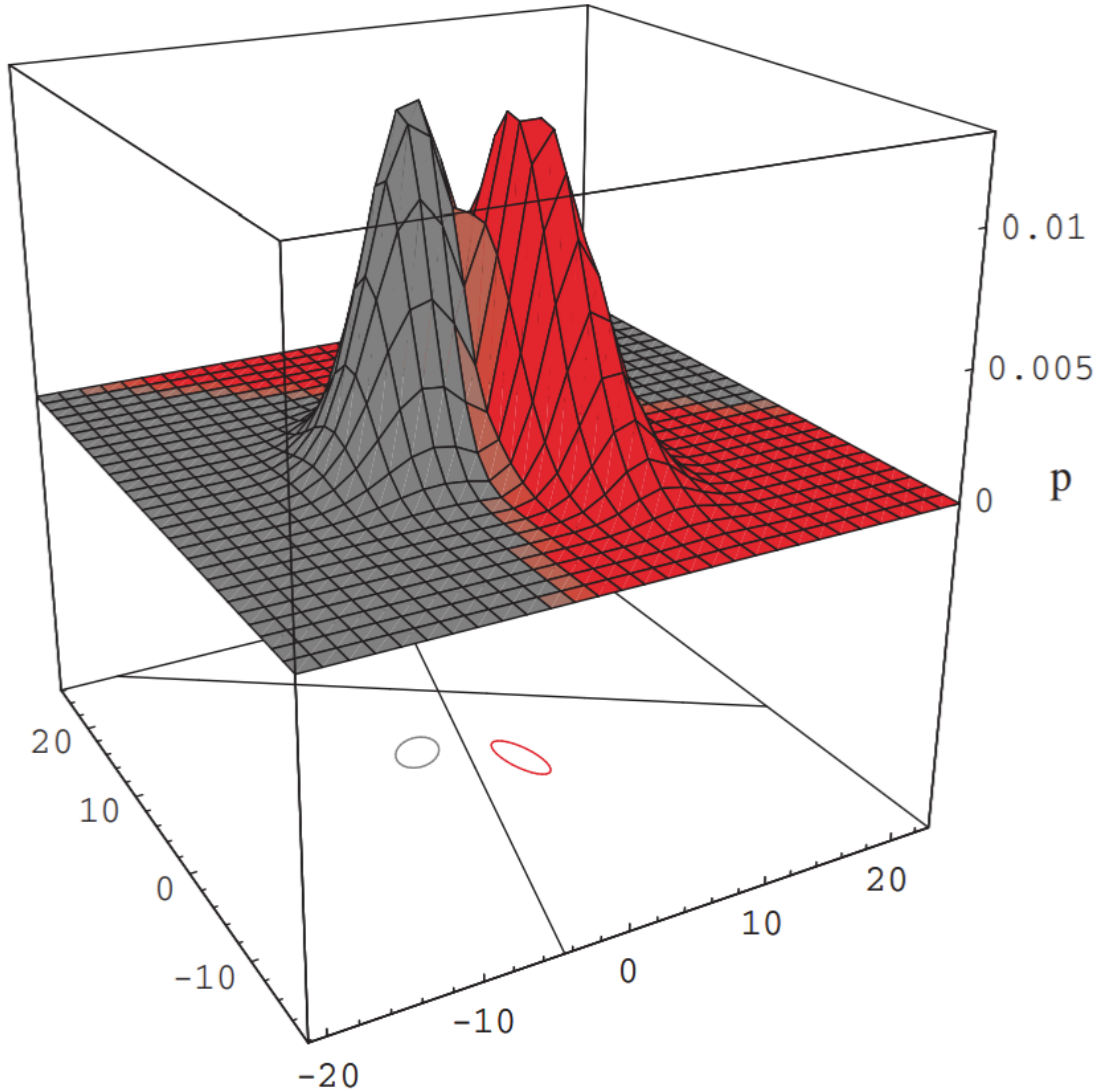
Multivariate Gaussian – Applications

- 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

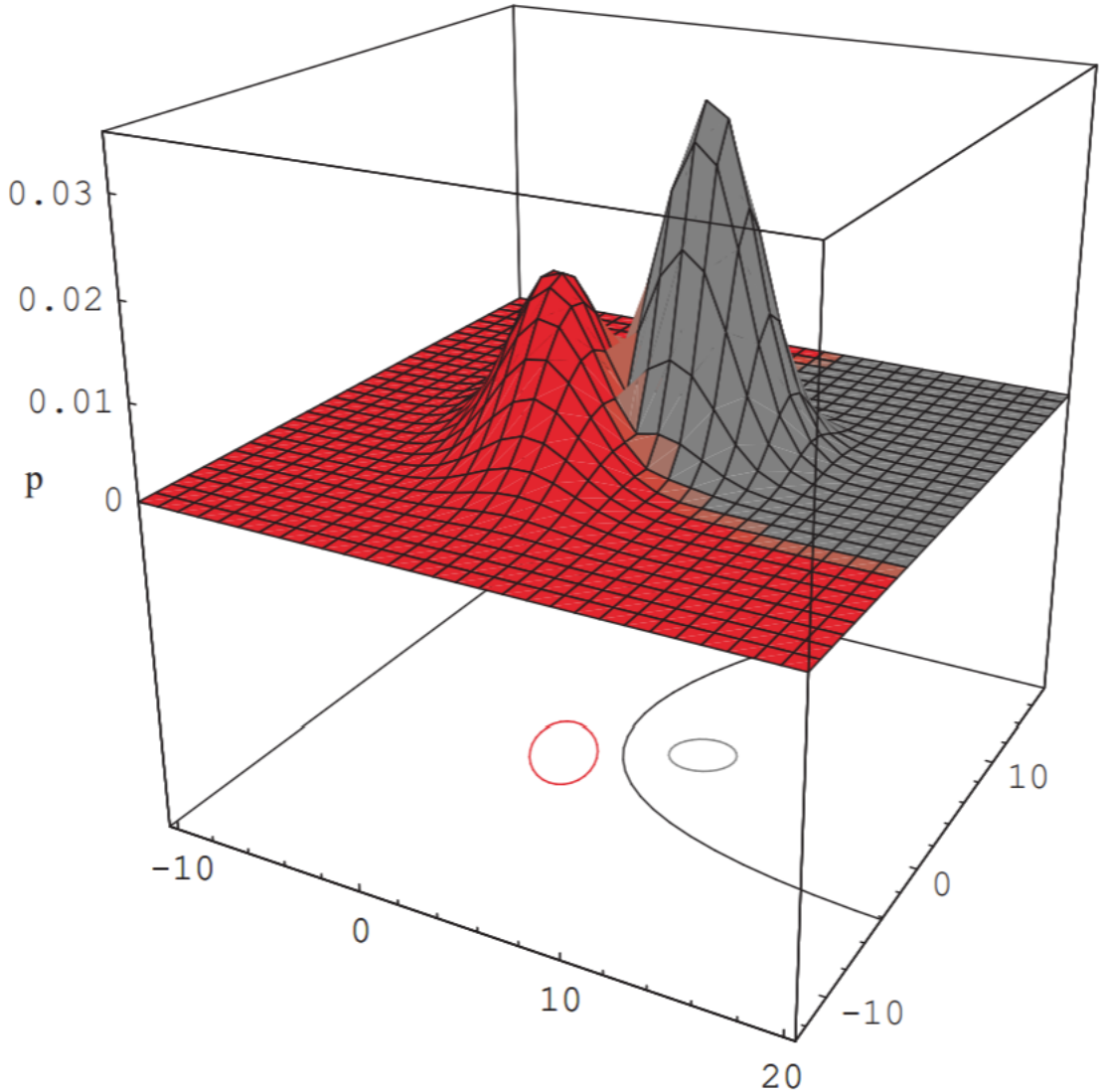


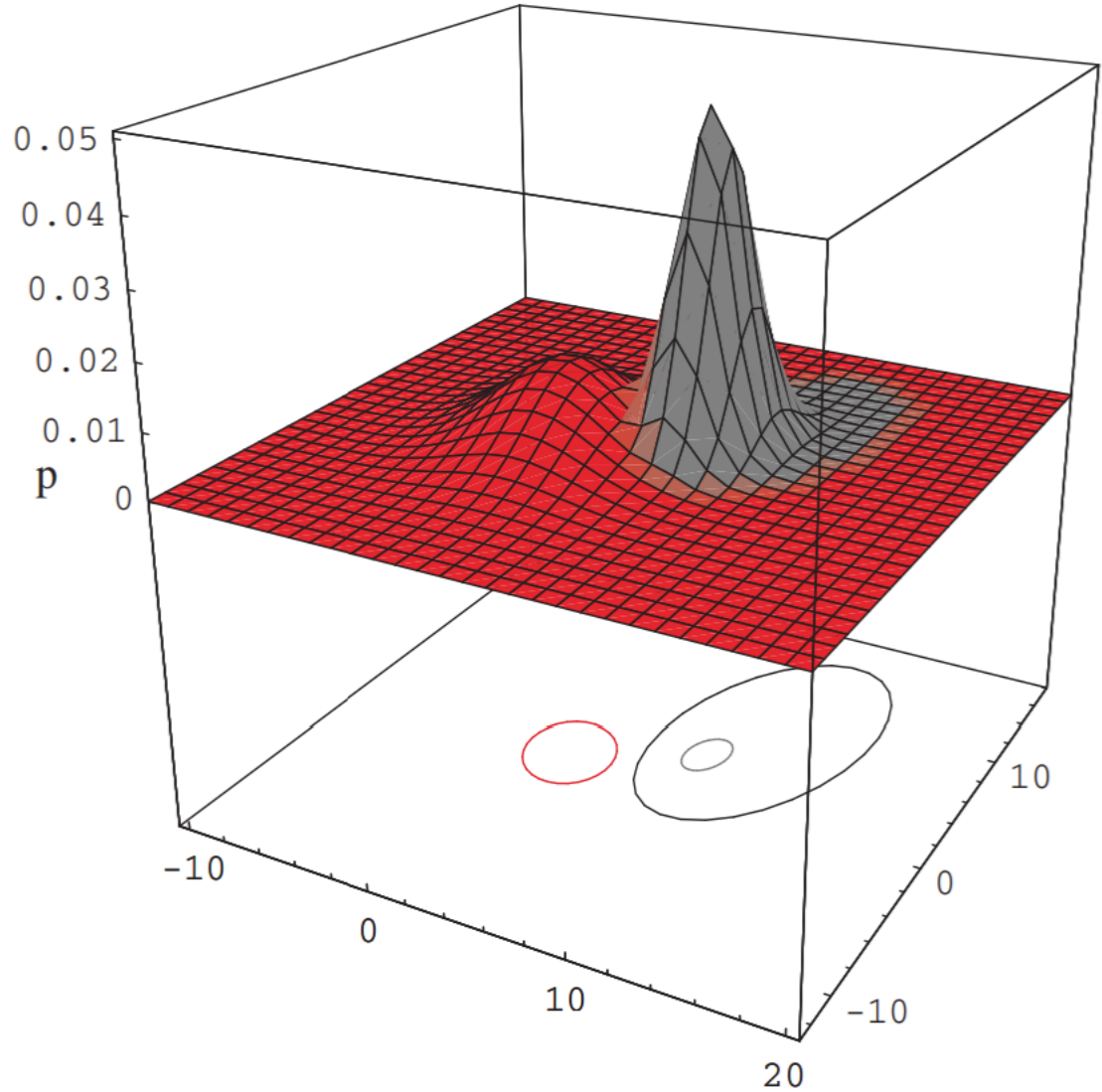


http://mi.eng.cam.ac.uk/~mjfg/local/4F10/lect2.pdf

Multivariate Gaussian – Applications

- Multivariate Gaussian for maximum-likelihood classification
  - Decision boundaries

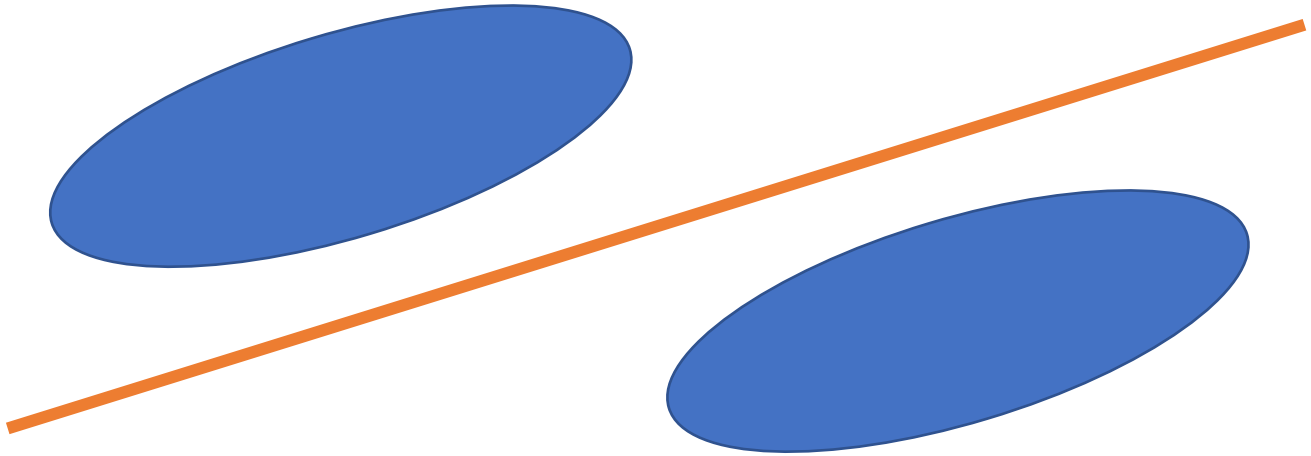




http://mi.eng.cam.ac.uk/~mjfg/local/4F10/lect2.pdf

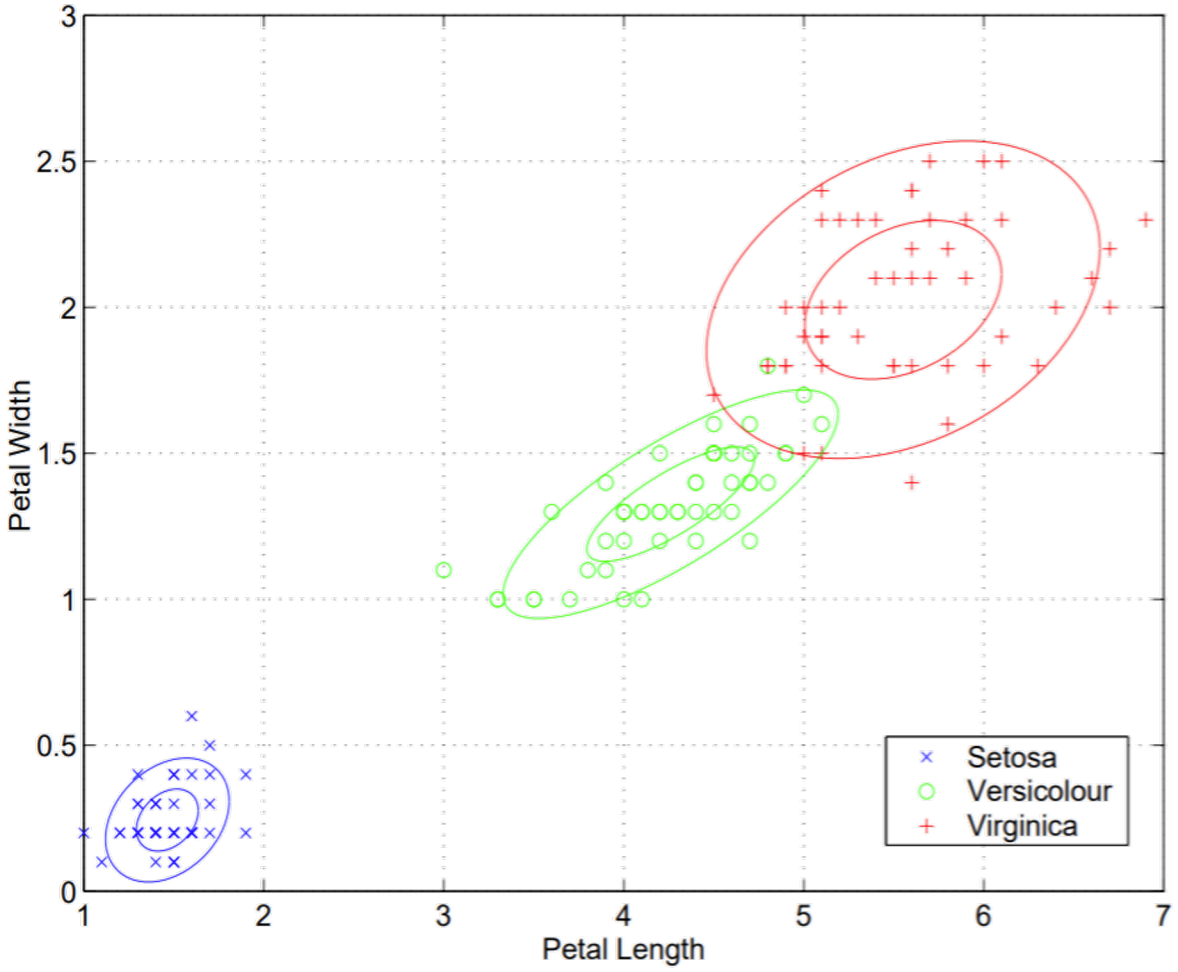
Multivariate Gaussian – Applications

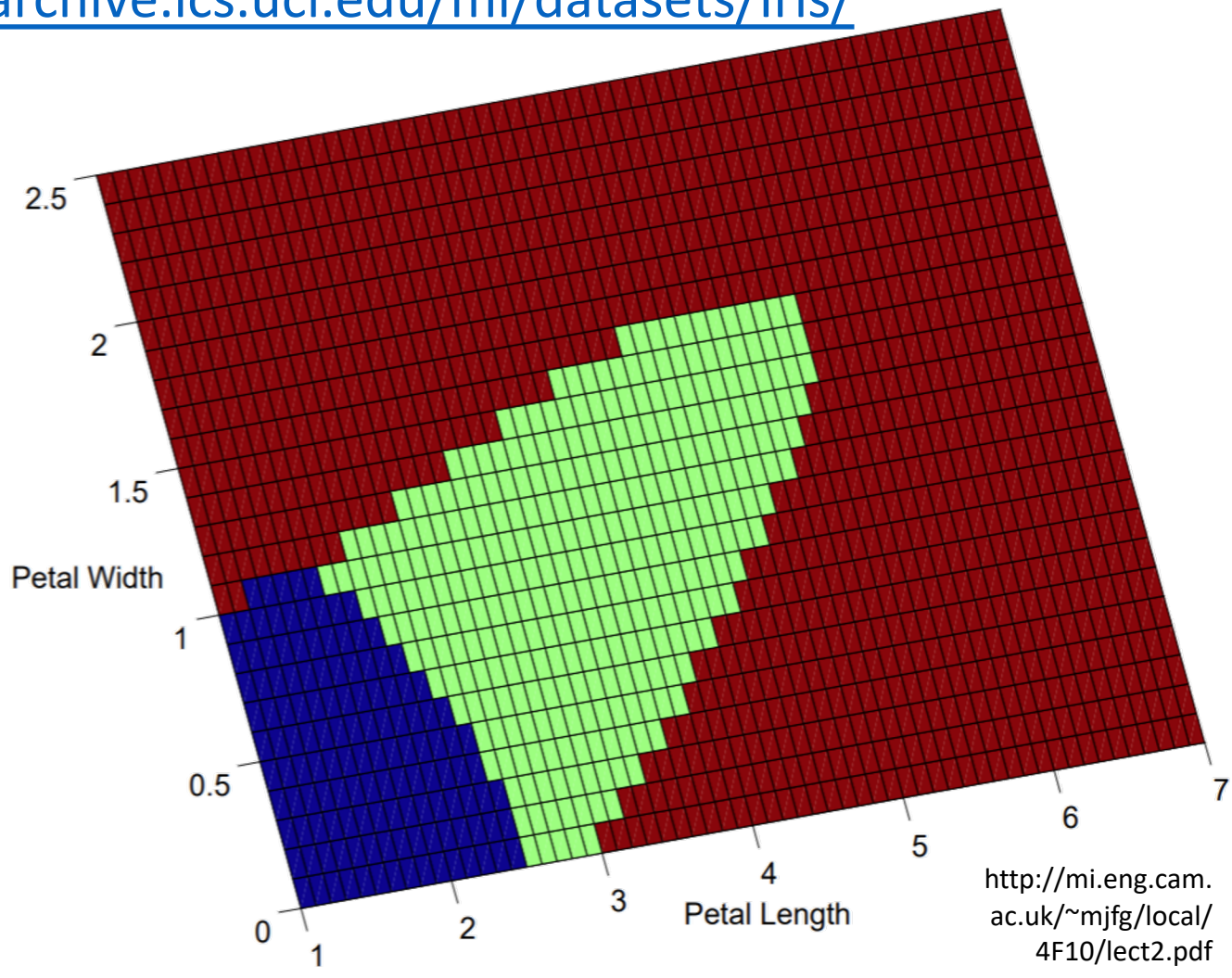
- Multivariate Gaussian for maximum-likelihood classification
  - Decision boundaries
  - When $C_1 = C_2 = C$, then decision boundary is:
    - $0 = \log (P(x|Class_1) / P(x|Class_2))$
      $=$
      $- 0.5 (x-m_1)^T C^{-1} (x-m_1)$
      $+ 0.5 (x-m_2)^T C^{-1} (x-m_2)$
    - 0
      $=$
      $+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





http://mi.eng.cam.ac.uk/~mjfg/local/4F10/lect2.pdf

Datasets

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

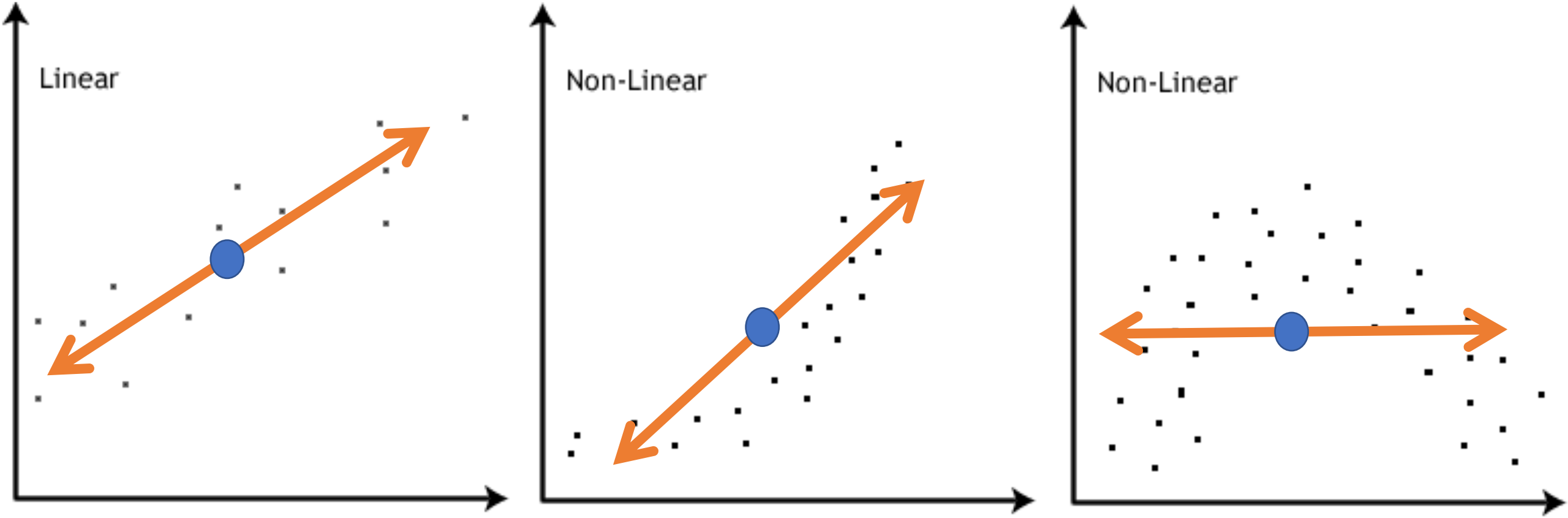
Principal Component Analysis (PCA)

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)

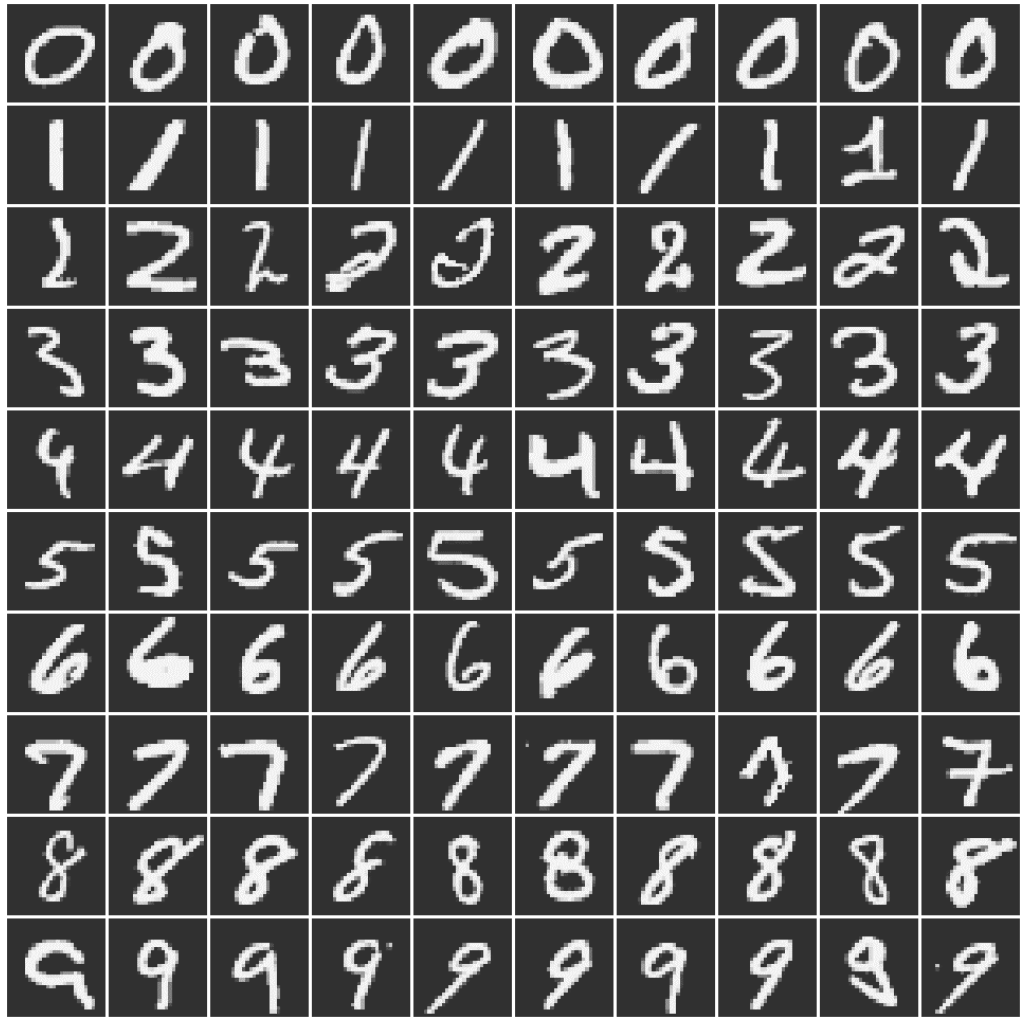
- Modes of variation
  - Set of vectors (directions and magnitudes) that are used to depict the variation in a population or sample, around the mean



https://statistics.laerd.com/spss-tutorials/pearsons-product-moment-correlation-using-spss-statistics.php

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



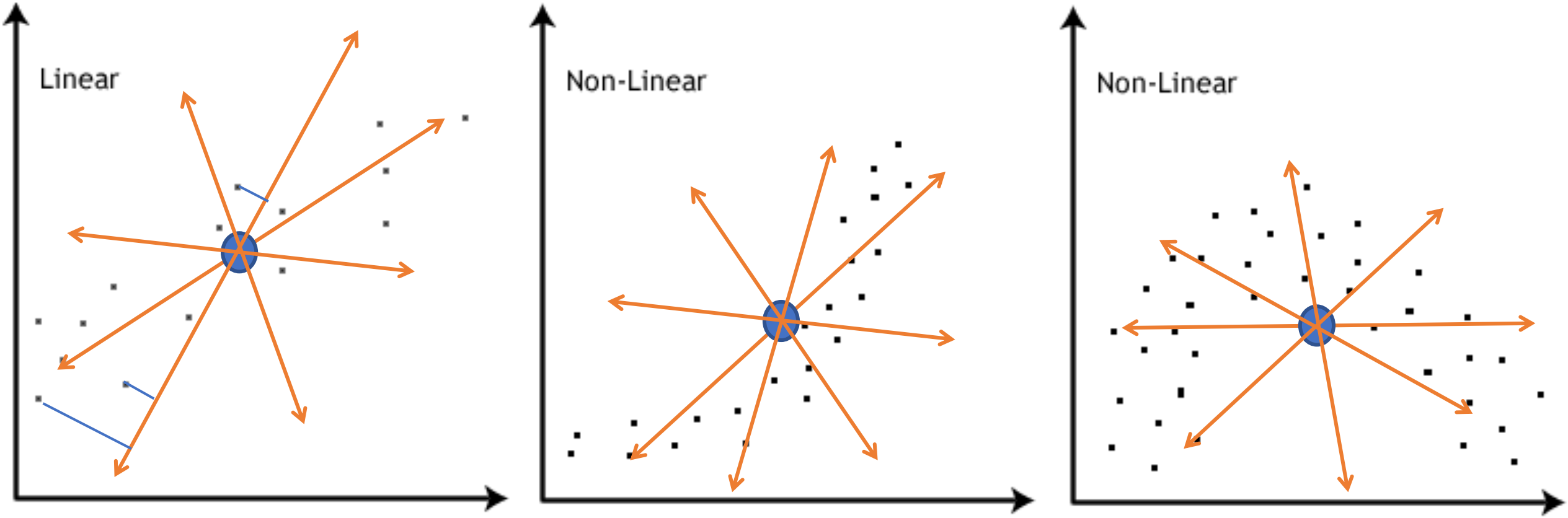
Principal Component Analysis (PCA)

- Directions of maximal variance
  - Consider a general multivariate random variable X with PDF P(X)
  - 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 $N \to \infty$ for the data, sample mean $\to \mu$, and sample covariance $\to C$

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)

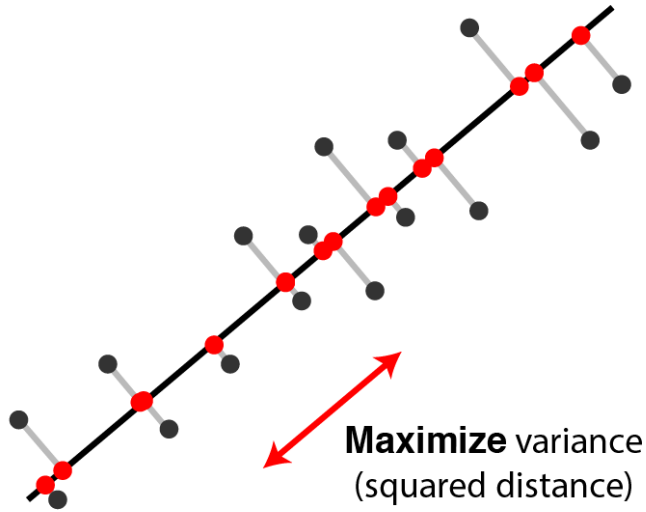
- Directions of maximal variance

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 $\sum_i \langle x_i, v\rangle v = 0$



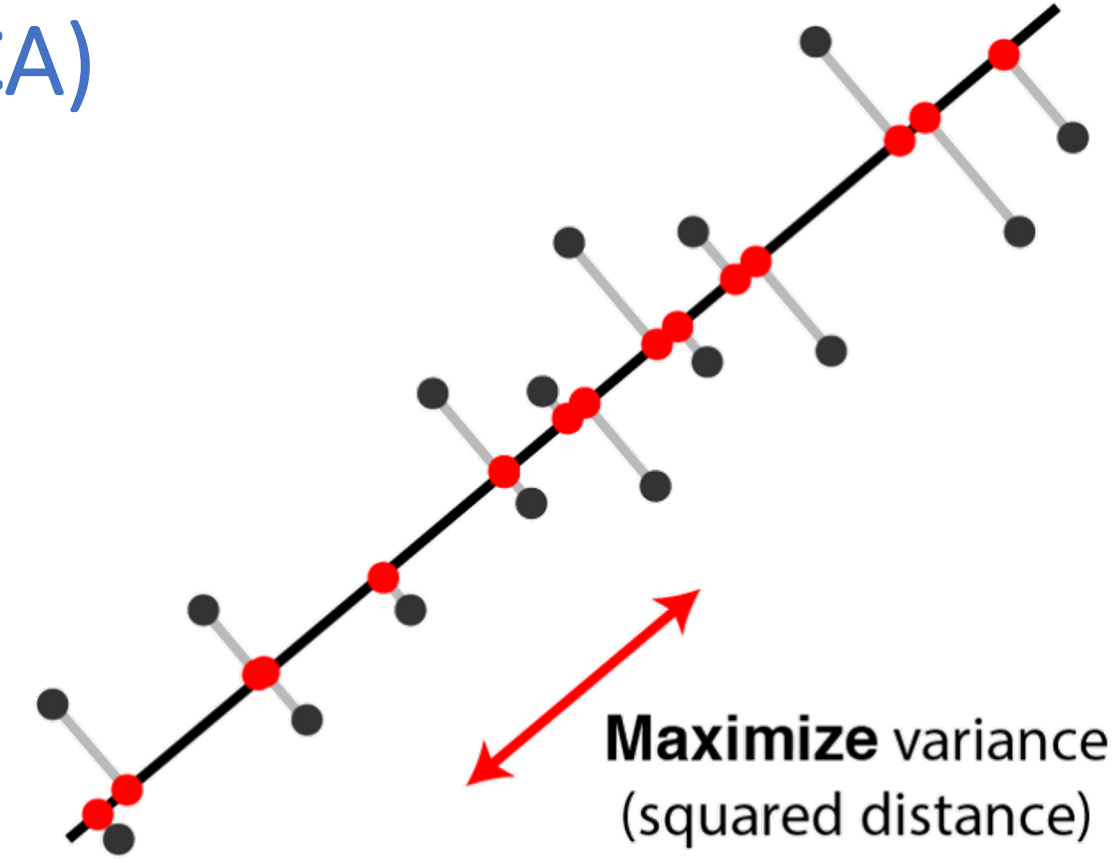
Maximize variance (squared distance) of red dots in this direction

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



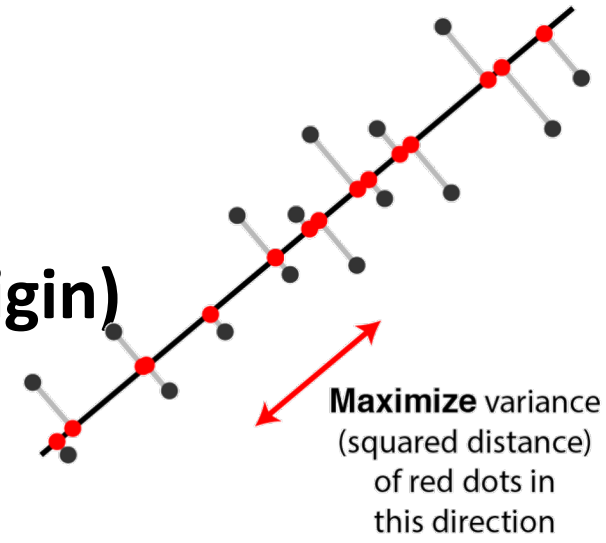
Maximize variance (squared distance) of red dots in this direction

Optimal direction

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

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., $C_{11} \geq C_{22} \geq \cdots$

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

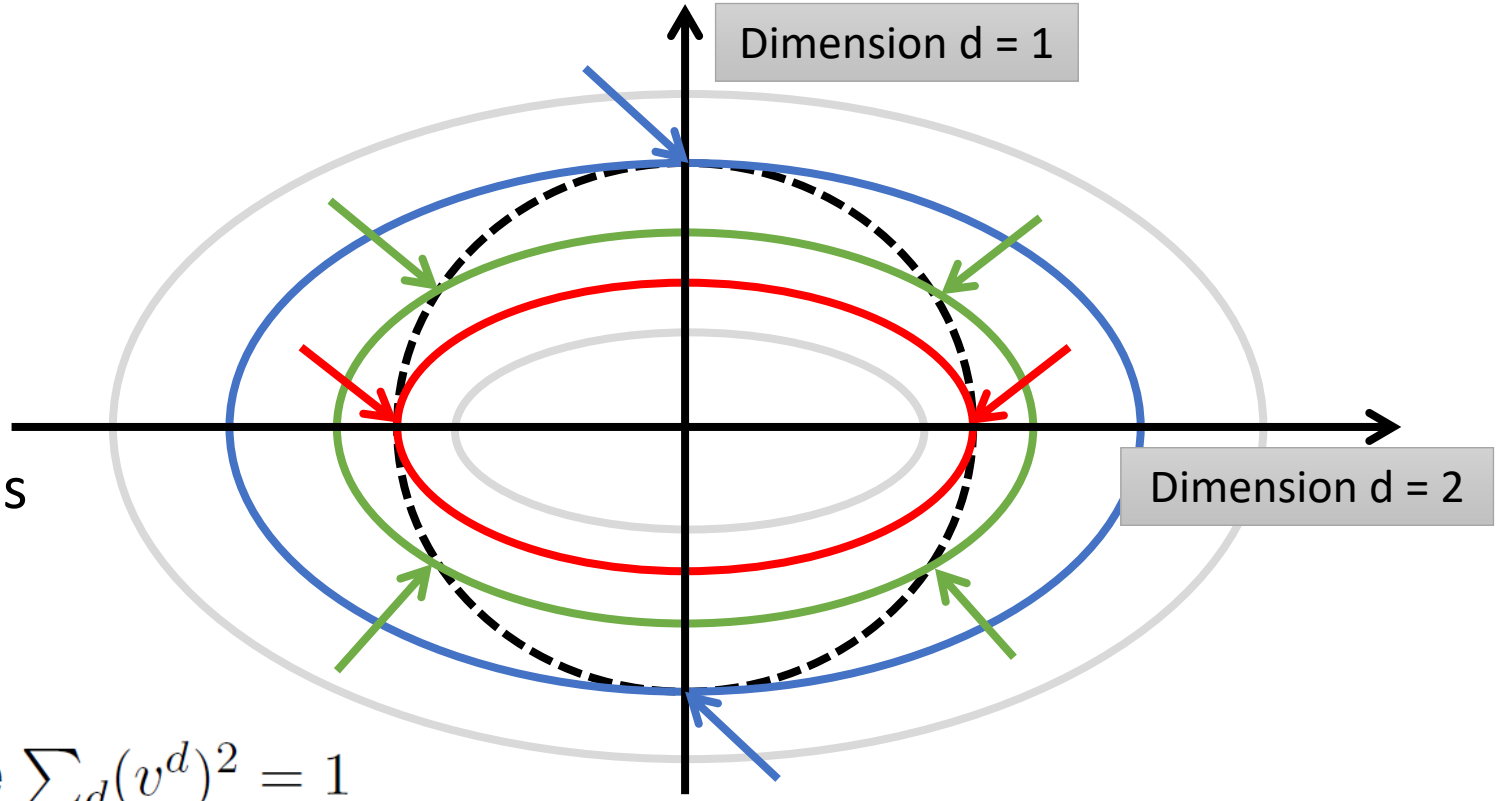
The unit-norm constraint on v is: $\sum_d (v^d)^2 = 1$

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

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

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 $1/\sqrt{C_{dd}}$

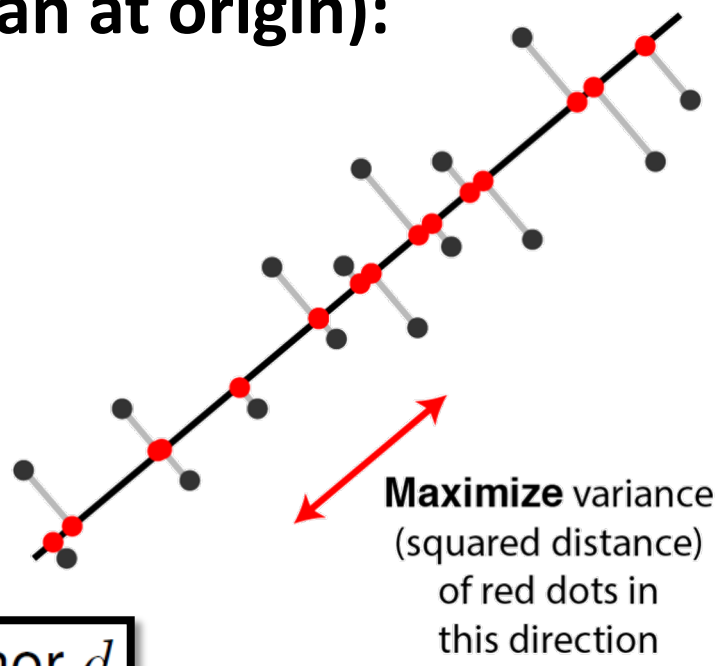
Principal Component Analysis (PCA)

- 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
$= \arg\max_{u:\|u\|_2=1, u\perp v} \sum_i \langle x_i, u\rangle^2$
$= \arg\max_{u:\|u\|_2=1, u\perp v} \sum_d C_{dd}(u^d)^2$, 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, $\max_v v^\top C v = \max_v v^\top Q\Lambda Q^\top v = \max_{u:=Q^\top v} u^\top \Lambda u$

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

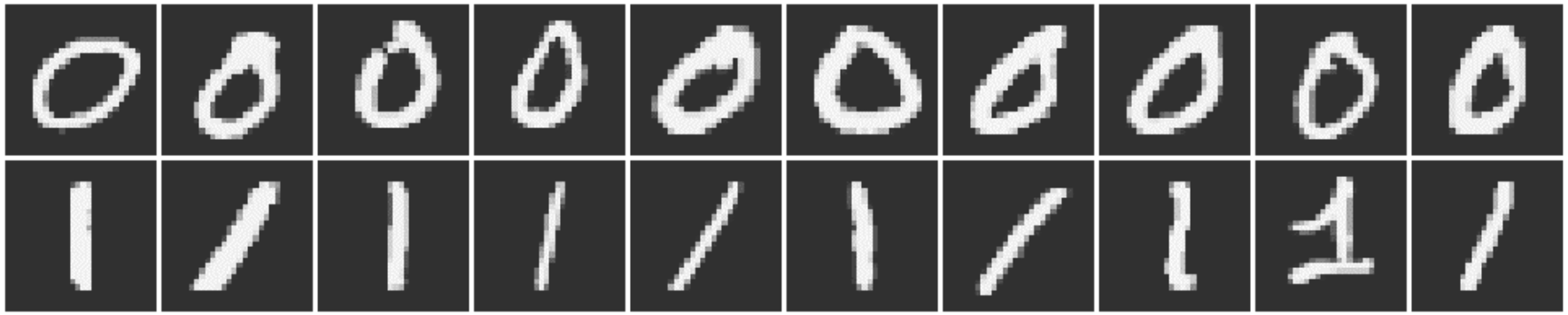
that is equivalent to $v = Qu =$ first column of Q (in the original basis)

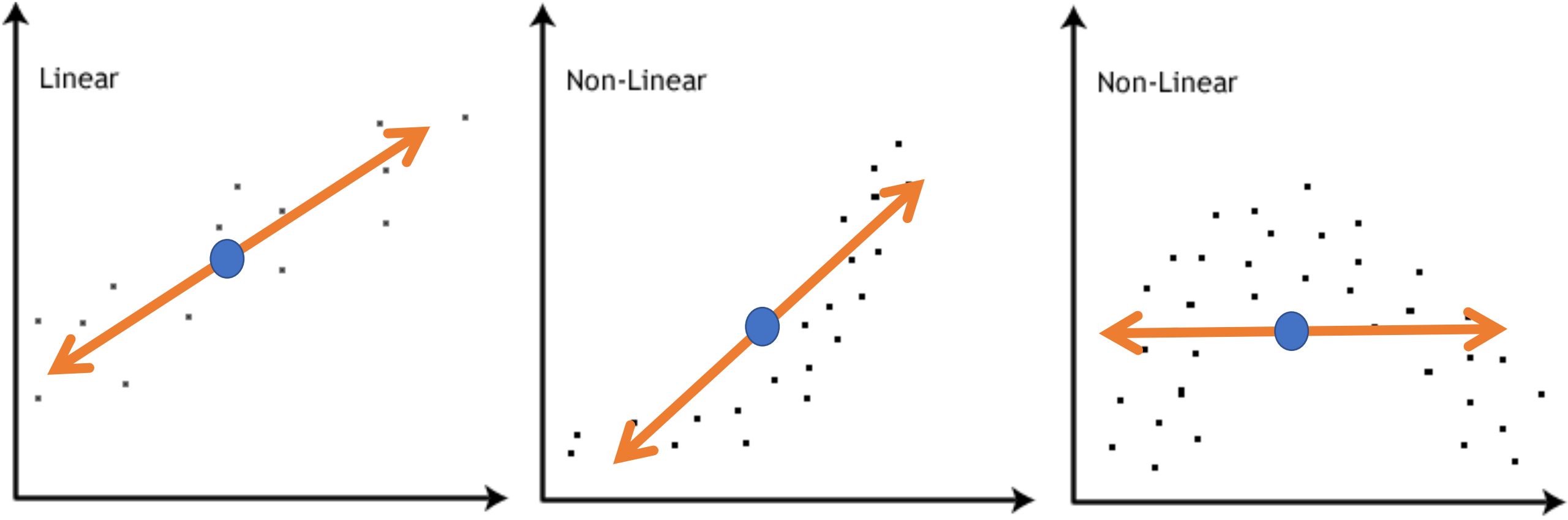
The maximal variance then is Λ_{11} that is the eigenvalue corresponding to the principal eigenvector

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

Principal Component Analysis (PCA)

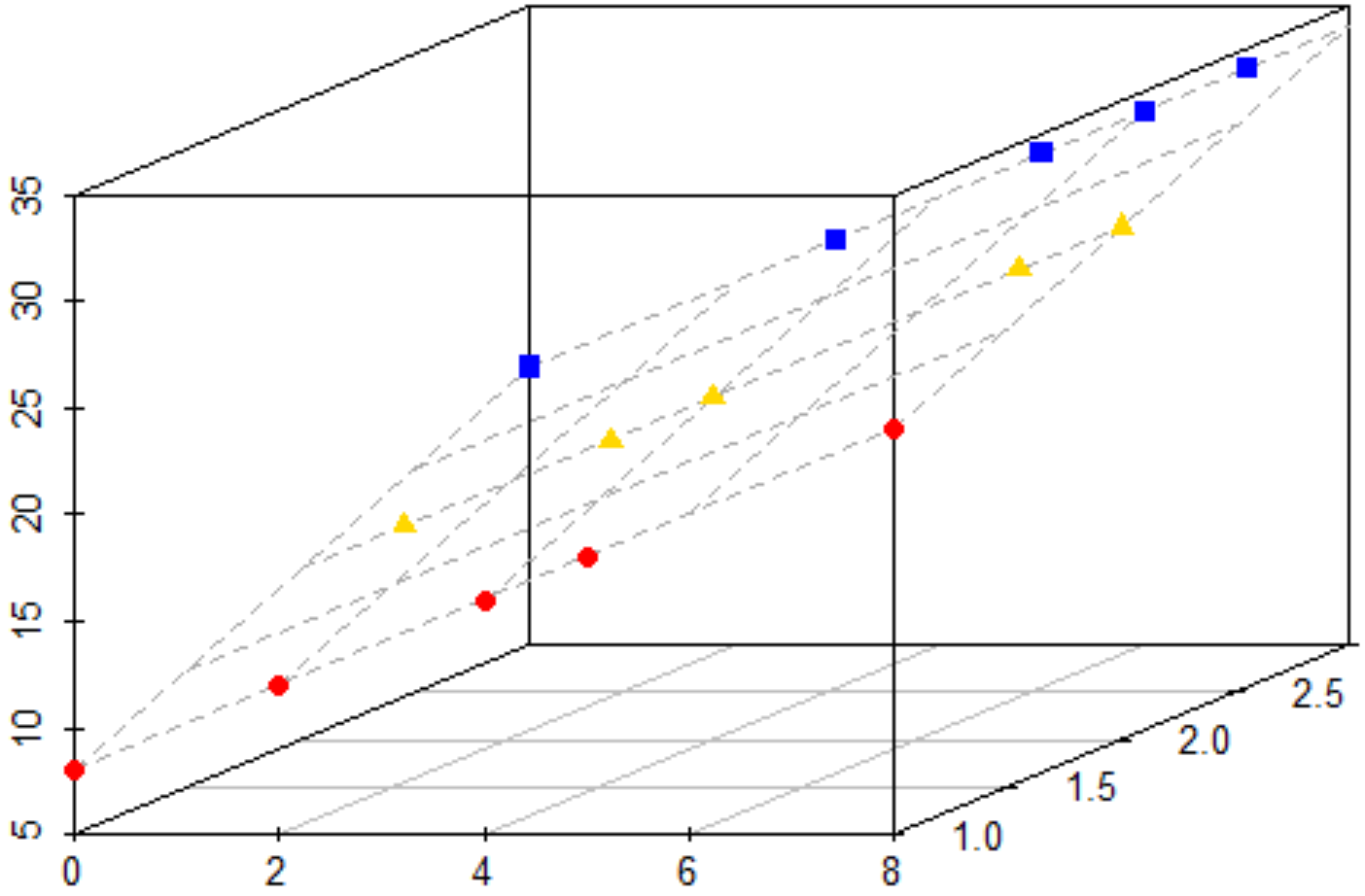
- Directions of maximal variance





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

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

Optimal direction

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

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 $(v_1^d)^2 + (v_2^d)^2 + \cdots + (v_D^d)^2 = 1$ (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 the optimization problem looks very similar to the one we had solved before

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

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 $C_{11} \geq C_{22} \geq \cdots$)

Why ? Proof by contradiction:
Suppose $\exists$ a solution (meeting all constraints) with some $a^1 + a^2 = 2 - \delta$ where $0 < \delta \leq 2$
Then, $\sum_{d>2}^{D} a^d = \delta$ and objective function value is $\sum_{d=1}^{D} C_{dd} a^d$
Then, for some $d > 2$, we can reducing a^d to zero and increase $a^1 + a^2$ to increase (not decrease) the objective function
We can repeat this procedure for all $d > 2$ until $a^1 + a^2 = 1$ and all other $a^d = 0$

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):**

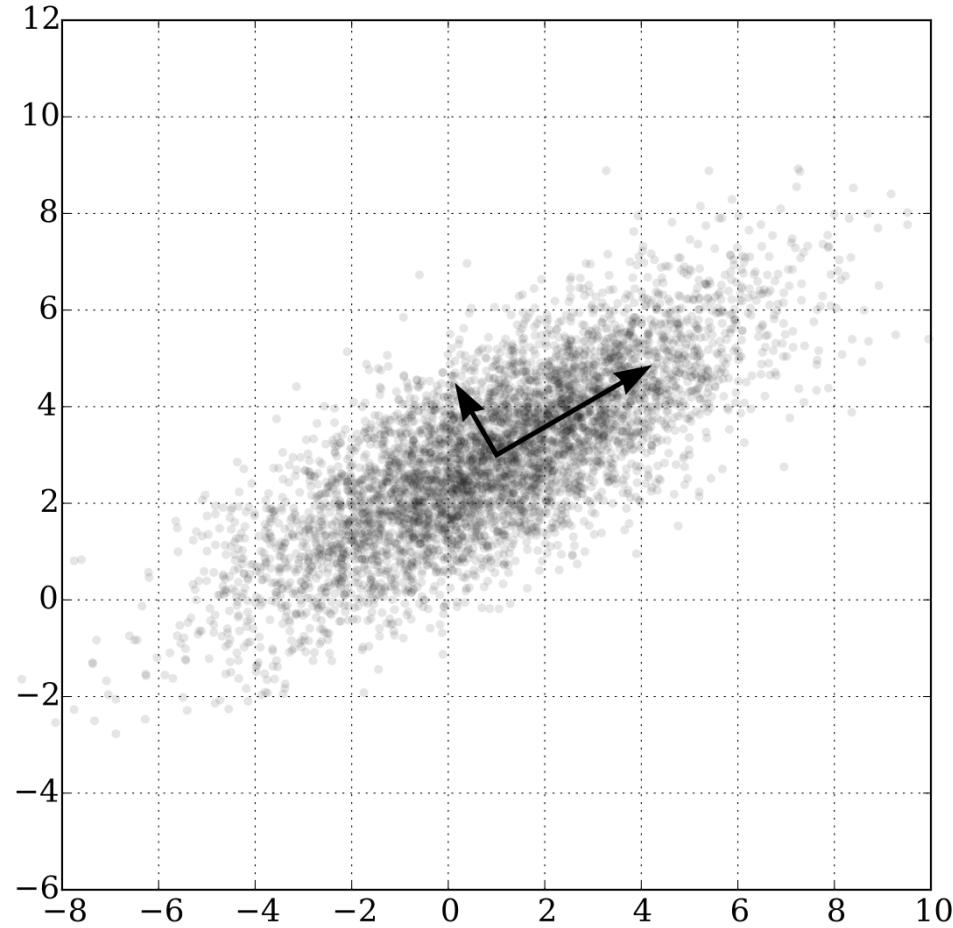
The solution for this problem is $a^1 = a^2 = 1$ and all other $a^d = 0$ (because $C_{11} \geq C_{22} \geq \cdots$)

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

  - 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

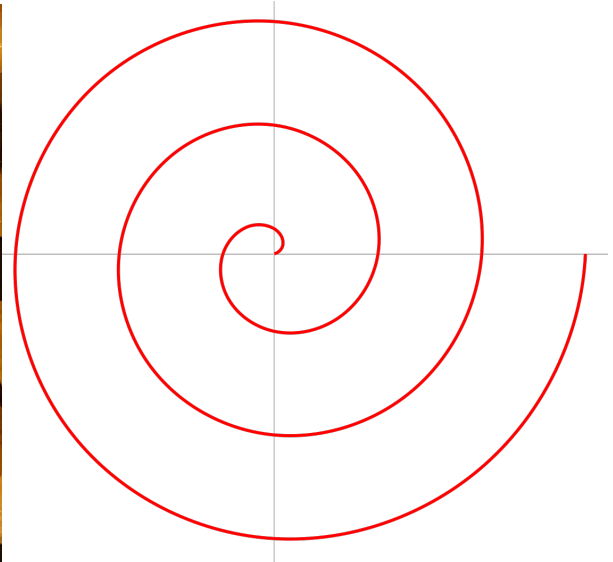
Principal Component Analysis (PCA)

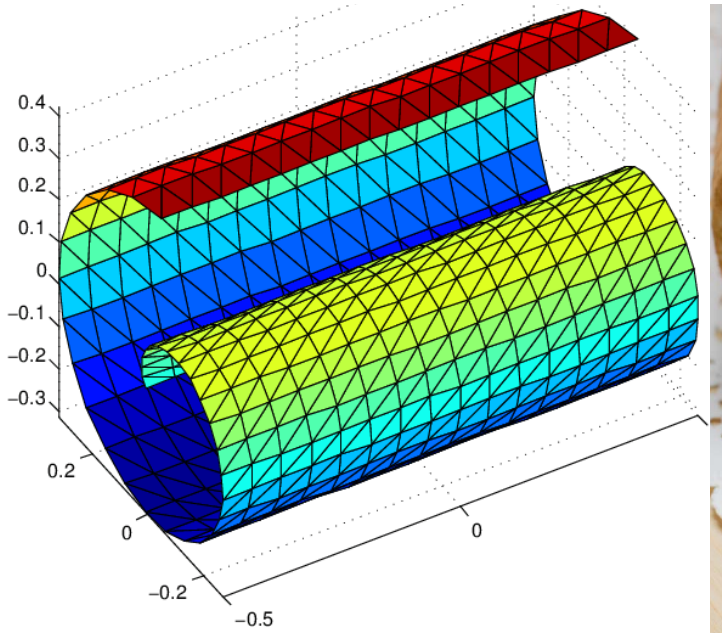
- PCA applied to data from a multivariate Gaussian distribution
  - **Consider X is multivariate Gaussian**
  - If X := AW + b, then:
    - Principal modes of variation are **directions** given by **eigenvectors** of covariance matrix $C := AA^T$
    - Principal modes of variation are along axes of hyper-ellipsoids 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 $\mathbf{S^2}$



Principal Component Analysis (PCA)

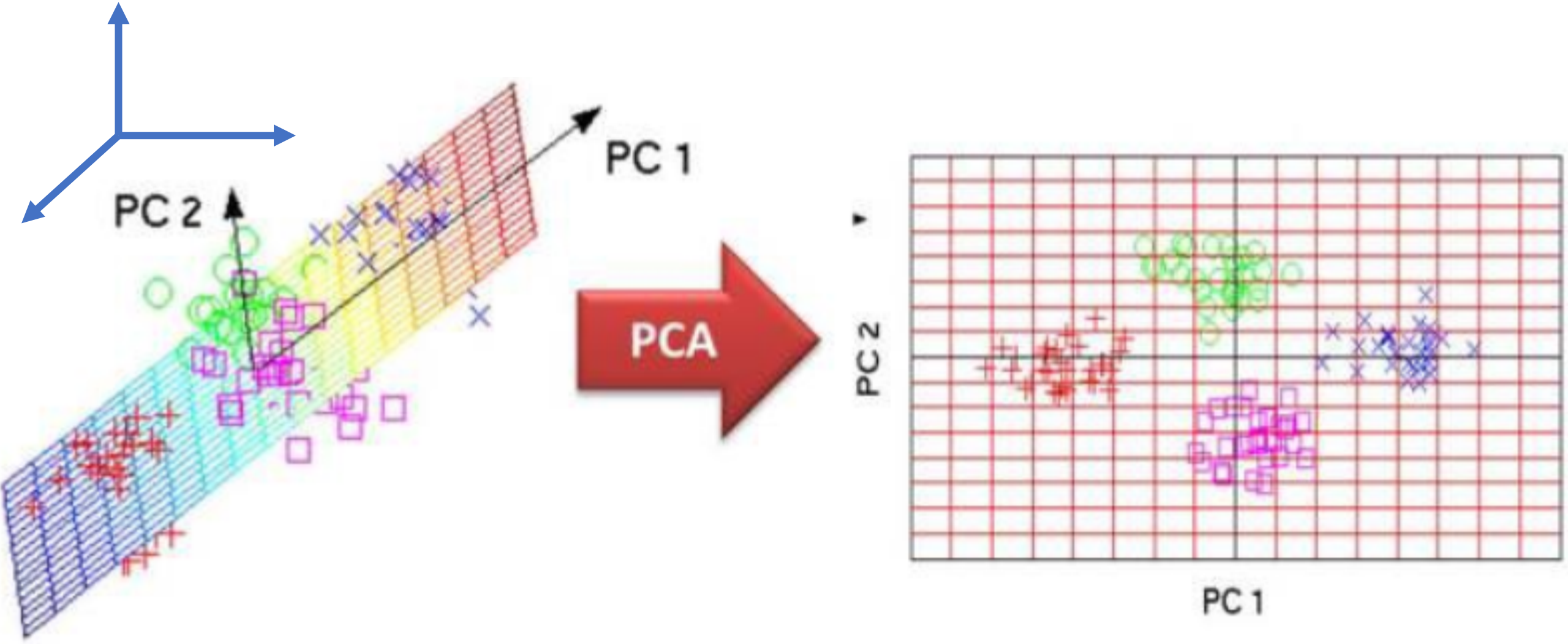
- Applications: **Dimensionality reduction**
  - Intrinsic dimension: Minimum number of variables (degrees of freedom) required to represent the signal
    - Consider a multivariate random vector X of N scalar variables: $x = (x_1, \ldots, x_N)$
    - Consider a function g(.), and M<N scalar variables $a_1, \ldots, a_M$ such that every x~P(X) can be written as $x = g(a_1, \ldots, a_M)$ for some $a_1, \ldots, a_M$, 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



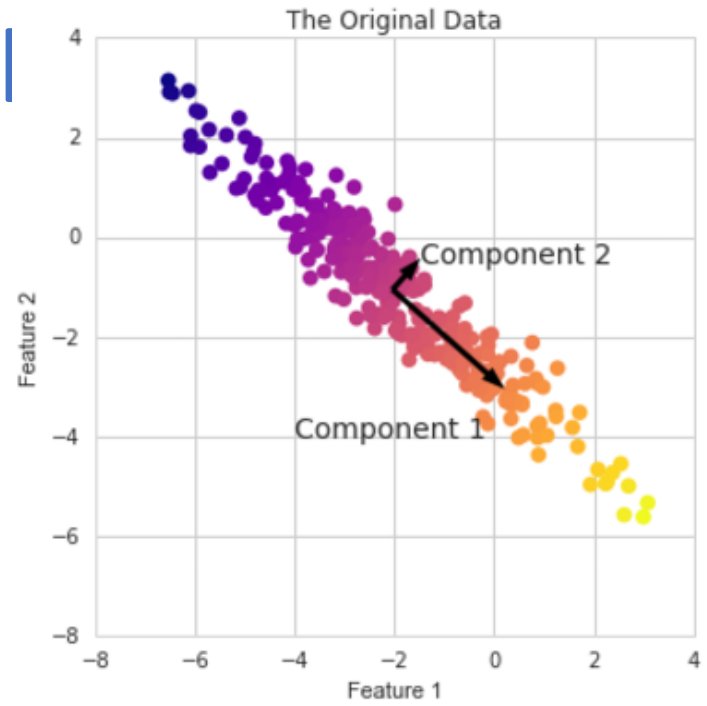
https://medium.com/analytics-vidhya/dimensionality-reduction-using-principal-component-analysis-pca-41e364615766

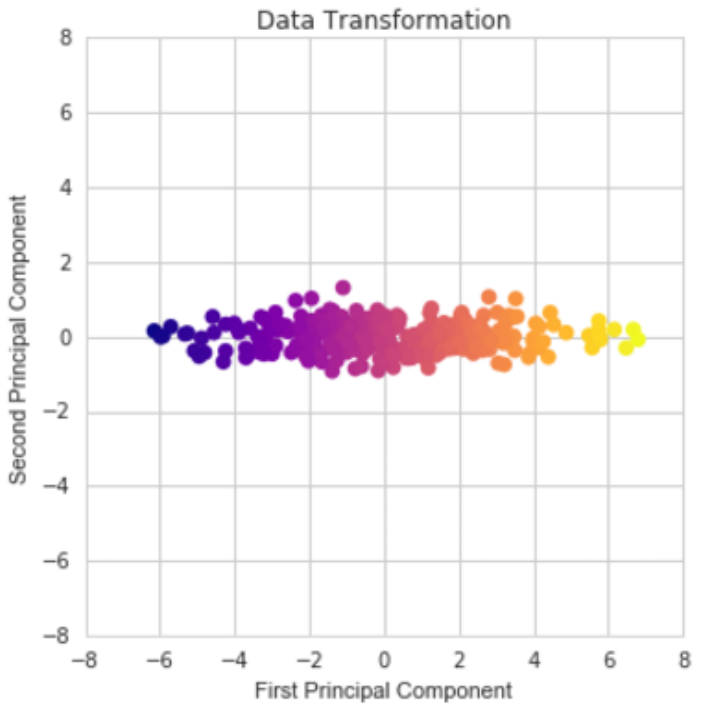
Principal Component Analysis (PCA)

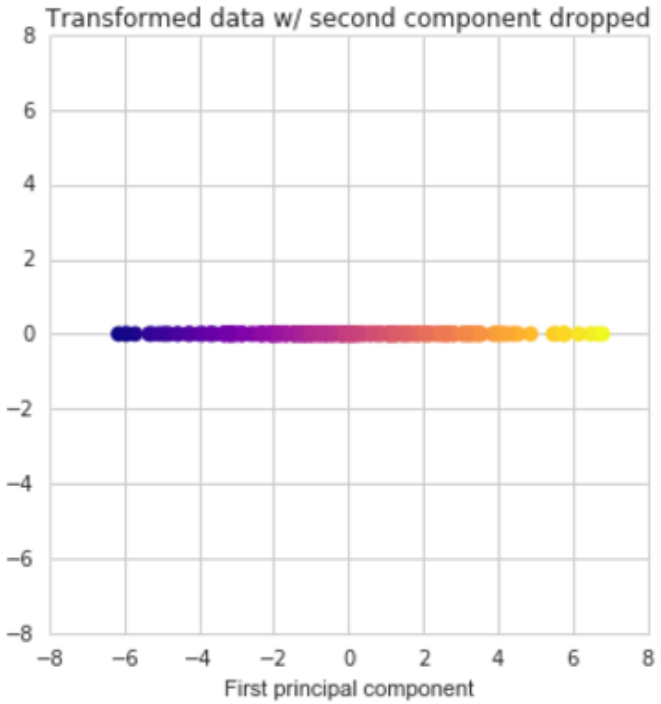
- Applications: Dimensionality reduction
  - Acquired data is corrupted with errors
    - e.g., measurement errors
    - Such errors make the signal representation seem to of a dimension higher than intrinsic dimension
  - Dimensionality reduction:
    Transformation of data from a high-D space into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

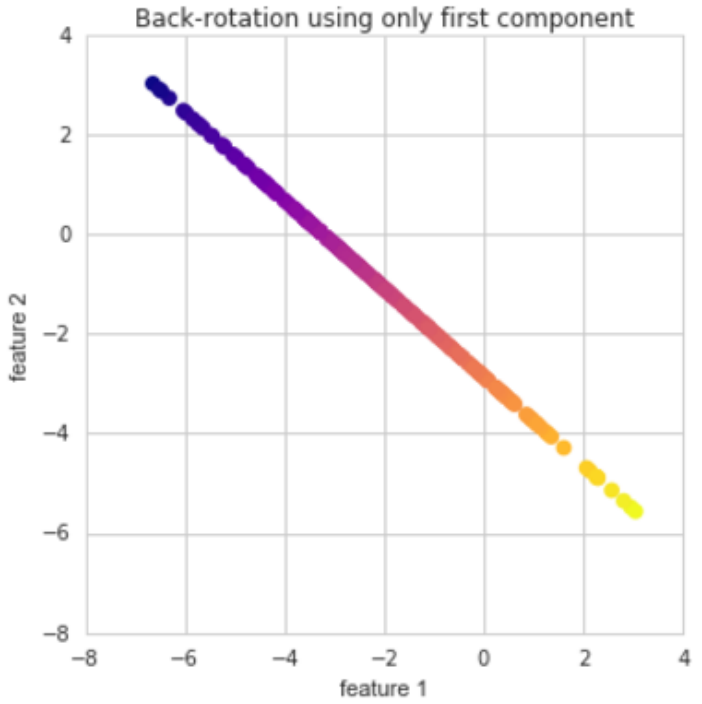
Principal Component Anal

- 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





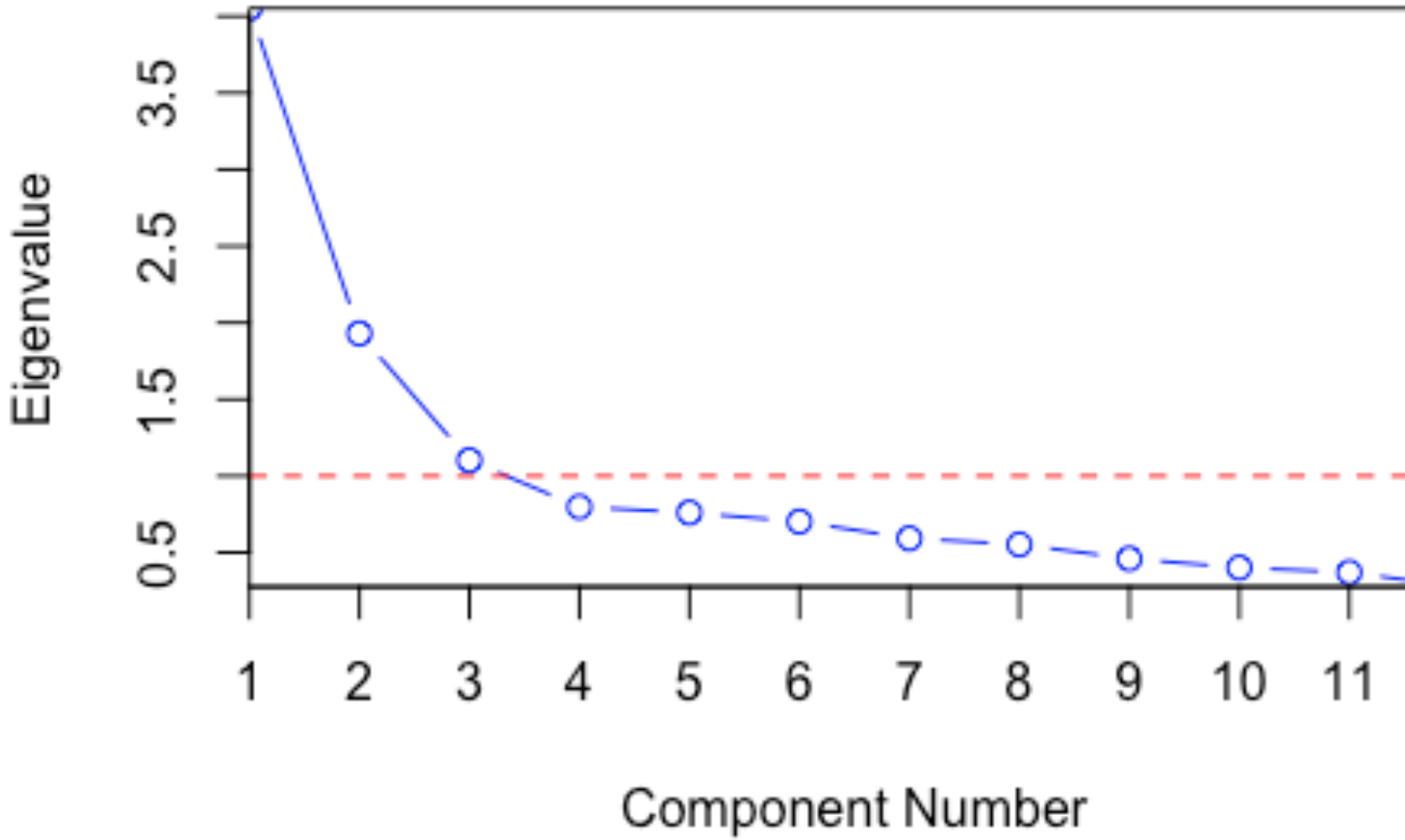




http://bennymachinelearning.blogspot.com/2017/08/machine-learning-principal-component.html

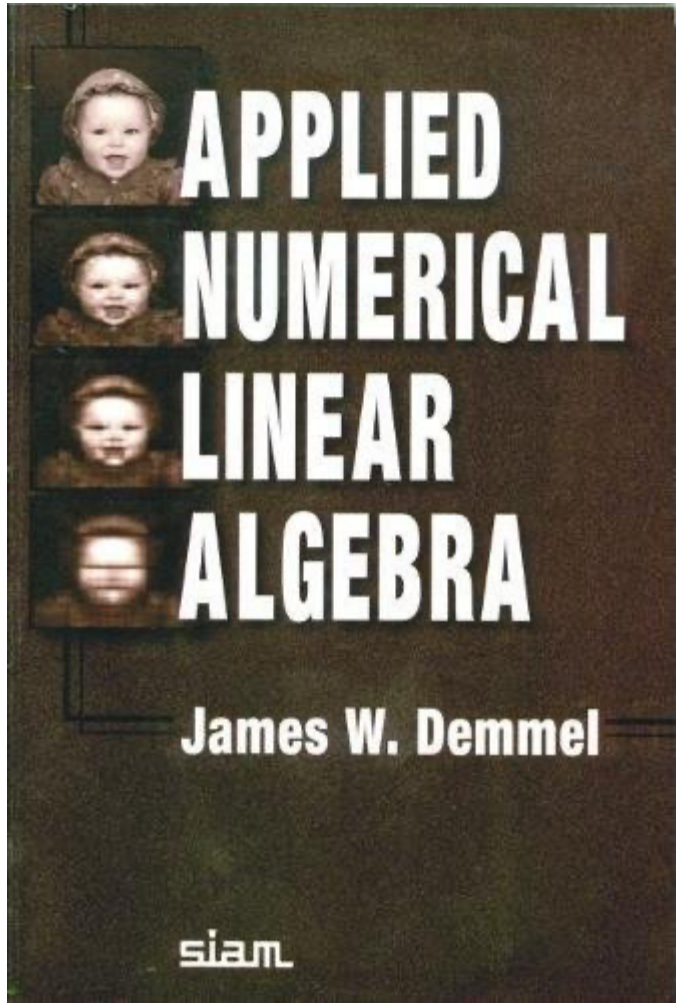
Principal Component Analysis (PCA)

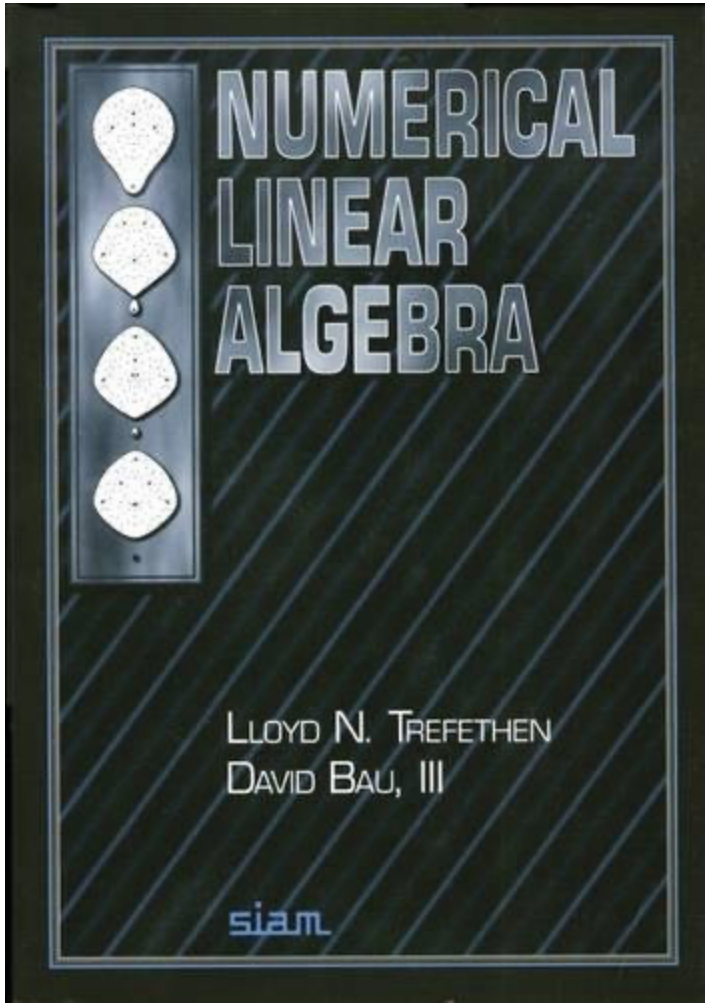
- Applications: Dimensionality reduction
  - Using PCA
  - X may be N dimensional
  - PCA can find an M-dimensional space (often when M << N) that captures most of the variability (total dispersion) in the data



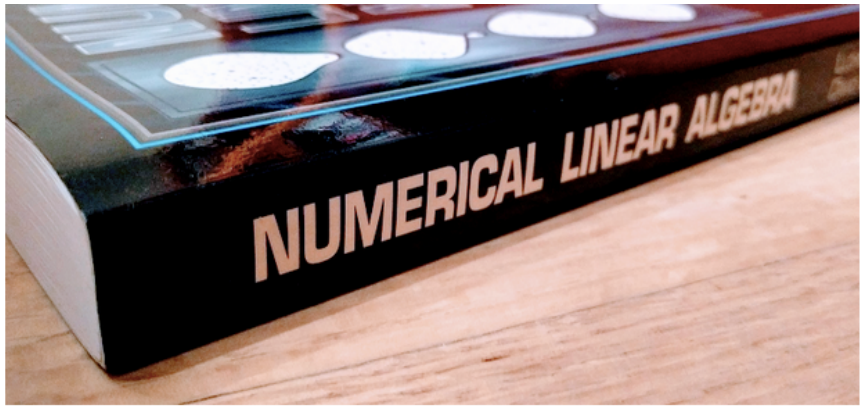
Singular Value Decomposition (SVD)

- 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)

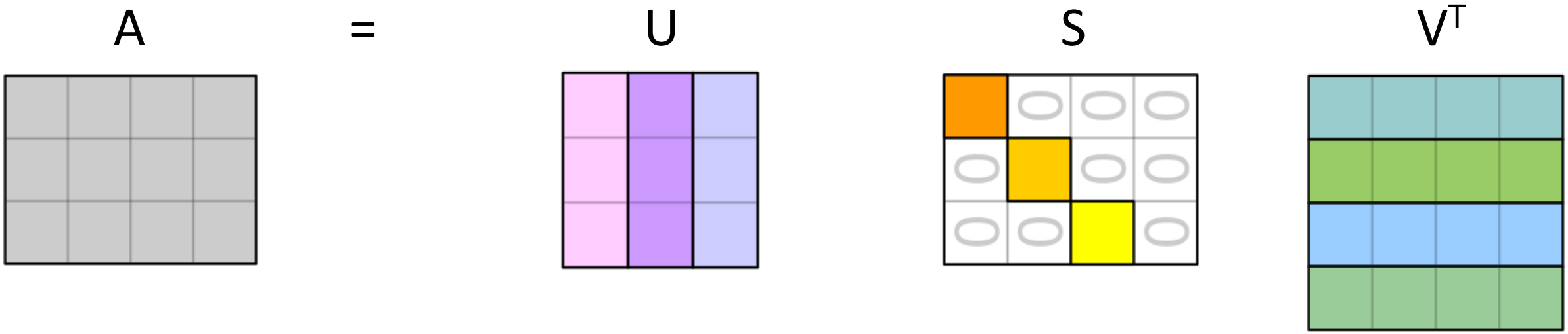
- Matrix factorization
- Let matrix A be size MxN
- When A is real valued, then SVD of A = U S V^T, where:
  - V is orthogonal of size NxN
    - When A is complex: V is unitary
  - U is orthogonal of size MxM
    - When A is complex: U is unitary
  - S is (rectangular) diagonal with size MxN
    - Values 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 $M \leq N$, $A = \sum_{m=1}^{M} s_m u_m v_m^\top$

Singular Value Decomposition (SVD)

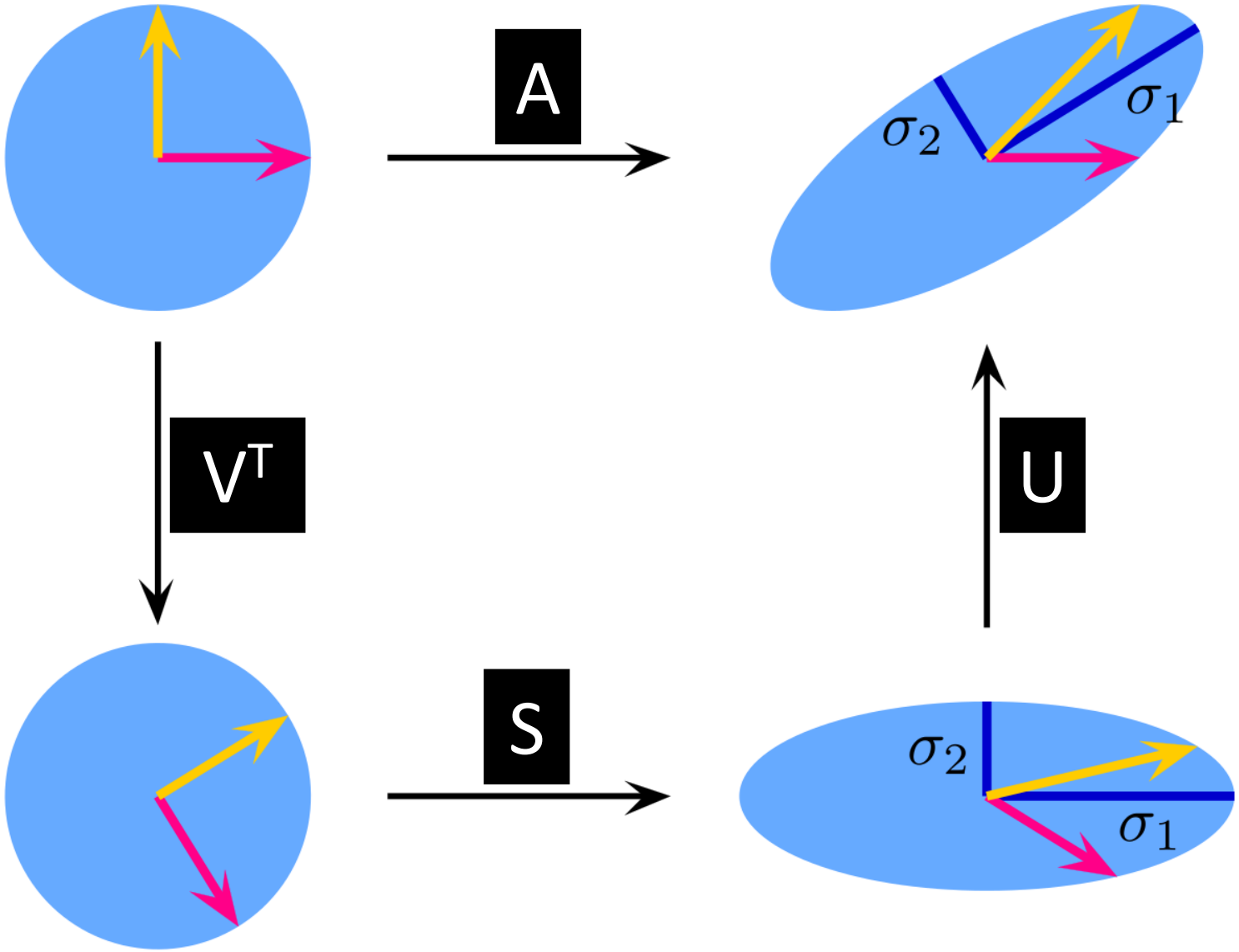
- $A = U\ S\ V^T$
- The matrices in pictures

A = U S V^T



Singular Value Decomposition (SVD)

- Geometric interpretation of the action of a matrix A on a vector



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: Снáйперская Винтóвка 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.

Main page
Contents
Current events
Random article
About Wikipedia
Contact us
Donate

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 $\|A\|_2 := \max_{x \neq \mathbf{0}}(\|Ax\|_2 / \|x\|_2) \geq 0$

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

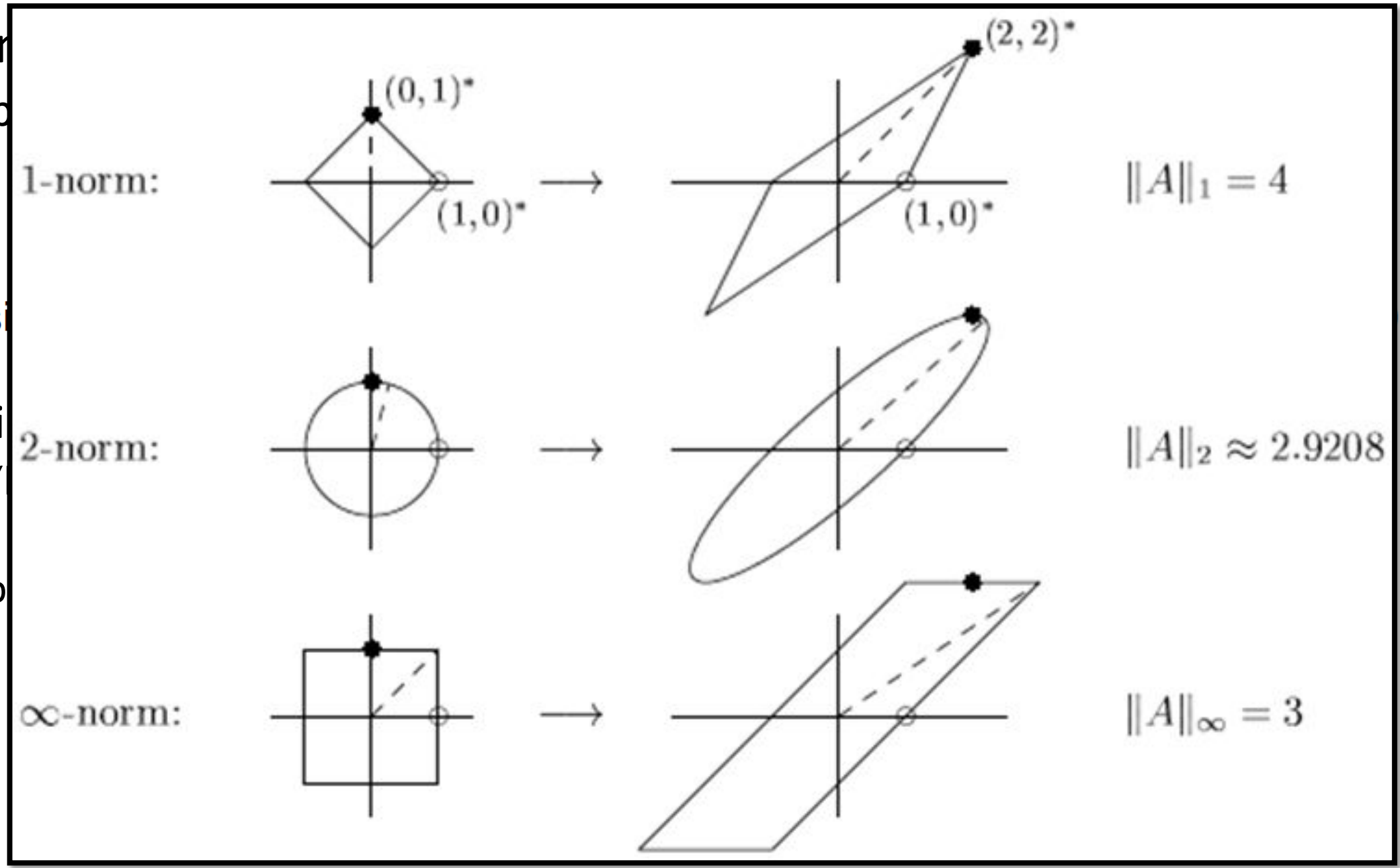
Singular Value Decomposition (SVD)

- Matrix norm
  - Induced b

For a vector $x \in$

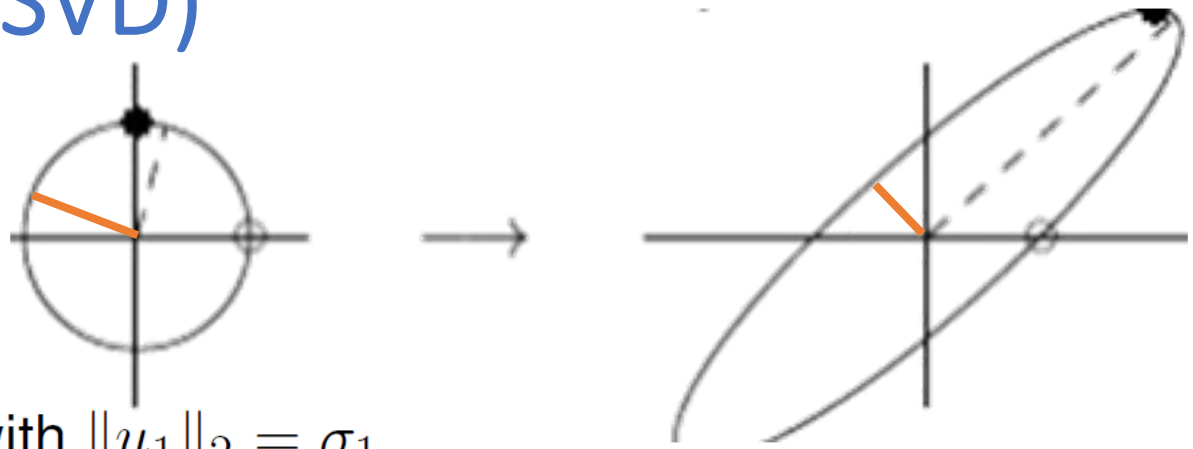
For matrix A of si

  - Geometri
    - Apply "
    - Let y :=
    - Then, p



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

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

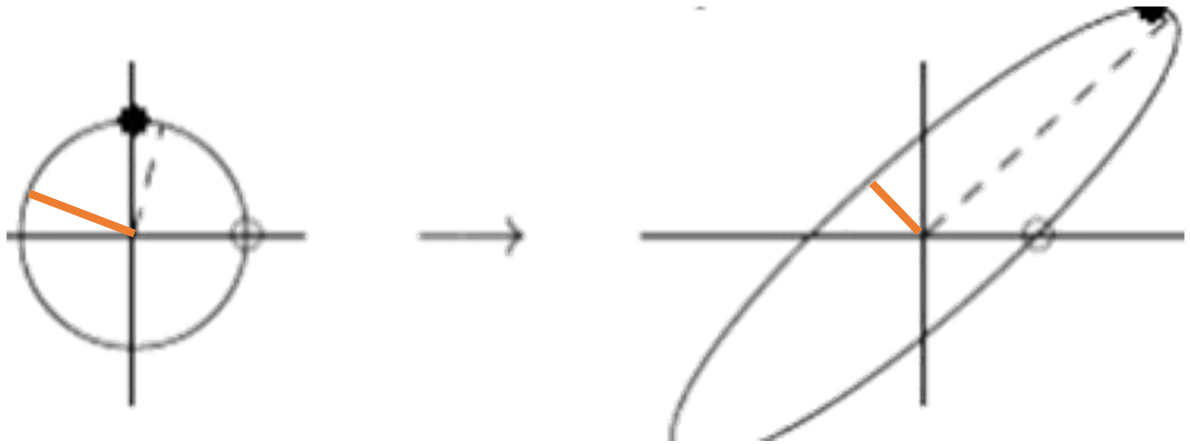
Because U and V are orthogonal matrices, we get $\|S\|_2 = \|A\|_2 = \sigma_1$

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, $U^\top AV = S = \begin{bmatrix} \sigma_1 & \mathbf{w}^\top \\ \mathbf{0} & B \end{bmatrix}$ 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
$S = I_{M\times M}[\sigma_1, \mathbf{0}^\top]^\top_{M\times 1} I_{1\times 1}$ or $S = I_{1\times 1}[\sigma_1, \mathbf{0}^\top]_{1\times N} I_{N\times N}$

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

For our original matrix A, by the induction hypothesis,

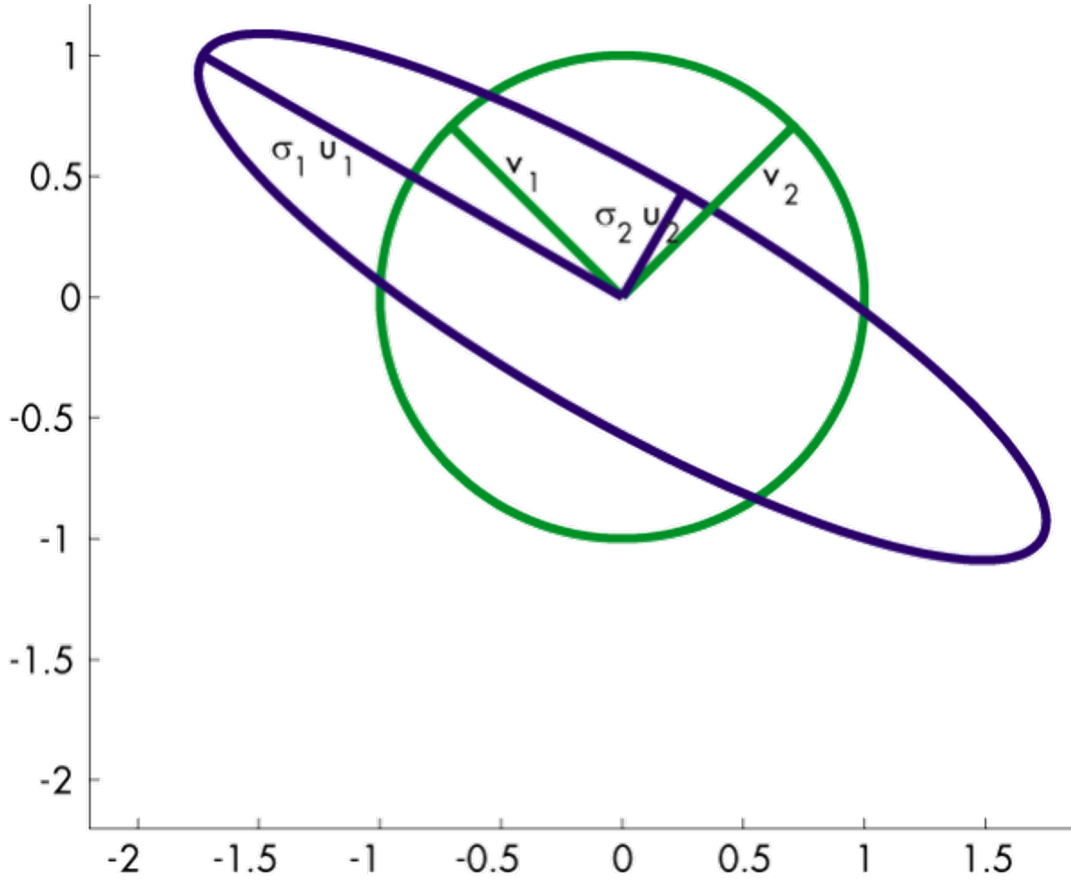
the remaining submatrix B has a factorization of the form $EDF^\top$

Then,

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

Singular Value Decomposition (SVD)

- Existence
  - 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)
  - Av_2
    $= USV^T v_2$
    $= U S [0\ 1\ 0 \dots 0]^T$
    $= U [0\ S_{22}\ 0 \dots 0]^T$
    $= S_{22}\ u_2$
  - Thus, Av_1 is along u_1 , and, hence, orthogonal to all other columns of U
  - Thus, Av_2 is along u_2 , and, hence, orthogonal to all other columns of U
  - …



Singular Value Decomposition (SVD)

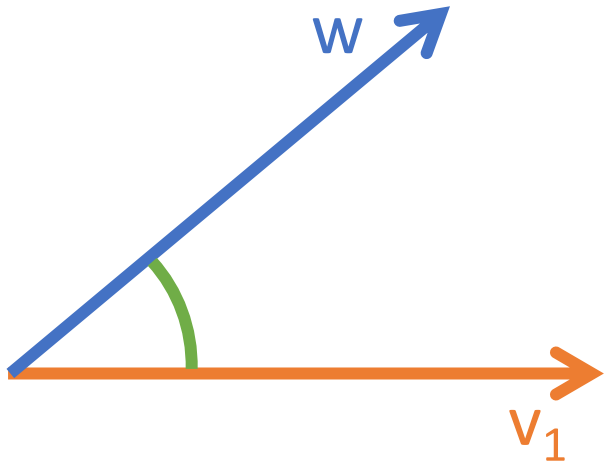
- Properties of singular values, vectors

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) ?

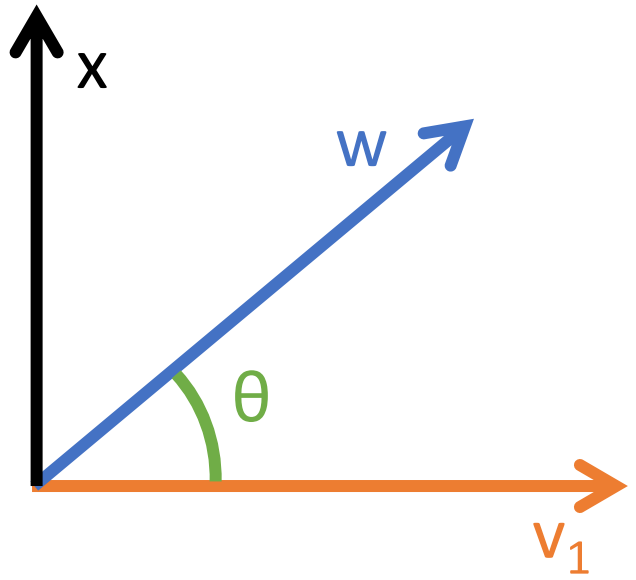
Suppose $\exists$ a unit-norm vector w, linearly independent with v_1, for which also $\|Aw\|_2 = \sigma_1$

What happens then ?



Singular Value Decomposition (SVD)

- Properties of singular values, vectors



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

What happens then ?

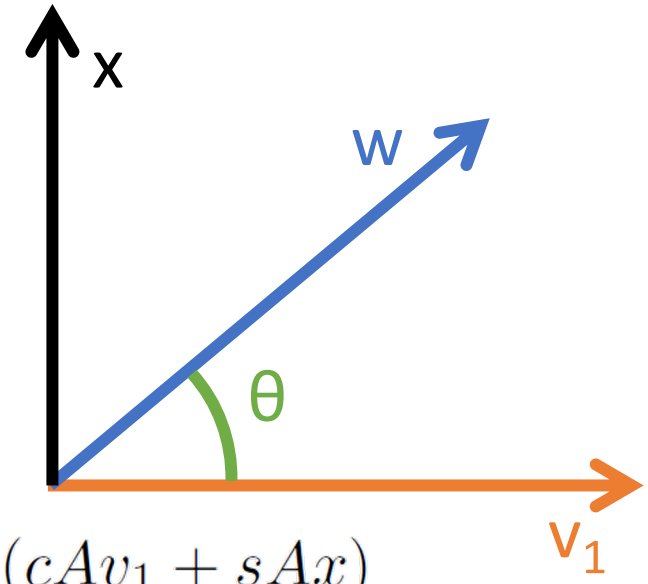
Define a unit vector x that is in the same direction as the component of w orthogonal to v_1

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)

- Properties of singular values, vectors



Let angle between w and v_1 equal θ.

Then, (see the picture) we can rewrite

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

$= cv_1 + sx$

where $c^2 + s^2 = 1$.

Then,

σ_1^2

$= \|Aw\|_2^2$

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

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

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

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

$\Downarrow$

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

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

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

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

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

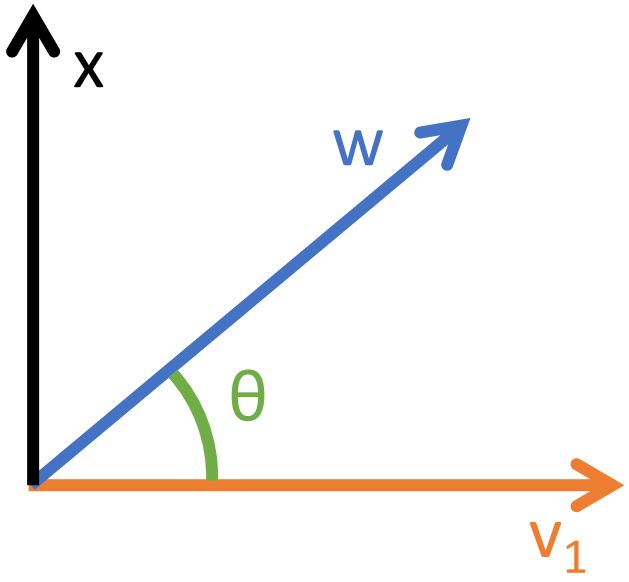
$= [0, y^\top]S^\top S[1, 0, \cdots, 0]^\top$ (where we define $[0, y^\top] \equiv x^\top V$)

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

$= 0$

Singular Value Decomposition (SVD)

- Properties of singular values



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:

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

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

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

Singular Value Decomposition (SVD)

- Properties of singular values
  - Why is norm(B) <= norm(A) ?

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

    - We know that A = USV^T , where U and V are orthogonal, and S is as shown above
    - Let β := norm(B)
    - By definition of norm(B),
      there exists a unit-norm column-vector “y” such that norm(By) = β
    - Use that “y” to construct a *longer (but still) unit-norm* column vector x := V [0,y^T]^T
    - norm (A x)
      = norm (USV^T V[0,y^T]^T)
      = norm (S [0,y^T]^T)
      = norm ([0, (By)^T]^T)
      = norm (By)
      = β
    - Thus, there exists an “x” such that norm(Ax) = β,
      which implies that norm(A) cannot be less than β, i.e., norm(A) >= β = norm(B)

Singular Value Decomposition (SVD)

- Properties of singular values

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 MxN, where M < N.
  - We use A := USV^T , where:
    - S is MxN (rectangular) diagonal. U is MxM orthogonal. V is NxN 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 MxM, and W'' is Mx1)
  - Covariance(X) = C = AA^T = U SS^T U^T = A'A'^T, where:
    - SS^T is square diagonal of size MxM
    - For matrix C to be SPD, the rank of S needs to be M (M non-zero singular values)