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

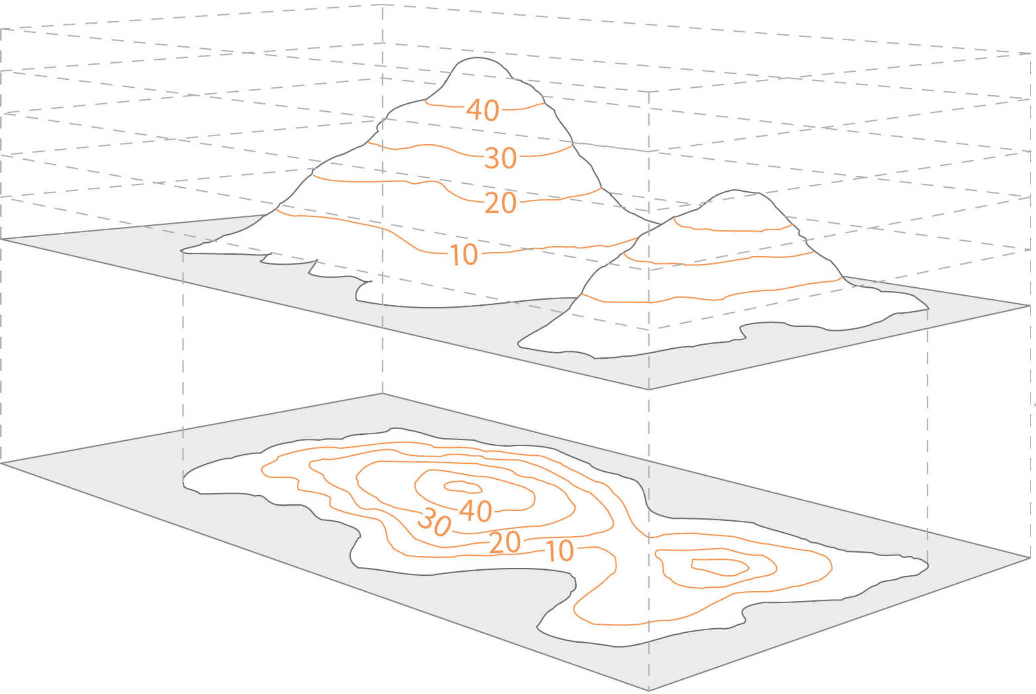
Multivariate Gaussian – Identity A

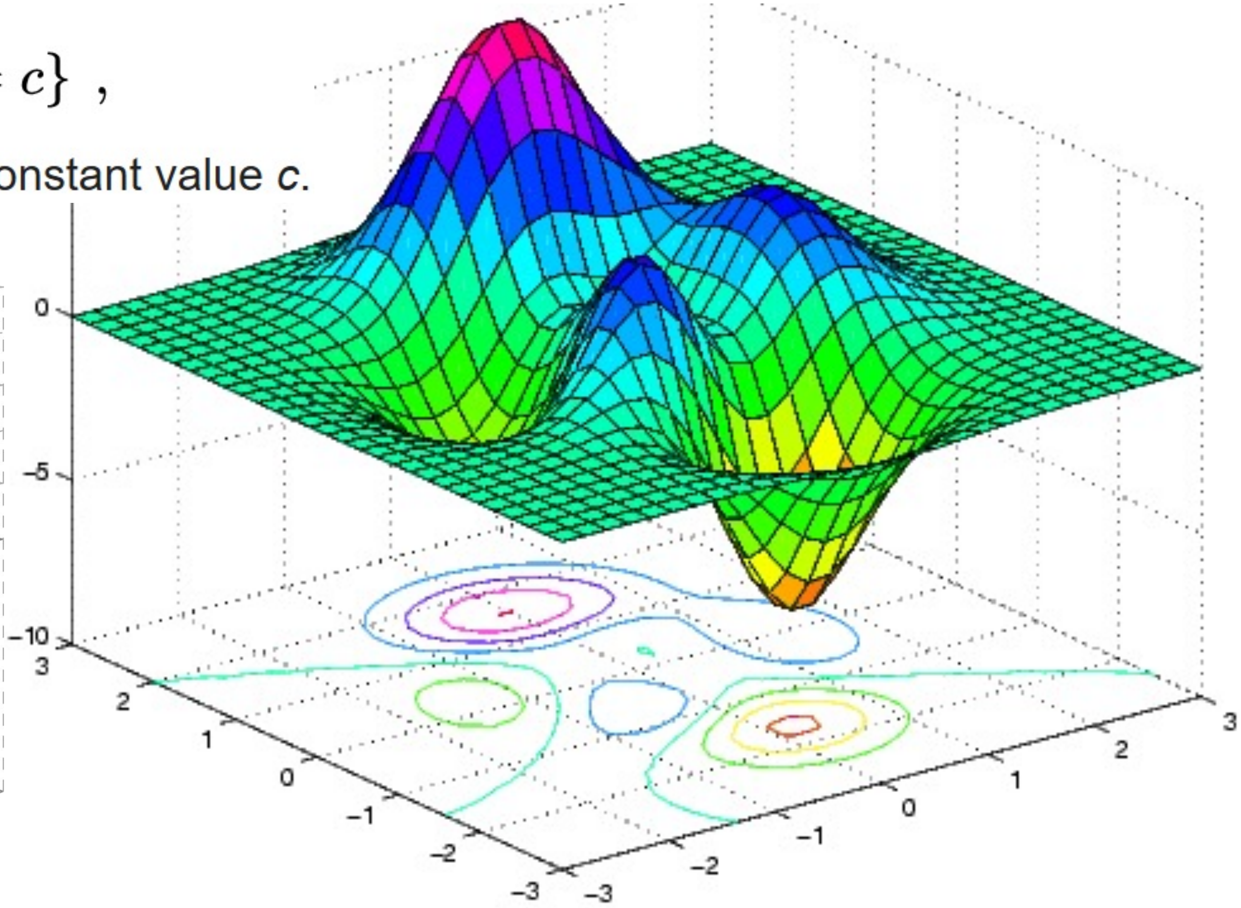
- What are the level sets of the PDF ?

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

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

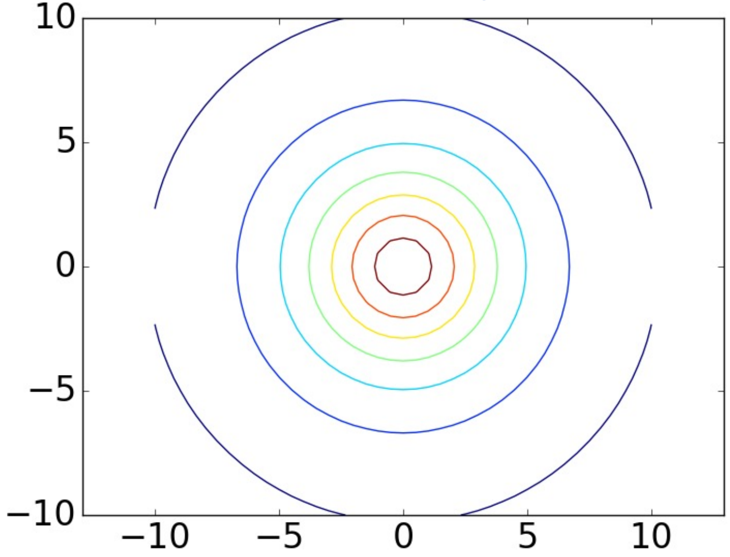
that is, a set where the function takes on a given constant value *c*.

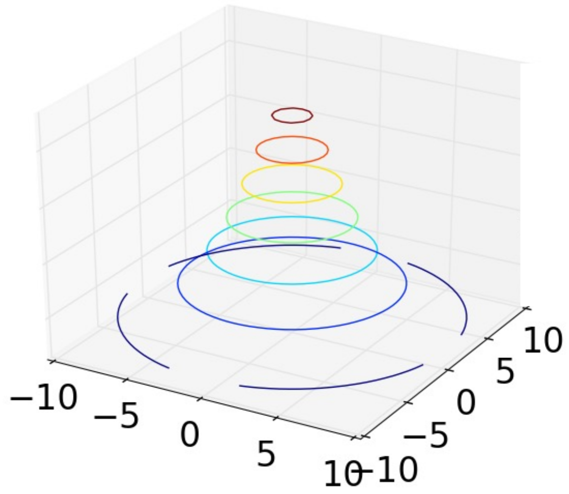




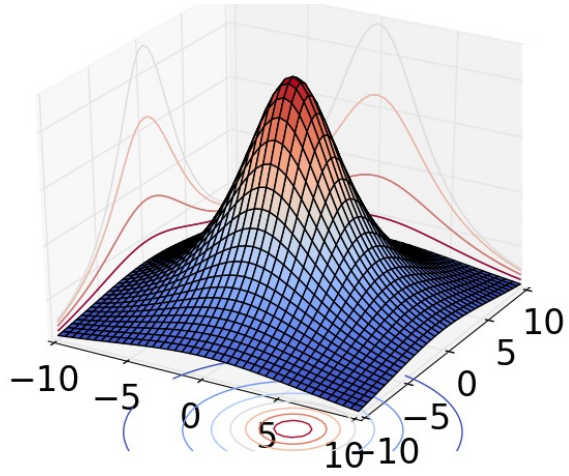
Multivariate Gaussian – Identity A

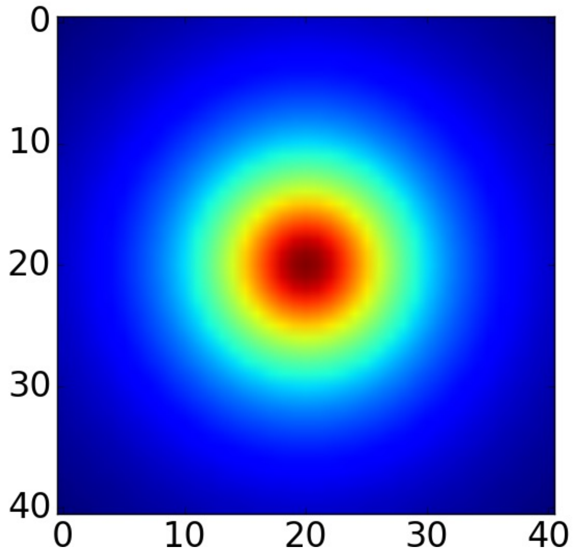
- Isotropic / spherical multivariate Gaussian
  - Level sets





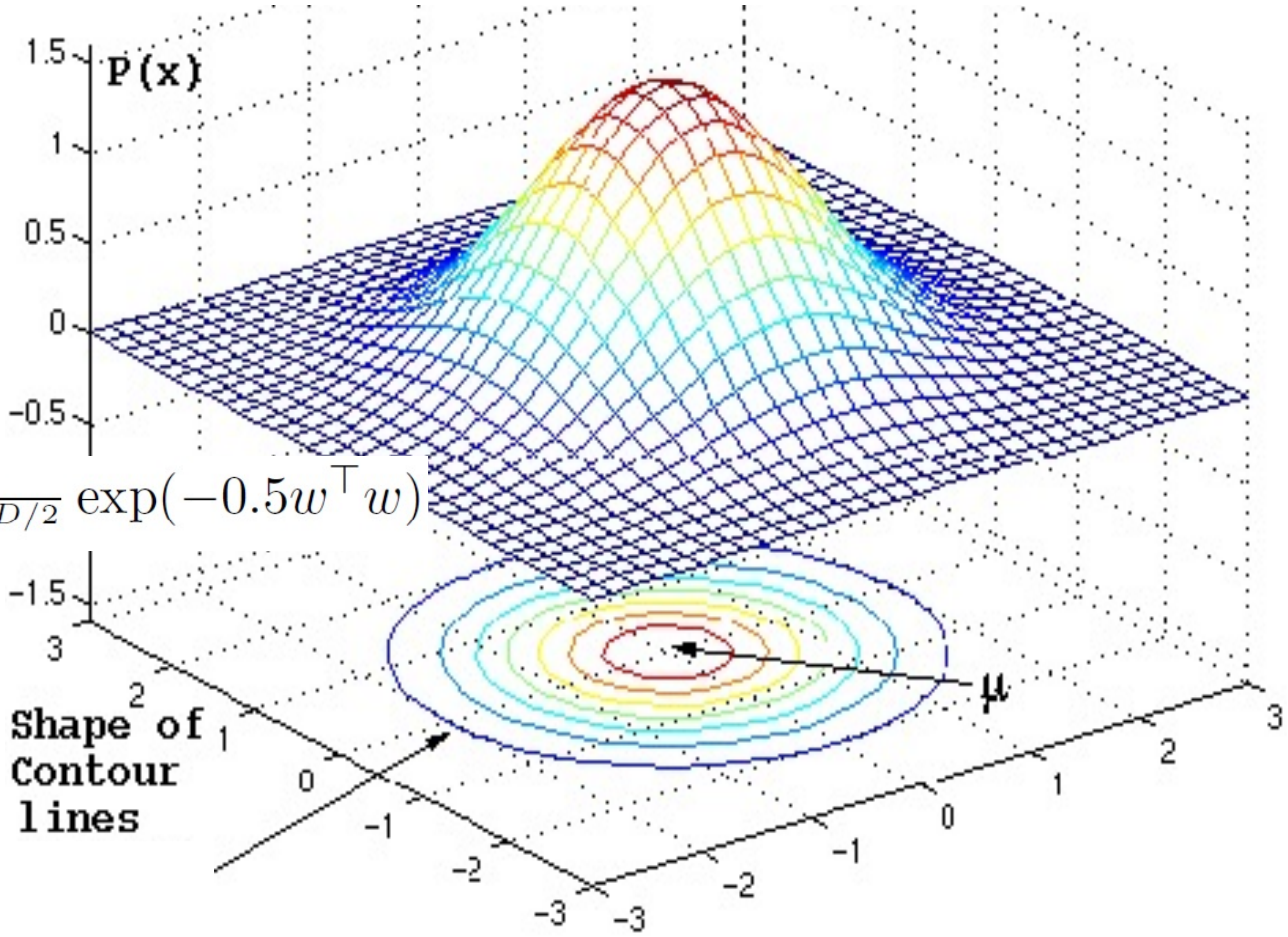
$$p(w) = \prod_d p(w_d) = \frac{1}{(2\pi)^{D/2}} \exp(-0.5 w^\top w)$$





Multivariate Gaussian – Identity A

- Isotropic / spherical multivariate Gaussian
  - Level sets



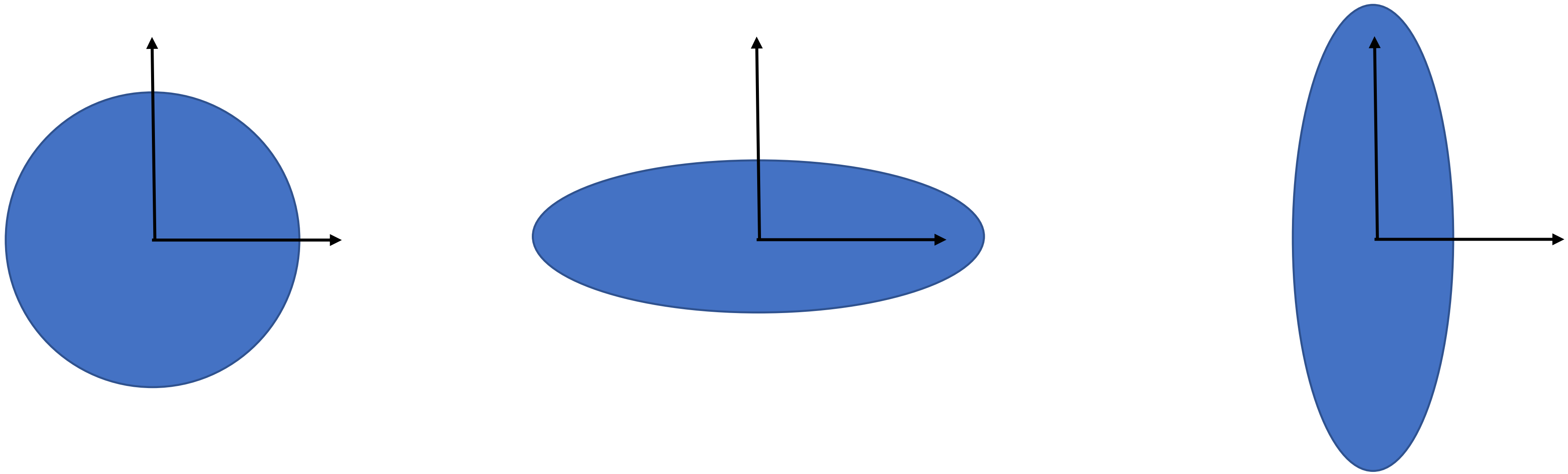
$$p(w) = \prod_d p(w_d) = \frac{1}{(2\pi)^{D/2}} \exp(-0.5 w^\top w)$$

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:
    - Center 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)

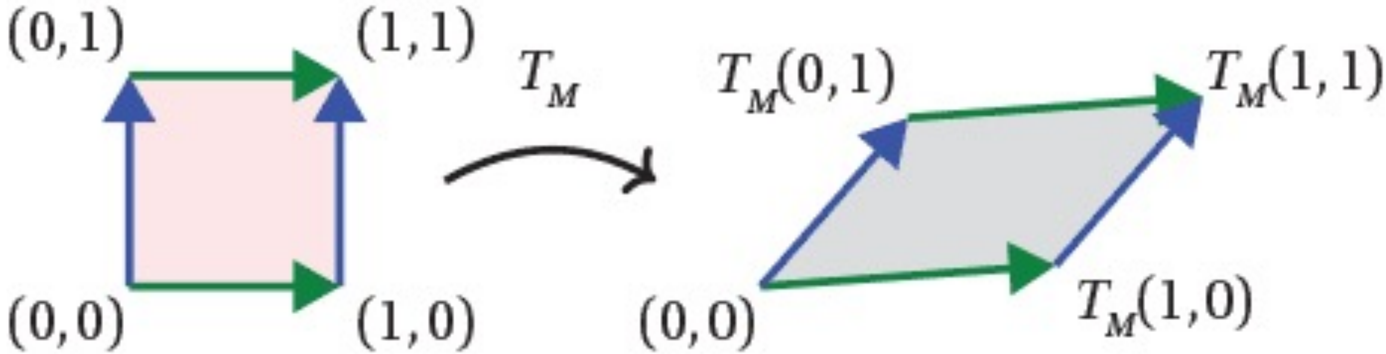


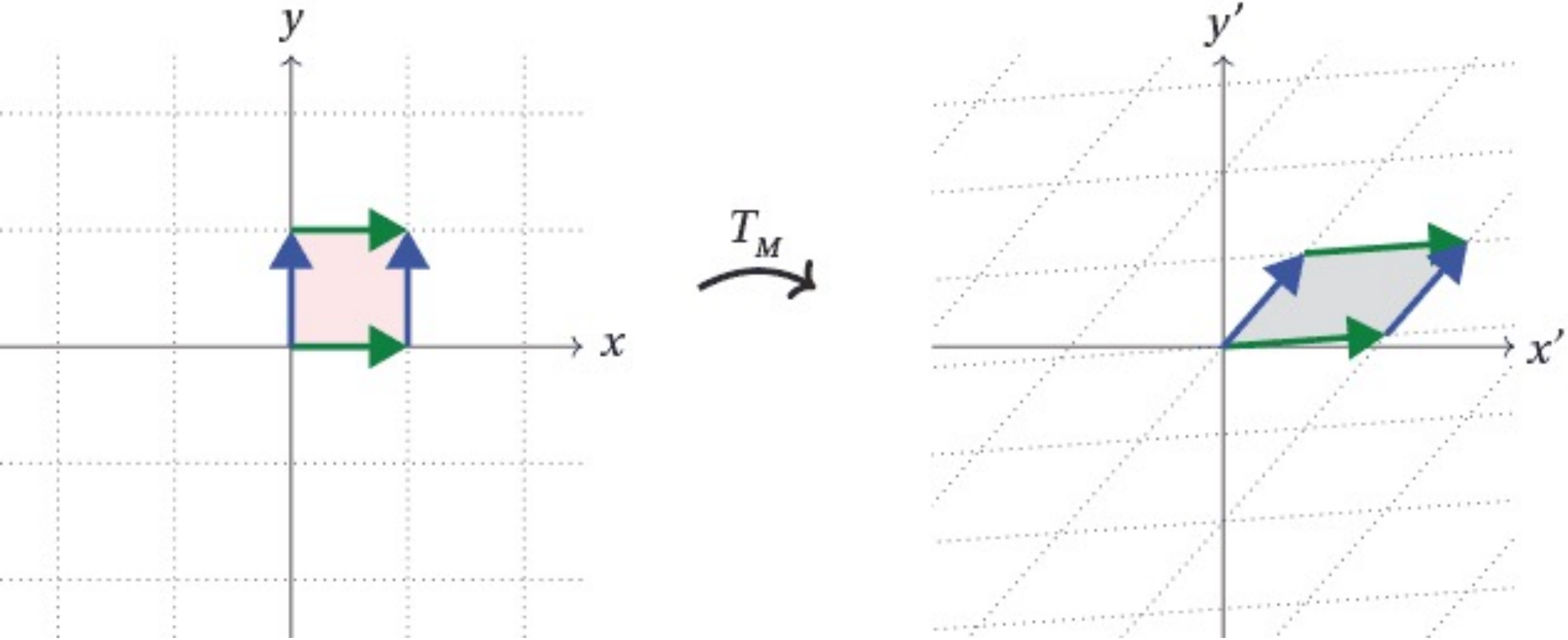
Multivariate Gaussian – Non-Singular A

- $X = A\,W + \mu$
- 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

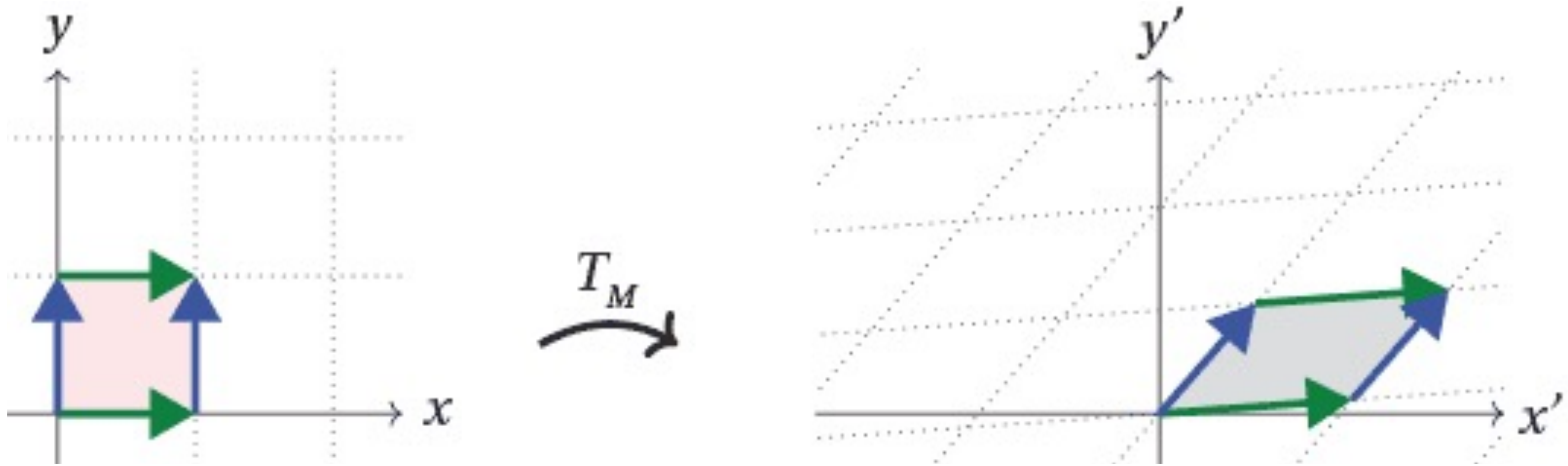
- Linear transformation
  $W := A^{-1} X$





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

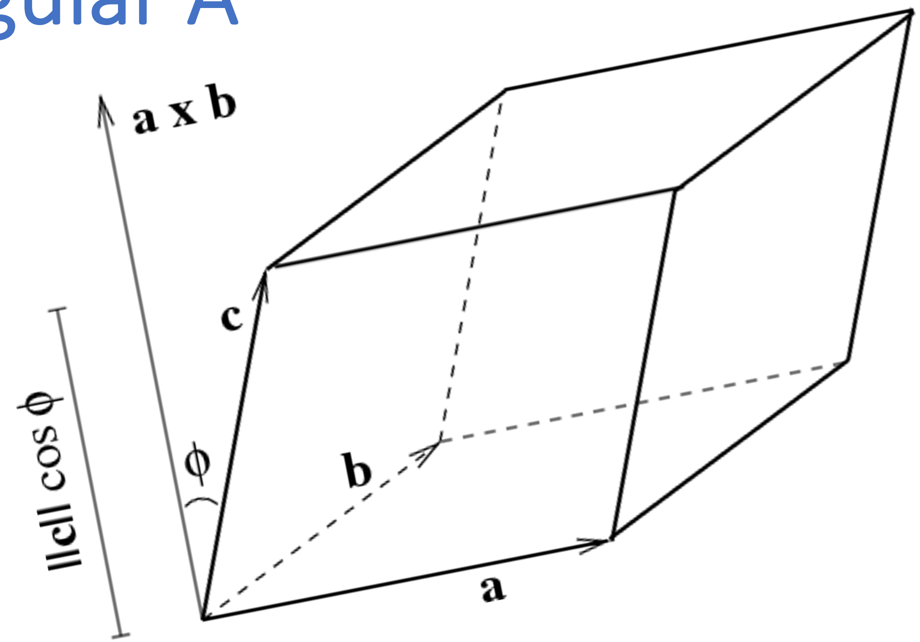
- Volume of a parallelepiped (in 3D)
  - Scalar triple product

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

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

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

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



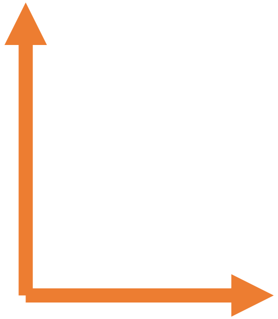
$$\begin{aligned}\text{Volume} &= \text{area of base}\cdot\text{height} \\ &= \|\mathbf{a}\times\mathbf{b}\|\,\|\mathbf{c}\|\,|\cos\phi| = |(\mathbf{a}\times\mathbf{b})\cdot\mathbf{c}|\end{aligned}$$

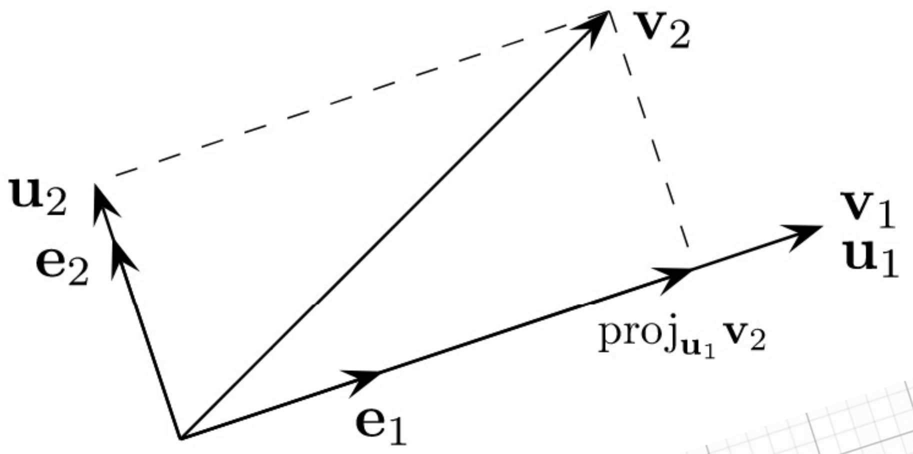
Multivariate Gaussian – Non-Singular A

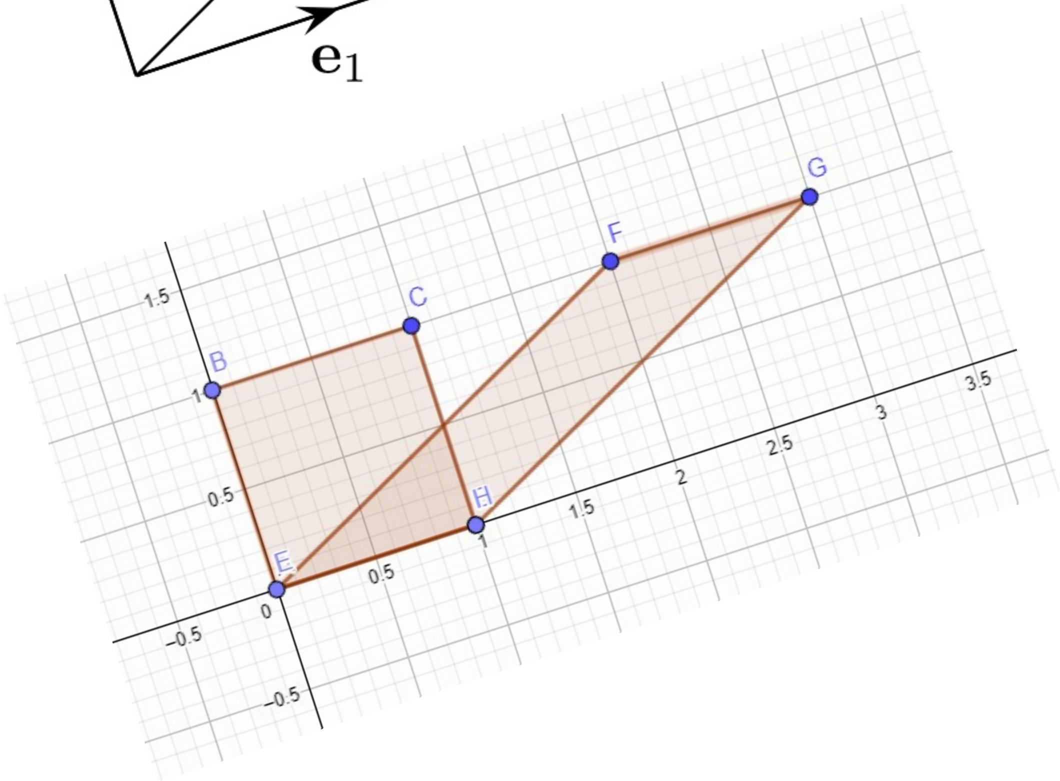
- Why is volume of hyper-parallelepiped given by determinant of matrix with columns as sides of hyper-parallelepiped ?
  - The following is an argument (not a proof; a separate inductive proof exists):
  - 2 important properties from linear algebra:
    Adding multiples of one column/side vector to another:
    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 $\{u_1, u_2\}$





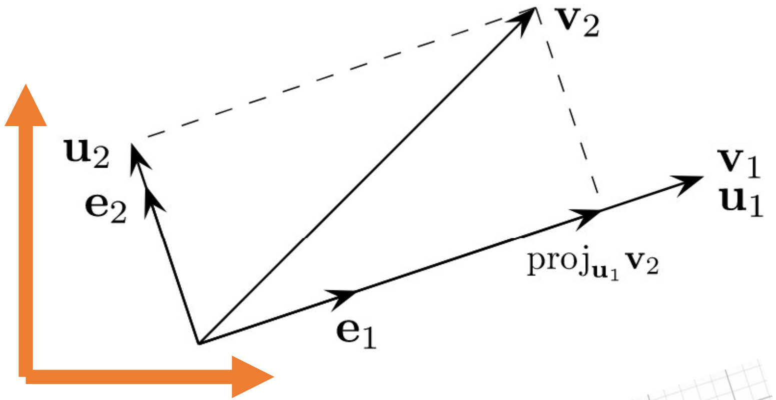


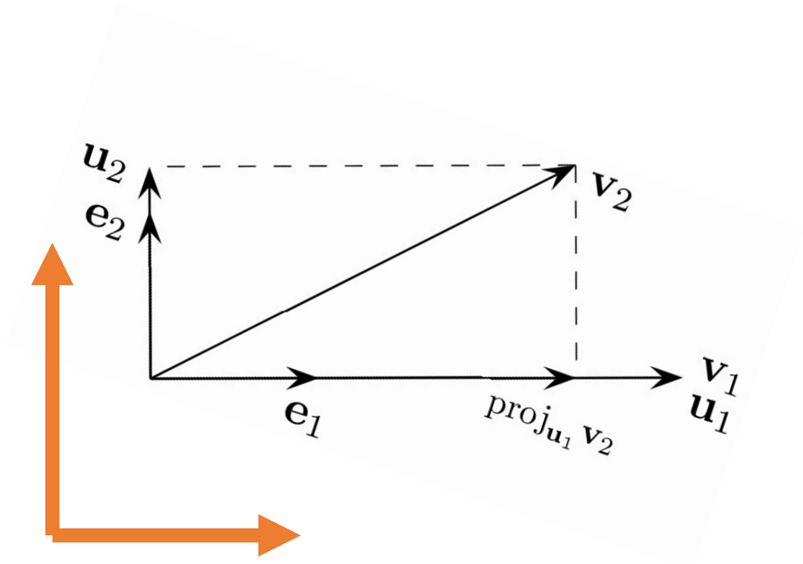
Multivariate Gaussian – Non-Singular A

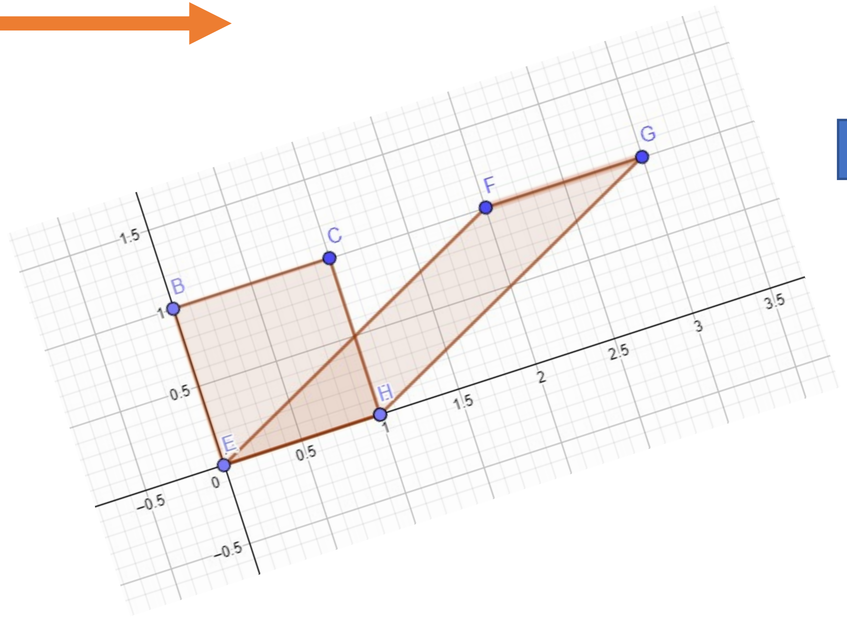
- Why is volume of hyper-parallelepiped given by determinant of matrix with columns as sides of hyper-parallelepiped ?
  - The following is an argument (not a proof; a separate inductive proof exists):
  - 2 important properties from linear algebra:
    Adding multiples of one column/side to another:
    - 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

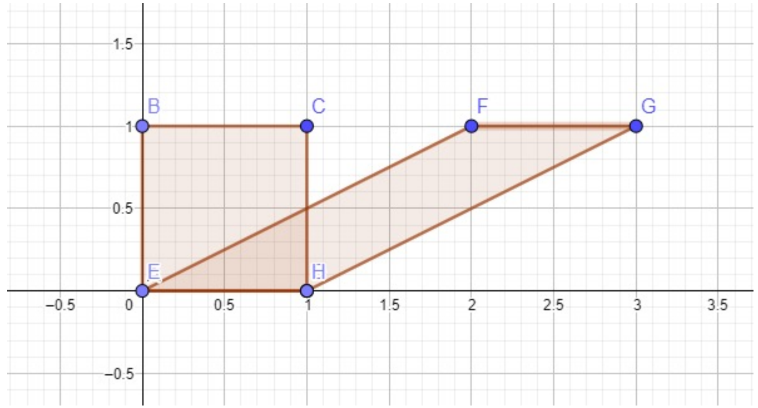
Multivariate Gaussian – Non-Singular A

- 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 magnitude (= product of diagonal-entries' magnitudes) = volume of a hyper-rectangle (= product of side lengths)
  - Now trace back all operations

Multivariate Gaussian – Non-Singular A

- X = A W + μ
- What is the PDF q(X) for non-singular square matrix A and μ = 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 $Y = X + \mu$ 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,
    Z
    $= B(AW + \mu) + c$
    $= (BA)W + (B\mu + c)$, 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}
  $= E_{P(Y)} [(Y_i - E[Y_i]) (Y_j - E[Y_j])]$
  $= E_{P(Yi,Yj)} [(Y_i - E[Y_i]) (Y_j - E[Y_j])]$
  $= \mathrm{Cov}(Y_i, Y_j)$

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 $a^\top Ca = E[a^\top (X - E[X])(X - E[X])^\top a] = E[(f(X))^\top f(X)] \geq 0$ that is the variance of a scalar RV $f(X) = (X - E[X])^\top a$

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.,
      $E_{P(W)}\,[\, A_{11}W_1 + A_{12}W_2 + \ldots A_{1D}W_D \,]$
      $= A_{11}\, E_{P(W)}\,[W_1] + A_{12}\, E_{P(W)}\,[W_2] + \ldots + A_{1D}\, E_{P(W)}\,[W_D]$
    - 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 $AA^\top$

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

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

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