10 Multivariate Gaussian

Generalizes a univariate Gaussian.

Consider a vector random variable $X = [X_1, X_2, \cdots, X_D]^\top$. Nothing but a joint RV with d RVs. Represent as a $d \times 1$ vector.

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

Formula for the PDF using Transformation of RVs

Example 2 (Zero Mean + Anisotropic): What is the PDF $q(X)$ for arbitrary non-singular SQUARE A and $\mu = 0$?

– Recall: Given PDF $p(w)$ and the transformation $X = g(W)$, the PDF $q(x) = p(g^{-1}(x)) \left| \frac{d}{dx} g^{-1}(x) \right|$
– In our case, $X = g(W) = AW$
– The inverse transformation $g^{-1}(X) = W = A^{-1}X$
– In the univariate case, we wanted the *magnitude* of the *derivative* of the inverse transformation: $\frac{\partial}{\partial y} g^{-1}(y)$
– In the multivariate case, we want the *volume* captured by the columns of the *Jacobian* of the inverse transformation: vol$(\frac{d}{dX} A^{-1} X) =$ vol$(A^{-1}) = \det(A^{-1}) = 1/\det(A)$

** Geometric intuition for vol$(A^{-1}) = 1/\det(A)$ (Note: determinant is defined only for a square matrix)
** Observe that the linear transformation A maps an infinitesimal hyper-cube $\delta \times \cdots \times \delta$ to an infinitesimal hyper-parallelepiped. If the axes of the hyper-cube were the cardinal axes, then the axes of the hyper-parallelepiped are the columns of A !!
** The volume of the hyper-parallelepiped is $\delta^d \det(A)$. In 3D, the volume can also be written as the scalar triple product $a_1 \cdot (a_2 \times a_3)$ where a_i is the i-th column of A

** Why is the volume equal to the determinant ?
The following is some intuition (not a proof; a separate inductive proof exists):
Adding multiples of one column to another:
1) keeps the determinant unchanged because the determinant function is multi-linear.
2) corresponds to a skew translation of the parallelepiped, which does not affect its volume.
Using Gram-Schmidt orthogonalization, we can transform matrix A to an orthogonal matrix A_{ortho} (NOT orthonormal; that would have determinant 1). This doesn't change the determinant or the volume.
We can rotate A_{ortho} to make it to diagonal form. Rotation doesn't change the determinant or the volume.
For this diagonal matrix, the determinant (= product of diagonal entries) equals the volume of a "rectangle" (= product of side lengths).

** Thus, $|dw| = \delta^d \implies |dx| = \delta^d \det(A)$
** Thus, $\frac{|dw|}{|dx|} = 1/\det(A)$

– Finally, the transformation of variables gives :
$q(X) = p(A^{-1}X) \frac{1}{\det(A)} = \frac{1}{(2\pi)^{d/2} \det(A)} \exp(-0.5 X^\top (A^{-1})^\top A^{-1} X)$
– Simplify: Let $C := AA^\top$. Then, $C^{-1} = A^{-T} A^{-1}$ and $\det(C) = \det(A) \det(A^\top) = (\det(A))^2$
– So, the multivariate-Gaussian PDF $q(X) = \frac{1}{(2\pi)^{d/2} |C|^{0.5}} \exp(-0.5 X^\top C^{-1} X)$, where C has a special name.

Property: The mean of $X = AW$ is zero

Proof: $E[AW] = AE[W] = A \cdot 0 = 0$
Note: $E[X] = [E[X_1], E[X_2], ..., E[X_d]]^\top$ (recall: all X_i share the same probability space).

Example 3 (Nonzero mean + Anisotropic): If X is multivariate Gaussian with zero mean, then $Y = X + \mu$ is multivariate Gaussian with PDF $p(y) = \frac{1}{(2\pi)^{d/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$
Proof:
Y is multivariate Gaussian because Y can be expressed as $AW + \mu$, where W_n is i.i.d. standard normal.
PDF $p(y) = \frac{1}{(2\pi)^{d/2}|C|^{0.5}} \exp(-0.5(y-\mu)^\top C^{-1}(y-\mu))$ because of the transformation of the variables $Y = X + \mu$

Property: The *mean vector* of $X = AW + \mu$ is μ.
Proof: $E[AW + \mu] = AE[W] + \mu = \mu$

Property: If Y is multivariate Gaussian, then $Z = BY + c$ is multivariate Gaussian, where A, BB are square invertible.
Proof: Because Y is multivariate Gaussian, $Y = AW + \mu$. Thus, $Z = B(AW + \mu) + c = (BA)W + (B\mu + c)$

Covariance Matrix

For any multivariate RV X, the definition of covariance is $C := E[(X - E[X])(X - E[X])^\top]$. This leads to a matrix C, where the outer-product structure implies that $C_{ij} = E[(X_i - E[X_i])(X_j - E[X_j])]$ which equals Cov(X_i, X_j).

Cov$(W) = E[WW^\top] = I$ because:
(i) Cov$(W_i, W_i) = 1$ and
(ii) Cov$(W_i, W_{j \neq i}) = 0$ because of independence of W_i and W_j

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 $C_{ij} =$ Cov(X_i, X_j).

More properties of C:

1) $C = E[XX^\top] - E[X](E[X])^\top$
Proof: Expand the terms in the definition.

2) C is symmetric
Proof: $C_{ij} =$ Cov$(X_i, X_j) =$ Cov$(X_j, X_i) = C_{ji}$

3) C is positive semi-definite (PSD)
Proof: For any $d \times 1$ non-zero vector a, $a^\top Ca = E[a^\top (X - E[X])(X - E[X])^\top a] = E[(f(X))^\top f(X)] \geq 0$ that is the variance of a scalar RV $f(X) = (X - E[X])^\top a$

Marginal PDFs

Property: The 1D marginal PDF of the multivariate Gaussian Z, 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) the transformations of scaling and translation on a univariate Gaussian RV leads to another univariate Gaussian RV,

(3) sum of two univariate Gaussian RVs leads to another univariate Gaussian RV (using concepts on convolution).

Property (Rotation+Reflection): If $\mu = 0, A = R$, 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 $W = R^\top Y$.
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.

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

Proof: Transformation of random vectors. $|\det(S)| = \prod_d S_{dd}$. Inverse transformation is $W = S^{-1} Y$.
So, $Q(y) = (1/(2\pi)^{d/2})(1/\det(S)) \exp(-0.5 (S^{-1} y)^\top S^{-1} y) = (1/(2\pi)^{d/2}(1/\det(S)) \exp(-0.5 y^\top (S^2)^{-1} y)$
Thus, $Y := SW$ is a zero-mean anisotropic (axis-aligned) multivariate Gaussian.

Property (Scaling and then Rotation+Reflection): If $\mu := 0, A := RS$, then $Y := RSW$ has the PDF
$P(y) = (1/(2\pi)^{d/2})(1/\det(S)) \exp(-0.5 y^\top (RS^2 R^\top)^{-1} y)$

Proof: Transformation of random vectors. $\det(RS) = \prod_d S_{dd}$. Inverse transformation is $W = (RS)^{-1} Y = S^{-1} R^\top y$.
So, $Q(y) = (1/(2\pi)^{d/2})(1/\det(S)) \exp(-0.5 (S^{-1} R^\top y)^\top S^{-1} R^\top y) = (1/(2\pi)^{d/2}(1/\det(S)) \exp(-0.5 y^\top (RS^2 R^\top)^{-1} y)$
Thus, $Y := RSW$ is zero-mean rotated+reflected anisotropic multivariate Gaussian with covariance $C = RS^2 R^\top$.

Property: Marginal PDFs of the multivariate Gaussian Z in n-dimensions, over any chosen subset of the variables, are (multivariate) Gaussian.

Proof: Choose the transformation B as the projection matrix of size $m \times n$ where $m < n$ with ones on diagonal and zeros elsewhere.
If we consider multivariate Gaussian $Z := \mu + AW$, then $BZ = B\mu + BAW = \mu' + A'W$, where $A' := BA$

SVD for any $m \times n$ matrix $A' = USV^\top$ exists, where
U is $m \times m$ orthogonal (real) / unitary (complex) s.t. $UU^* = U^* U = I$.
S is $m \times n$ diagonal with non-negative reals (singular values).
V is $n \times n$ orthogonal (real) / unitary (complex).

FYI: Motivation for SVD: The image of the unit sphere under any $m \times n$ matrix is a hyper-ellipse.
FYI: The singular values are uniquely determined.
FYI: For a square matrix whose singular values are distinct, the SVD is unique upto sign of singular vectors, i.e., columns of U and V.
FYI: Rank of a matrix = number of non-zero singular values

So, $A'W = USV^\top W$.
We know that $V^\top W$ is also a zero-mean isotropic Gaussian in n-dimensional Euclidean space
$S(V^\top W)$ extracts the first m components from $V^\top W$, which are i.i.d. standard normal, and then scales them
$S(V^\top W)$ will be a zero-mean anisotropic (axis-aligned) multivariate Gaussian in m-dimensional Euclidean space
$U(S(V^\top W))$ will be a zero-mean rotated anisotropic Gaussian in m-dimensional Euclidean space

Important: Marginal PDFs being Gaussian doesn't imply the joint PDF is multivariate Gaussian. See example below.

Example: Let X be a Gaussian random variable with zero mean, and $Y = B(X)$ where $B(X)$ is $+X$ or to X with equal probability.
Note: Y also has a Gaussian PDF.
The joint PDF $P(X, Y)$ for this case is like a cross $\times$. Thus, the joint PDF $P(X, Y)$ isn't multivariate Gaussian because for this PDF $P(x, y) = 0, \forall |x| \neq |y|$, which isn't the case for the multivariate Gaussian.
Moreover, Cov$(X, Y) = E[X \cdot B(X)] - E[X]E[B(X)] = 0 - 0 = 0 \neq B(E[X^2]) - B(E[X])^2$

Note: Covariance and correlation are guaranteed to be informative only for linearly dependent random variables. In the above case $Y := B(X)$, RVs X and Y aren't deterministically and linearly related, i.e., $B(X)$ is neither a deterministic nor a linear function of X.

Eigen Decomposition

For any $N \times N$ matrix A, an eigenvector is a non-zero vector v s.t. $Av = \lambda v$. Then λ is the associated eigenvalue.

For a general $N \times N$ matrix M, the eigen-decomposition is $M = Q \Lambda Q^{-1}$, where the columns of M are the eigenvectors. However, the eigenvectors needn't be orthogonal and Q^{-1} needn't equal $Q^\top$.

Every $N \times N$ *real symmetric* matrix M (like the covariance matrix C) has an eigen-decomposition $M = Q \Lambda Q^\top$, where Q is an orthogonal matrix (i.e., $Q^\top Q = Q Q^\top = I$).

Every $N \times N$ *real SPD* matrix M (like the covariance matrix C) has an eigen-decomposition $M = Q \Lambda Q^\top$, where Q is orthogonal and all eigenvalues $\Lambda_{dd} > 0$

Contours of the Multivariate Gaussian PDF

Property: If X is multivariate Gaussian in 2D with a **diagonal** (invertible) covariance matrix, then the iso-probability contours of $P(X)$ are ellipses whose axes are aligned with the cardinal axes.

Proof:
(1) C^{-1} is SPD because C is SPD; this can be seen from the (unique) Cholesky decomposition of any SPD matrix $C = M^\top M$, where M = upper triangular with positive diagonal entries.
More important to this case, M is invertible (otherwise C won't be invertible).
Thus, $C^{-1} = M^{-1} M^{-\top} = N N^\top$ that is also SPD (it is symmetric; $\forall x \neq 0, x^\top N N^\top x > 0$ because N are $N^\top$ invertible)
It will also turn out that N will be upper triangular with positive diagonal entries.

(2) The contour $\{x \in \mathbb{R}^D : P(x) = \alpha\}$ is the same as
$\{x : (x - \mu)^\top C^{-1} (x - \mu) = \beta\}$ (β must be positive because C^{-1} is SPD)
$\{x : \sum_d (x^d - \mu^d)^2 / (\sigma^d)^2 = \beta\}$
$\{x : \sum_d (x^d - \mu^d)^2 / (\beta(\sigma^d)^2) = 1\}$
This is the equation of an ellipse in $\mathbb{R}^D$ with center μ and axes of lengths $\sigma^d \sqrt{\beta}$ aligned with the coordinates axes.

Mahalanobis Distance

Term $(x - \mu)^\top C^{-1} (x - \mu)$ appearing in the exponent = squared Mahalanobis distance of a point x from the mean μ.
Multidimensional generalization of measuring distance along a dimension.

For a diagonal C, this measures distance in terms of the units of the standard deviation of the data along that dimension. That is, how many standard deviations away is the point x from the mean μ. This introduces scale invariance.

Mahalanobis distance (for diagonal C) rescales the units along each dimension, based on the variance of the data along that dimension

Mahalanobis distance reduces to the Euclidean distance when $C = I$

A Gaussian's iso-probability contour is the locus of points with the same Mahalanobis distance to the mean.

Property: The Mahalanobis distance is a true distance metric.

Proof:
Given a *non-singular* covariance C, let Mahalanobis distance be $d(x,y) := \sqrt{(x-y)^\top C^{-1}(x-y)}$

1) Positivity: $d(x,y) \geq 0, \forall x, y$. Follows from the positive semi-definiteness of C^{-1}
$d(x,y) = 0$ when $x = y$
$d(x,y) > 0$ when $x \neq y$ (note: C is non-singular)

2) Symmetry: Follows from definition.

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

Proof for the case when the covariance C is diagonal.
Let $u := x - z$ and $v := z - y$
Then $u + v = x - y$
LHS $= \sqrt{(u+v)^\top C^{-1}(u+v)}$
RHS $= \sqrt{u^\top C^{-1} u} + \sqrt{v^\top C^{-1} v}$

We know that all distances are positive. So, showing LHS < RHS is same as showing LHS^2 < RHS^2

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 $a, b \in \mathbb{R}^D$

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.

Scaling and Rotating the coordinate frame

(1)
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 $x^\top C^{-1} x = x^\top S^{-2} x$

(2)

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 $P(X)$.

Connections to PCA

Suppose $A = USV^\top$ was applied to W that had a spherical PDF
Then,
1) Rotation $V^\top$ doesn't change the structure of the spherical PDF. The covariance C is still the identity I.
2) Scaling S scales each dimension d by S_{dd} making the PDF anisotropic. Consider distinct $S_{11} > S_{22} > \cdots$
This yields covariance $C = S^2$ such that Cov$(X_d, X_e) = 0$ and Cov$(X_d, X_d) = S_{dd}^2 =$ Var(X_i) where $X = SW$
3) Rotation U rotates the anisotropic PDF so that the variance S_{dd}^2 is now along U_i
Now $C = US(US)^\top = US^2U^\top$
Given data, we can empirically estimate $\widehat{C}$ as the sample covariance that will tend to equal C asymptotically
Given $\widehat{C}$, we can get back the vectors U and variances S^2 by performing an eigen decomposition of $\widehat{C}$, producing eigenvectors U (upto sign) and eigenvalues S_{ii}^2

ML Estimation for Mean and Covariance

MLE for mean is sample mean. Prove.

Note: $\frac{d}{d\mu}(x-\mu)^\top C^{-1}(x-\mu) = 2C^{-1}(x-\mu)$ (Rely on symmetry of C)

MLE for covariance is sample covariance. Prove.

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

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

PCA: Directions of maximal variance

Suppose we have data $\{x_i\}_{i=1}^n$ drawn from a Gaussian PDF with mean $\mu = 0$ and *diagonal* covariance C

For a Gaussian PDF, a diagonal covariance implies that (i) $X^d = \sqrt{C_{dd}}W^d$, where W_d is a standard-Normal RV and (ii) the ellipsoidal data distribution is s.t. the ellipsoidal axes are aligned with the coordinate axes

Find the direction v ($\| v \|_2 = 1$) s.t. the data projected on the subspace v (containing the origin = mean) has the maximal variance

Projected data = $\langle x_i, v\rangle v$
Mean of the projected data $= \sum_i \langle x_i, v\rangle v = \langle \sum_i x_i, v\rangle v = 0$
Projected data is 1D
Distance of projected data from the mean $= \| \langle x_i, v\rangle v \|_2 = |\langle x_i, v\rangle|$
Variance of projected data $= \sum_i \langle x_i, v\rangle^2$
Optimal direction $= \arg\max_{v:\|v\|_2=1} \sum_i \langle x_i, v\rangle^2$
$= \arg\max_{v:\|v\|_2=1} \sum_i (x_i^\top v)^2$
$= \arg\max_{v:\|v\|_2=1} \sum_i (x_i^\top v)^\top (x_i^\top v)$
$= \arg\max_{v:\|v\|_2=1} \sum_i v^\top x_i x_i^\top v$
$= \arg\max_{v:\|v\|_2=1} v^\top (\sum_i x_i x_i^\top) v$
$= \arg\max_{v:\|v\|_2=1} v^\top C v$ (this is the connection between sample covariance C and direction v maximizing variance of projected data)
$= \arg\max_{v:\|v\|_2=1} \sum_d C_{dd}(v^d)^2$ (because C is diagonal)

This is maximized when $v^d = 1$ for $d = \arg\max_e C_{ee}$ and $v^d = 0$ otherwise

Think: I'll put all "weight" on that component of v that is associated with the maximum of the diagonal elements in C

Think: Constraint set = hypersphere. Contours of the objective function are ellipsoids with the *minor* axis being the dimension $\equiv \arg\max_d C_{dd}$. Thus, the point on the hypersphere that maximizes the objective function lies at the intersection of the *minor* axis with the hypersphere.

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 $u^c = 1$ for $c = \arg\max_{d\neq e} C_{dd}$ and $u^c = 0$ otherwise

Think: I'll put all "weight" on that component of u that is (i) *not* the d chosen before and (ii) associated with the maximum of the *remaining* diagonal elements in C

Similar arguments hold for 3rd, 4th, ... directions

Thus, *for a diagonal covariance matrix the cardinal directions are the directions maximizing variance*.
These directions = *principal components of variation*.

Rotating the coordinate frame by pre-multiplication with a orthogonal matrix U simply rotates the principal components.

Another perspective:
When the SPD covariance matrix is $C = Q\Lambda Q^\top$ (non-diagonal), where the diagonal of Λ has sorted values,
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 first principal eigenvector is given by $u = [1, 0, \cdots, 0]^\top$ that makes $v = Qu =$ first column of Q
Simlarly, the remaining principal eigenvectors will be the other columns of Q

PCA and Eigen decomposition

The *principal directions* U for data modeled by $X := AW + \mu$ are obtained by performing an eigen-decomposition of the (empirical) covariance matrix $C = US^2U^\top$

The *variances* diag(S^2) along the principal directions for data modeled by $X := AW + \mu$ are obtained by performing an eigen-decomposition of the (empirical) covariance matrix $C = US^2U^\top$