CS215
Transformation of R.V.
and
Multivariate Gaussian

By: Harsh Shah

October 2021

Contents

1 Transformation of Random Variable

Given any continuous r.v. X with PDF $P_X(x)$ and given any function $g(X)$(defined on range of X) we intend to find PDF associated with the r.v. $Y = g(X)$.
For simplicity, let's assume $g(.)$ is monotonic increasing.
Then by probability mass conservation,

$$P(a < X < b) = P(g(a) < Y < g(b)) = \int_{g(a)}^{g(b)} Q(y)dy$$

But $y = g(x)$, hence,

$$P(a < X < b) = \int_a^b P(x)dx = \int_{g(a)}^{g(b)} P(g^{-1}(y))\frac{d(g^{-1}(y)dy)}{dy}$$

Therefore,

$$\int_{g(a)}^{g(b)} P(g^{-1}(y))\frac{d(g^{-1}(y)dy)}{dy} = \int_{g(a)}^{g(b)} Q(y)dy$$

for any a, b. Hence,

$$Q(y) = P(g^{-1}(y))\frac{d(g^{-1}(y))}{dy}$$

To handle monotonically decreasing $g(.)$ as well,

$$Q(y) = P(g^{-1}(y))\left|\frac{d(g^{-1}(y))}{dy}\right|$$

Example: Let $X = G(x; 0, 1)$ and $Y = X^2$(non-monotonic). We aim to find the PDF associated with Y.
Now,

$$P_Y(y) = G(\sqrt{y}; 0, 1)\left|\frac{d(\sqrt{y})}{dy}\right| + G(-\sqrt{y}; 0, 1)\left|\frac{d(-\sqrt{y})}{dy}\right|$$

Expression of $G(x; 0, 1)$ can be substituted to reach the result.

2 Multi-variate Gaussian Disribution

2.1 Definition

Let X be a vector of random variables of dimension D given by

$$X = [X_1; X_2 \ldots; X_D]$$

A r.v. X has a joint PDF as multi-variate Gaussian distribution $\exists$ finite i.i.d. standard Gaussian r.v. $W_1, W_2, \ldots W_N$ with $N > D$ such that

$$X = AW + \mu$$

Level sets of a function: $L_c(f) = \{(x_1, x_2 \ldots x_n 0 | f(x_1, x_2 \ldots x_n) = c\}$ A spherical multi-variate Gaussian has spherical level sets. Now let's consider various forms A can take resulting in different PDFs.

2.2 A is diagonal

In this case, the X_i are independent and given by

$$X_i = A_{ii} W_i + \mu_i$$

for $i = 1, \ldots, D$
The standard deviation of distribution of X_i is A_{ii}.
The level sets of $Q(X)$ are hyper-ellipsoids in D dimensions with axis aligned along cardinal axis.

$$Q(X) = G(\mu_1, A_{11}) \cdot G(\mu_2, A_{22}), \ldots G(\mu_D, A_{DD})$$

2.3 A is non-singular square matrix

Let's take $\mu = 0$ for simplicity.
Similar to univariate case, where scaling was determined by $\left|\frac{d(g^{-1}(y))}{dy}\right|$, the scaling for multi-variate case is determined by determinant of matrix of derivatives, Jacobian matrix.
Also, $W = A^{-1}X$, which is a linear transformation of vector X. A^{-1} maps a hypercube to parallelepiped. If the vectors describing the hypercube are along cardinal axis, then the parallelepiped is described by vectors which are columns of A^{-1}.
We intend to find the volume of the parallelepiped formed due to this transformation.
Claim: The volume of parallelepiped described by column vectors of matrix A^{-1} is given by $\det(A^{-1})$
Proof: Addition of any scaled column of a matrix M to another column does not change the determinant.
Therefore by Gram-Schmidt orthogonalization process the columns of A^{-1} can be constructed to be orthogonal to each other, without changing the determinant. Then multiplying by an orthogonal matrix would rotate the orthogonal vectors(to align them with cardinal axis), and this operation would not change the determinant as well. Now the result matrix is diagonal square matrix and the volume of the parallelepiped described by the column vectors is given by product of diagonal elements.

From the above result, an infinitesimal volume δ^D after transformation becomes $\delta^D \cdot det(A^{-1})$.
The PDF for X is now given by

$$P(X) = \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{det(A)} \cdot exp(0.5 \cdot (A^{-1}X)^T \cdot A^{-1}X)$$

Let $C = A \cdot A^T$. Then $det(A) = \sqrt{det(C)}$. The above expression can we rewritten as

$$P(X) = \frac{1}{(2\pi)^{D/2}} \cdot \frac{1}{\sqrt{det(C)}} \cdot exp(0.5 \cdot X^T \cdot C^{-1} \cdot X)$$

In case of $\mu \neq 0$, substitute $X := X - \mu$
Mean vector?

$$E_{P(X)}[X] = E_{P(W)}[A \cdot W + \mu] = A \cdot E[W] + \mu = \mu$$

3 Covariance of vector r.v.

Let Y be any column vector of r.v., then the covariance matrix C for Y is given by

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

And,

$$C_{ij} := E_{P(Y_i,Y_j)}[(Y - E[Y])(Y - E[Y])^T] = cov(Y_i, Y_j)$$

In case of multi-variate Gaussian random vector X, $C = AA^T$ turns out to be the covariance matrix.

Properties of covariance matrix,

- $C = E[X \cdot X^T] - E[X] \cdot (E[X])^T$
- C is symmetric, since $cov(X_i, X_j) = cov(X_j, X_i)$
- C is positive semi-definite(For any column vector a, $a^T \cdot C \cdot a \geq 0$)

Some terminology:

- Orthogonal matrix: A matrix with columns as unit normal vectors $\iff M^T M = I$
- Rotation matrix: Orthogonal matrix with determinant 1. If determinant is -1, the matrix can model reflection and rotation. At times, the term rotation is used for improper rotation(reflection) as well
- Reflection matrix: Symmetric orthogonal matrix
- Diagonalizable matrix: $\exists$ invertible matrix P such that $P^{-1}MP = D$, where D is a diagonal matrix
- Defective matrix: Non-diagonalizable matrix

Important results:

- If A,which is a $n \times n$ matrix, is diagonalizable, then it has n-linearly independent eigenvectors.
- Invertible does not imply diagonalizable, and nor does diagonalizable imply invertible
- Every real symmetric matrix is diagonalizable. Further the matrix P in this case is orthogonal.
- **Spectral theorem**: If A is $n \times n$ real symmetric square matrix, the it has n-linearly independent real valued orthogonal eigenvectors, with real eigenvalues.(Can be proved by carrying out simplification on $\lambda v^T v^*$ where v is eigenvector and lambda is corresponding eigenvalue)
- A symmetric real-valued positive definite matrix has all eigenvalues positive.(Can be proved using spectral theorem and definition of positive definite)

Now consider multi-variable Gaussian r.v., $Y := AW + \mu$, we intend to find level sets of the corresponding PDF,

$$(y-\mu)^T C^{-1}(y-\mu) = a > 0$$
$$(y-\mu)^T (Q^T)^{-1} D^{-1} Q^{-1}(y-\mu) = a$$

Let $y' = Q^{-1}(y-\mu)$, then

$$(y')^T D^{-1} y' = a$$

From the above expression, the level sets are hyper-ellipsoids with half axes being $\sqrt{D_{ii}}$ aligned with columns of $Q^{-1} = Q^T$.

4 Marginal PDF

For a multi-variate Gaussian vector X, each marginal PDF (i.e., $P_{X_i}(x_i)$) is itself a univariate Gaussian distribution. Proof: $X_i = \sum_{j=1}^{n} A_{ij} W_j + \mu_i$. Since sum of Gaussian random variables is also a Gaussian, X_i is also Gaussian.
Further, any subset of the r.v. in X is also a multi-variate Gaussian(Proof: Multiply by projection matrix(say $B_{m\times n}$) in which every row has exactly one element non-zero and unity. The position of the 1 determines the random variables in the subset. Note that BA will be of rank m because A is invertible).
Important: Marginal PDF being Gaussian **does not** imply joint PDF is multi-variate Gaussian distribution.

5 Conditional PDF

Given a multi-variate Gaussian random vector X and let X_1, X_2 be partition of elements of X. Then the PDF, $P(X_1|X_2 = x_2)$ is also a multi-variate Gaussian distribution(Note that X_1 and X_2 can be vectors). Intuitively, each slice of a bivariate Gaussian distribution(parallel to cardinal axes) has a univariate Gaussian distribution. Surprisingly, even a slice not parallel to cardinal axis, gives a Gaussian distribution.

6 ML estimation for multi-variate Gaussian

6.1 ML estimation of mean

We have to find μ at which the likelihood function($\Pi_{i=1}^{n} Q(y_i)$) is maximum. Taking log and equating the derivative w.r.t. μ to 0 , we have,

$$\sum_i \frac{d(y_i-\mu)^T C^{-1}(y_i-\mu)}{d\mu} = 0$$

Now consider derivative w.r.t. a for quadratic $a^T B a = \sum_{i,j} B_{ij} a_i a_j$,

$$k^{th}\text{component of Jacobian}(J_k) = \sum_i [B_{ik} a_i] + \sum_j [B_{kj} a_j] = 2\sum_i B_{ik} a_i$$

(since B is symmetric). Hence,

$$J_k = B_k(k^{th} row) \cdot a$$

Using the above result we have,

$$\sum_i C^{-1}(y-\mu) = 0$$

$$\mu = \frac{1}{N}\sum_i y_i$$

6.2 ML estimation of covariance

For this, we require to find derivative of the likelihood function w.r.t. C, and hence partial derivative w.r.t. C_{ij}.
It can be proved that,

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

$$\frac{d[log(|C|)]}{dC} = C^{-T}$$

Upon simplifiation,

$$C_{ML} = \frac{1}{N}\sum_i (y_i-\mu)(y_i-\mu)^T$$

7 Mahalanobis distance

Let C be a symmetric positive definite matrix. Then Mahalanobis distance between vectors x,y is given as

$$d(x,y;C) = (x-y)^T C^{-1}(x-y)$$

Mahalanobis distance is a true distance metric, since it has the following required properties:

1. $d(x,y;C) = 0 \iff x = y$ (identity of indiscernibles)
2. $d(x,y;C) = d(y,x;C)$ (symmetry)
3. $d(x,y;C) \leq d(x,z;C) + d(z,y;C)$ (triangle inequality)(can be proved using Cauchy-Schwarz inequality)
4. $d(x,y;C) \geq 0$ (non-negativity) (can be proved from above 3 properties)

It can be observed that the level sets of multi-variate Gaussian distribution with covariance matrix C are points with equal Mahalanobis distance. The variance along each axis of the PDF is equal to the eigenvalue corresponding to the axis(each axis corresponds to an eigenvector of C).

8 Applications

1. Anomaly detection: Given a bunch of sample points drawn from a multi-variate Gaussian distribution, we can construct a model the distribution by finding mean and covariance matrix. Then any new point can be associated with the probability of it being drawn from the same distribution and hence can be classified as outlier or inlier.(To get A sunch that $AA^T = C$ carry out eigenvector decomposition of $C = QDQ^T$. Then A is given by $A = Q\sqrt{D}$)
2. Classification: Once the models of multi-variate Gaussian distributions are constructed, given a new data-point, its probability of belonging to each class can be found.

8.1 Decision boundaries

Decisions boundaries refer to the set of points at which PDF's corresponding to two different classes are equal.
In case of multi-variate Gaussian distribution, for two classes, the surface formed is called hyper-quadric surface.

9 Principal Component Analysis

9.1 Modes of variations

Set of vectors representing variations in a given dataset around the mean of the dataset.
We intend to find the k orthogonal unit vectors such that the projection of data point(after subtraction of mean) on the hyperplane spanned by the k vectors yields the maximum variance.
Mathematically,

$$argmin_{v_i} \sum_{i=1}^{n} || \sum_{j=1}^{k} < x_i, v_j > v_j ||^2$$

With constraints,

$$< v_i, v_j >= \delta_{ij}$$

It can be proved that v_i **are the orthogonal eigenvectors of** C **having** k **largest eigenvalues.** These vectors can be found by eigenvector decomposition(singular value decomposition) of the covariance matrix($C = UDU^T$), and sorting the columns of U (the eigenvectors) w.r.t. their eigenvalues, and then finally selecting the top k eigenvectors.

9.2 Dimensionality reduction

Let the matrix of eigenvectors having top k eigenvectors be $M_{n \times k}$.
Any data point x_i having n dimensions can be reduced to a vector of k dimensions by projecting on the hyperplane spanned by the above found eigenvectors such that the loss of information is minimum.
The reduced vector(whose i^{th} element is component along i^{th} column vector of M) is given by $x_{i,reduced} = M^T x_i$.

9.3 Reconstruction of the data point

To get back the approximated x_i from $x_{i,reduced}$:

$$x_{i,approx} = M x_i$$

Also add mean, found earlier, to get the original data point.

10 Singular Value Decomposition(SVD)

Given any real valued matrix $A_{m \times n}$, the matrix can be factorized as $A_{m \times n} = U_{m \times m} S_{m \times n} V^T_{n \times n}$ where,

- U is orthogonal (unitary in case A has complex entries)
- S is rectangular diagonal (the entries of S are non-negative real valued even if A is complex-valued)
- V^T is orthogonal(unitary in case A has complex entries)

If $m \leq n$, then (x_i is i^{th} column of matrix X)

$$A = \sum_{i=1}^{m} s_{ii} u_i v_i^T$$

10.1 Matrix norm

Let the two norm of a vector $v_{n\times 1}$ be denoted as $||v||_2$. For any matrix $A_{m\times n}$ the induced norm is defined as

$$||A||_2 = max_{x\neq 0} \frac{||Ax||_2}{||x||_2}$$

10.2 SVD analysis

Let A be a real valued $m \times n$ matrix. Let $||A||_2 = \sigma_1$ and v_1 be the corresponding unit vector to evaluate $||A||_2$ and $u_1 = Av_1$.
Construct orthogonal matrix U and V such that first column of U is $u_1/||u_1||_2$ and that for V is v_1. Now,

$$U^T AV = \begin{bmatrix} \sigma_1 & w^T \\ 0 & B \end{bmatrix} = S \tag{1}$$

where B is a sub-matrix of shape $(m-1) \times (n-1)$
Note that, $||S||_2 = ||A||_2 = \sigma_1$(since A is multiplied by orthogonal vectors).
Using the definition of induced norm of matrix, and the above fact, it can be proved that $w = 0$.

Hence we have,

$$S = \begin{bmatrix} \sigma_1 & 0 \\ 0 & B \end{bmatrix} \tag{2}$$

Now, using induction it can be proved that B matrix(if exists) is rectangular diagonal, hence the proof of SVD of $A(= USV^T)$.
Therefore,

$$Av_i = S_{ii} u_i$$

Geometrically, A maps orthogonal vectors v_i to columns of U(i.e., u_i, respectively) such that u_i are also orthogonal.

10.3 Analysis of right singular vectors(columns of V)

Let $||A||_2 = \sigma_1$(unique). Is the corresponding vector v_1 unique?(upto sign)
Suppose there exists another vector w such that $||Aw||_2 = \sigma_1$.
It can be proved(by breaking w into components perpendicular and parallel to v_1) that there exists a unit vector x perpendicular to v such that $||Ax||_2 = \sigma_1$.
Consider,

$$\sigma_1 = ||Ax||_2 = ||USV^T x||_2 = ||SV^T x||_2 = ||By||_2$$

where,(using $v_i x^T = 0$)

$$[0, y^T]^T = V^T x$$

It can be shown by contradiction that $||B||_2 \leq ||A||_2$(else a vector can be constructed,$m = V[0, y^T]^T$ which would yield $||Am||_2 > \sigma_1$) . But since,

$$\sigma_1 = ||By||_2$$

σ_1 would be repeated singular value(found after decomposing B). The above argument can be carried out recursively and it can be concluded that,

1. If v_1 is not unique then $sigma_1$ is not simple singular value(i.e., repeated in S)
2. If all singular values are distinct then all singular vectors are unique.