Answer:
Multivariate calculus and statistics

- Sec1.1: Now we need multivariate calculus and multivariate statistics, so we learn them (or review them) now.

  Since, we now need multivariate calculus and multivariate statistics, we learn them (or review them).

- Sec 1.3: Paragraph 1 a is mentioned as scalar and used as a vector
- Sec 1.4: For instance, x^2 may be generalized to XX^T

  For instance, x^2 may be generalized to X^TX

- Sec 2.2: By the same reasoning, if Y is a scaler,

  *By the same reasoning, if Y is a **scalar**,*

  And since usually scalars are denoted by lower case letters y would be a better notation.

  Similarly, the other scaler in the same section should be changed to scalar

  For scaler Y , we can define the Hessian or second derivative

  *For **scalar** Y , we can define the Hessian or second derivative*

- Sec 2.8: Then $AI = (Ae1, \cdots, Ae_n)$

  Then $AI = (Ae_1, \cdots, Ae_n)$

- Sec 2.14: A symmetric matrix is positive definite, denoted $S > 0$,

  *A symmetric matrix is positive definite, denoted **by** $S > 0$,*

- Sec 2.19 No consistency for the notation of X_{ij}. Either lowercase or uppercase needs to be used through the section.

<u>Overfitting:</u>

1. Page1: Background materials: Background **material.** ()
2. Page1: Sec 1.1: an n x p matrix: **a** nxp matrix
3. Page 2: Sec 2.1 : beta = min || Y- X*beta|| : beta = min ||Y- X(**transpose**) *beta||
4. Page 3: Sec 2.2: Pythagarean theorem: **Pythagorean** theorem. Appears twice.
5. Page 3: Sec 2.2: overfitting the data in training, you will not be able to do so in testing: overfitting the data in training **but** you will not be able to do so in testing
6. Page 3: Sec 2.3: Testing output parameters not explained.
7. Page 4: Sec 2.3: Wrong indentation. Second line.
8. Page 4: Sec 2.3: Thus again the overfitting is 2p: Thus**, the overfitting is again** 2p.
9. Page 4: Sec 2.3: if will decrease first: **it** will decrease first
10. Page 4: Sec 2.3: Now consider a general estimator β: **Now,** consider a …
11. Page 4: Sec 2.3: Again the overfitting: **Again,** the overfitting

12. Page 4: Sec 2.4: the second term is 0, and the: the second term is **0 and** the
13. Page 5: Sec 2.5: Comma before beginning of every equation.
14. Page 5: Sec 3.1: Comma before equation.
15. Page 6: Sec 3.2: First line, epsilon notation mismatch
16. Page 6: Sec 3.3: In the case of yi that may depend on Xi . We can consider the following quantity as a measure of overfitting : In the case of yi that may **depend on Xi, we** can consider the following quantity as a measure of **overfitting,**
17. Page 6: Sec 3.3: Unnecessary comma usage in second paragraph.
18. Page 6: Sec 3.4: One may wonder it may not be realistic to: One may wonder **that** it may not be realistic to
19. Page 7: Sec 3.6: Comma before equations

Correlation:

1. Page 1: Background materials: Background **material**
2. Page 1: Sec 1.1: a joint probability density: **is/be** a joint probability density
3. Page 2: Sec 1.2: In the following, we assume n: In the following **example**, we assume n
4. Page 2 and Page 3: Table numbers are improper
5. Page 3: Sec 1.3: each variable corresponds: **Each** variable corresponds
6. Page 3: Sec 1.4: Comma after every equation,
7. Page 3: Sec 1.4: Indentation
8. Page 4: Sec 2.2: For instance, if X is the height,: **Let X be the height,**
9. Page 4: Sec 2.2:: then the covariance is in the unit of inch × pound.: t**hen the unit of covariance is inch × pound.**
10. Page 5: Sec 3.1: of people 40 years old. Given X = x, the: of people 40 years old **given** X = x, the
11. Page 6: Sec 3.1: weight of 40 years old if x = 40: weight of 40 **year** old if x = 40
12. Page 6: Sec 3.1: Thus it won't change : **Thus,** it won't change
13. Page 7: Sec 3.2: decomposed into the between group variance and the within group variance: decomposed into the **'between group variance'** and the **'within group variance'**
14. Page 8: Sec 3.4: Consider the scatter plot, it is possible that within a school: Consider the scatter **plot. It is** possible that within a school
15. Page 8: Sec 3.4: be the remainders. Then it is possible: be the **remainders, then** it is possible

Population:

1. Page 1: Background materials: Background **material**
2. Page 2: Sec 3: on the real line. Then: on the real **line, then**
3. Page 2: Sec 3: Full stop after formula not required.
4. Page 2: Sec 4: as a point. Then we have a distribution of points.: as a **point, then** we have a distribution of points.
5. Page 2: Sec 4: average heights of 30 years old:average heights of 30 **year** old
6. Page 3: Sec 5: n points into Ω. Then the x: n points into Ω**, then** the x
7. Page 4: Sec 6: position at time t, we may model: position at time **t. We** may model
8. Page 5: Sec 7: The same with random relocation in the vertical : **This is** same with random relocation in the vertical
9. Page 5: Sec 8: we want to know what happen as time goes by: we want to know what **happens** as time goes by
10. Page 5: Sec 8: how many people were at state c at time s?: how many people were **in** state c at time s?
11. Page 5: Sec 8: and can be generalize to : and can be **generalized** to

A Note on Multivariate Calculus and Multivariate Statistics

Ying Nian Wu, UCLA Statistics

Background materials for STATS 200A and 202A

Contents

1 Introduction

1.1 A dropped ball

Multivariate calculus does not involve new concepts beyond calculus and linear algebra. It is just that it may not be covered in either the calculus course or the linear algebra course. It falls through the crack and becomes the dropped ball, or a hole in one's undergraduate math education. However, it is immensely useful in statistics and machine learning, because we always deal with multiple variables at a time. The same may happen to multivariate statistics, which may not be covered in either the probability course or the statistics course.

But there is no need to worry. We can always fill the holes in our knowledge at an appropriate time, such as when we need to use them. In fact, having a purpose or motivation is perhaps the most important way to learn things. Now we need multivariate calculus and multivariate statistics, so we learn them (or review them) now.

1.2 Calculation

The calculation aspect of multivariate calculus is based on the matrix multiplication rule. If $C = AB$, then $C_{ij} = \sum_k A_{ik}B_{kj}$. In Einstein's notation, $C_{ij} = A_{ik}B_{kj}$, where the repeated index k is automatically summed. This rule also applies if A or B or both are vectors.

1.3 Vectors

Vectors are special cases of matrices. We usually use column vectors. For a vector v, we can multiply it by a number a. In matrix form, we need to write it as va. But for convenience, we sometimes write it as av, such as in $\sum_k a_k v_k$, to emphasize the meaning that the resulting vector is a linear superposition of some vectors.

The inner product between two vectors $\langle u, v\rangle = u^\top v = v^\top u$. The geometric meaning is $\langle u, v\rangle = |u||v|\cos\theta$, where $|u|$, $|v|$ are the lengths (or ℓ_2 norms) of u and v, i.e., $|u|^2 = \langle u, u\rangle = u^\top u$. If v is a unit vector, and u is another vector, then $\langle u, v\rangle v$ is the projection of u on v. If v is not a unit vector, we can normalize it to $v/|v|$.

A generalization of $|v|^2 = v^\top v$ is the quadratic form $v^\top S v$, where S is a symmetric matrix, and serves as a metric. We will see the meaning of S soon.

For a column vector v, we sometimes need to compute $vv^\top$, which is a squared matrix. This is a multivariate generalization of v^2.

1.4 Packing

A main advantage of matrix is that it packs all its elements so that people won't be drowned in the sea of subscripts.

With such packaging, we can generalize almost all the expressions in univariate calculus to multivariate calculus. We only need to take care of matching the dimensions of the matrices. For

instance, x^2 may be generalized to $XX^\top$, sx^2 may be generalize to $X^\top SX$. The matrix forms are more elegant.

When performing matrix calculations, we can treat the column vectors or the row vectors of a matrix as if they are numbers. In fact, we should always use this mental trick in our calculations.

1.5 Packing the operators

A matrix or a vector can even pack operators such as differentiations. The composition of these operators can be considered multiplication. Then we still have the same rule of matrix multiplication for the matrices or vectors that consist of such operators.

1.6 Noun or verb, change of viewpoint

A matrix collects a bunch of numbers, and can be viewed as a noun. But most of the time, we multiple a matrix to a vector to transform it to another vector, hence a matrix is a verb. The meaning of the verb sometimes means a change of viewpoint. Thus a mathematical expression in terms of such a matrix can often be read in terms of English. That is, matrices give us a mathematical language that encodes rich meanings.

1.7 Not just calculation

The textbook treatment of matrices emphasizes a lot on calculation, but less on the language and meaning aspect. For instance, we start the course by calculating the determinant of a squared matrix, and to use it to compute the inverse of the matrix. But we are usually not taught the geometric meaning of the determinant in terms of how much the matrix stretches or squeezes things. In this note, we shall minimize the efforts in calculation, but maximize our efforts in understanding the meaning.

2 Multivariate calculus

2.1 First derivative

Suppose $Y = (y_i)_{m\times 1}$, and $X = (x_j)_{n\times 1}$. Suppose $Y = h(X)$. We can define

$$\frac{\partial Y}{\partial X^\top} = \left(\frac{\partial y_i}{\partial x_j}\right)_{m\times n}.$$

Here is the key. The above definition is not even necessary, because it follows directly from matrix multiplication. Specifically, we can treat $\partial Y = (\partial y_i, i = 1, ..., m)^\top$ as a column vector, and $1/\partial X = (1/\partial x_j, j = 1, ..., m)^\top$ as another column vector. Now we have two vectors of operations, instead of numbers. The product of the elements of the two vectors is understood as composition of the two operators, i.e., $\partial y_i(1/\partial x_j) = \partial y_i/\partial x_j$. Then $\partial Y/\partial X^\top$ is a squared matrix according to the matrix multiplication rule.

2.2 Second derivative

By the same reasoning, if Y is a scaler, then the gradient $h'(X) = \partial Y/\partial X$ is a $n \times 1$ column vector, and $\partial Y/\partial X^\top$ is a $1 \times n$ row vector. For scaler Y, we can define the Hessian or second derivative

$$h''(X) = \frac{\partial^2 Y}{\partial X \partial X^\top} = \left(\frac{\partial^2 Y}{\partial x_i \partial x_j}\right)_{n\times n}.$$

Again, we can treat $\partial/\partial X$ as a column vector. Then $\partial^2/\partial X \partial X^\top = (\partial/\partial X)(\partial/\partial X)^\top$ is a squared matrix following the rule of vector multiplication.

2.3 Examples

If $Y = AX$, then $y_i = \sum_k a_{ik} x_k$. Thus $\partial y_i/\partial x_j = a_{ij}$. So $\partial Y/\partial X^\top = A$.
If $Y = X^\top S X$, where S is symmetric, then $\partial Y/\partial X = 2SX$, and $\partial^2 Y/\partial X \partial X^\top = 2S$.
If $S = I$, $Y = \|X\|^2$, $\partial Y/\partial X = 2X$.
The above results generalize the scalar results with almost no change in notation.

2.4 Chain rule

If $Y = h(X)$ and $X = g(Z)$, then $\partial y_i/\partial z_j = \sum_k (\partial y_i/\partial x_k)(\partial x_k/\partial z_j)$. Thus

$$\frac{\partial Y}{\partial Z^\top} = \frac{\partial Y}{\partial X^\top}\frac{\partial X}{\partial Z^\top}.$$

2.5 Product rule

If $Y = \langle h(X), g(X)\rangle = \sum_i h_i(X) g_i(X)$, then $\partial Y/\partial x_j = \sum_i [\partial h_i/\partial x_j g_i + h_i \partial g_i/\partial x_j]$. So

$$\frac{\partial Y}{\partial X^\top} = h(X)^\top \frac{\partial g(X)}{\partial X^\top} + g(X)^\top \frac{\partial h(X)}{\partial X^\top}.$$

2.6 Taylor expansion

For a scalar $f(X)$,

$$f(X) = f(X_0) + \langle f'(X_0), X - X_0\rangle + \frac{1}{2}(X - X_0)^\top f''(X_0)(X - X_0) + o(|X - X_0|^2).$$

This can be proved as follows. Let $u = (X - X_0)/|X - X_0|$ be the unit vector from X_0 to X, and let $t = |X - X_0|$. Then $X = X_0 + tu$. Let

$$g(t) = f(X) = f(X_0 + tu).$$

Then according to the univariate Taylor expansion

$$g(t) = g(0) + g'(0)t + g''(0)t^2/2 + o(t^2).$$

Let $Y = g(t) = f(X_0 + ut)$. According to the chain rule,

$$g'(t) = \frac{\partial Y}{\partial X^\top}\frac{\partial X}{\partial t} = f'(X)^\top u = u^\top f'(X),$$

where $f'(X) = \partial Y/\partial X$. Let $Z = f'(X)$, then $g'(t) = u^\top Z$, and

$$g''(t) = u^\top \frac{\partial Z}{\partial t} = u^\top \frac{\partial Z}{\partial X^\top}\frac{\partial X}{\partial t} = u^\top f''(X) u.$$

Thus the above Taylor expansion follows.

2.7 Gradient: steepest direction

Since $g'(t) = \langle f'(X), u\rangle = |f'(X)|\cos\theta$, where θ is the angle between the unit vector u and $f'(X)$. Its magnitude is maximized if u is aligned with $f'(X)$. Thus $f'(X)$ is the direction with the steepest change. We call $f'(X)$ the gradient.

2.8 Determinant

For an $n \times n$ invertible squared matrix A, let $Y = AX$, then A maps a region in the space of X to a region in the space of Y. The determinant measures the change of the volume caused by the transformation A.

Specifically, consider the unit cube spanned by $(e_i, i = 1, ..., n)$, where $I = (e_1, ..., e_n)$, i.e., e_i is the i-th column of the identity matrix. Then $AI = (Ae1, ..., Ae_n) = (a_1, ..., a_n)$, i.e., A maps e_i to a_i, the i-th column of A. Then A maps the unit cube spanned by $(e_i, i = 1, ..., n)$ to the parallelogram spanned by $(a_i, i = 1, ..., n)$. The volume of the unit cube is 1. Thus the determinant is the volume of this parallelogram.

2.9 Jacobian

Let $Y = h(X)$ where both X and Y are $n \times 1$. Assume that h is one-to-one differentiable mapping. According to the first order Taylor expansion, $h(X) = h(X_0) + h'(X_0)(X - X_0) + o(|X - X_0|)$, where $h'(X) = \partial Y/\partial X^\top$ is the first derivative. Thus locally around each X_0, the mapping h is a linear mapping by a matrix $A = h'(X_0)$.

Let D_X be a local region around X in the domain of X. Suppose h maps D_X to a region D_Y in the domain of Y. Then as the size of D_X goes to 0, $|D_Y|/|D_X| \to |h'(X)|$, where $|h'(X)|$ is the determinant of $h'(X) = \partial Y/\partial X^\top$. This is called the Jacobian of h.

2.10 Orthogonal matrix: viewpoint

An orthogonal matrix $Q = (q_1, ..., q_n)$ is such as $\langle q_i, q_j\rangle = \delta_{ij}$, where $\delta_{ij} = 1$ if $i = j$ and $\delta_{ij} = 0$ otherwise. That is $(q_1, ..., q_n)$ form an orthonormal basis.

For a vector v, we can view it in Q, so that $v = q_1u_1 + ... + q_nu_n = (q_1, ..., q_n)u = Qu$. In the last step of the calculation, we may treat each q_i as a number in our mental calculation. Each $u_i = \langle v, q_i\rangle = q_i^\top u$. Thus $u = Q^\top u$, where again we can treat each $q_i^\top$ as a number in our mental calculation. Thus, v becomes $u = (u_1, ..., u_n)^\top$ from the point of view of Q. $u = Q^\top v$ is analysis, i.e., we decompose v into pieces along $(q_1, ..., q_n)$. $v = Qu$ is synthesis, i.e., we put the pieces together to get back v. Clearly $Q^\top = Q^{-1}$, i.e., $QQ^\top = Q^\top Q = I$.

2.11 Symmetric matrix becomes a diagonal matrix

For any symmetric matrix S, we can diagonalize it by $S = Q\Lambda Q^\top$, where Q is an orthogonal matrix, and Λ is a diagonal matrix.

As a verb, we can let S act on a vector v to get Sv.

From the point of view of Q, v is u, i.e., $v = Qu$. Then $Sv = SQu = Q\Lambda u$. Thus from the point of view of Q, Sv is Λu.

Thus from the point of view of Q, the verb S is Λ, which acts on each element individually, i.e., $\Lambda u = (\lambda_1u_1, ..., \lambda_nu_n)^\top$ (again we can treat each u_i as a number in our mental calculation).

2.12 Power and square root

$S^t = Q\Lambda^t Q^\top$, where Λ^t is the diagonal matrix formed by $(\lambda_i^t, i = 1, ..., n)$. This can be easily proved. This underlies the power of the viewpoint Q, because Λ^t can be obtained by element-wise operation. This underlies the power method for computing Λ and Q, where t indexes the iterations of the algorithm.

We define $S^{1/2} = Q\Lambda^{1/2}Q^\top$. $S^{-1/2} = Q\Lambda^{-1/2}Q^\top$, where $\Lambda^{1/2}$ and $\Lambda^{-1/2}$ are obtained by element-wise operation.

2.13 Quadratic form as sum of squares, metric

For a quadratic form $X^\top SX$ where S is symmetric, we can find Q, so that $S = Q\Lambda Q^\top$. From the point of view of Q, X becomes Z, i.e., $X = QZ$. Then S becomes Λ, and the quadratic form $X^\top SX = Z^T \Lambda Z = \sum_i \lambda_i z_i^2$.

The contour $X^\top SX = c$ becomes $Z^\top \Lambda Z = c$ from the viewpoint Q, which is an ellipse.

Let $\tilde{z}_i = z_i\sqrt{\lambda_i}$, i.e., $\tilde{Z} = \Lambda^{1/2}Z$, then $X^\top SX = Z^\top \Lambda Z = \tilde{Z}^\top \tilde{Z} = \sum_i \tilde{z}_i^2$. That is, S (or more precisely S^{-1}) plays the role of a metric (or unit), which measures the magnitude of the component z_i by $z_i\sqrt{\lambda_i}$. It is as if different z_i are measured in different units $1/\sqrt{\lambda_i}$.

2.14 Positive definite

A symmetric matrix is positive definite, denoted $S > 0$, if all the $\lambda_i > 0$. It means $X^\top SX > 0$ for any non-zero X. In the above discussion of metric, we assume $S > 0$, so that $S^{-1} > 0$. If $S > 0$, the function $f(X) = X^\top SX$ is concave.

2.15 Hessian in second order Taylor expansion

In Taylor expansion,

$$f(X) = f(X_0) + \langle f'(X_0), X - X_0\rangle + \frac{1}{2}(X - X_0)^\top f''(X_0)(X - X_0) + o(|X - X_0|^2),$$

$f''(X)$ is the Hessian matrix. $g''(t) = (X - X_0)^\top f''(X_0)(X - X_0)$ is a quadratic form.

Let $f''(X_0) = Q\Lambda Q^\top > 0$, then from the viewpoint Q, the contour of the expansion is an ellipse.

If $f''(X) > 0$ for all X, then $f(X)$ is concave, and has a single global maximum.

2.16 Newton-Raphson

Take derivative of the Taylor expansion with respect to X, and set it to 0, we have

$$f'(X_0) + f''(X_0)(X - X_0) = 0.$$

Let X_1 be the solution to the above equation, then

$$X_1 = X_0 - [f''(X_0)]^{-1} f'(X_0).$$

X_1 is the center of the above-mentioned ellipse.

The change from X_0 to X_1 is an iteration of Newton-Raphson.

2.17 Least squares

Let the data frame be $(\mathbf{X}, \mathbf{Y})$, where $\mathbf{X}$ is $n \times p$ and Y is $n \times 1$. The model is $\mathbf{Y} = \mathbf{X}\beta_{\text{true}} + \epsilon$, where β is $p \times 1$, and ϵ is $n \times 1$.

Let $R(\beta) = \|\mathbf{Y} - \mathbf{X}\beta\|^2$ be the least squares loss function, then

$$R'(\beta) = -2\mathbf{X}^\top(\mathbf{Y} - \mathbf{X}\beta)$$

and

$$R''(\beta) = 2\mathbf{X}^\top\mathbf{X}.$$

We can derive these by the chain rule. Let $e = \mathbf{Y} - \mathbf{X}\beta$. Then

$$\frac{\partial R}{\partial \beta^\top} = \frac{\partial R}{\partial e^\top}\frac{\partial e}{\partial \beta^\top} = -2e^\top\mathbf{X}.$$

$R'(\beta) = \partial R/\partial\beta$, which is obtained by transposing $-2e^\top\mathbf{X}$.

$$R''(\beta) = \frac{\partial^2 R}{\partial\beta\partial\beta^\top} = \partial(-2\mathbf{X}^\top e)/\partial\beta^\top = -2\mathbf{X}^\top\mathbf{X} < 0.$$

The least square solution is

$$\hat{\beta} = (\mathbf{X}^\top\mathbf{X})^{-1}(\mathbf{X}^\top\mathbf{Y}).$$

It is the single global minimum of $R(\beta)$.

2.18 Similarity, diagonalization, and eigen vectors

For a squared matrix K, usually we can decompose it $K = V\Lambda V^{-1}$, where V is an invertible squared matrix. K is said to be similar to the diagonal matrix Λ. For symmetric K, $V = Q$.

Let $V = (v_1, ..., v_n)$. Then

$$KV = (Kv_1, ..., Kv_n) = V\Lambda = (\lambda_1 v_1, ..., \lambda_n v_n),$$

i.e., $Kv_i = \lambda_i v_i$ for $i = 1, ..., n$, and v_i are the right eigen vectors with eigen values λ_i.

Let $U = V^{-1}$, and let $U = (u_1, ..., u_n)^\top$. Then

$$UK = \Lambda U,$$

i.e., $u_i^\top K = \lambda_i u_i^\top$. u_i are called left eigen vectors with eigen values λ_i.

$K^t = V\Lambda^t V^{-1}$. This underlies the convergence rate of the Markov chain.

2.19 Matrix derivative

Suppose $X = (X_{ij})$ is an invertible squared matrix, and $Y = f(X)$ is a scalar function, we can define $f'(X) = (\partial Y/\partial x_{ij})$.

If $Y = f(X) = \log|X|$, where $|X|$ is the determinant of X. Then $f'(X) = X^{-1}$.

If $Y = f(X) = a^\top Xa$, then $f'(X) = aa^\top$.

In fact, we can write $Y = \text{tr}(aa^\top X)$, where $\text{tr}(A) = \sum_i a_{ii}$ denotes the trace of a matrix, and $\text{tr}(AB) = \text{tr}(A)\text{tr}(B)$. If $Y = f(X) = \text{tr}(AX)$, then $f'(X) = A$.

3 Multivariate statistics

3.1 Expectation of a random matrix

Consider a random matrix X. Suppose X is $m \times n$, and the elements of X are x_{ij}, $i = 1, ..., m$ and $j = 1, ..., n$. Usually we write $X = (x_{ij})_{m \times n}$ or simply $X = (x_{ij})$. We define

$$\mathrm{E}(X) = (\mathrm{E}(x_{ij})),$$

i.e., taking expectations element-wise. Let A be a constant matrix of appropriate dimension, then $\mathrm{E}(AX) = A\mathrm{E}(X)$. Let B be another constant matrix of appropriate dimension, then $\mathrm{E}(XB) = \mathrm{E}(X)B$.

The above result can be easily understood if we have iid copies $X_1, ..., X_n$, so that $\sum_{i=1}^n X_i/n \to \mathrm{E}(X)$, and $\sum_{i=1}^n AX_i/n \to \mathrm{E}(AX)$, but $\sum_{i=1}^n AX_i/n = A\sum_{i=1}^n X_i/n \to A\mathrm{E}(X)$. Thus $\mathrm{E}(AX) = A\mathrm{E}(X)$.

3.2 Variance of a random vector

Let X be a random vector. Let $\mu_X = \mathrm{E}(X)$. We define

$$\mathrm{Var}(X) = \mathrm{E}[(X - \mu_X)(X - \mu_X)^\top].$$

Then the (i, j)-th element of $\mathrm{Var}(X)$ is $\mathrm{Cov}(x_i, x_j)$. The diagonal elements are $\mathrm{Var}(x_i)$.

Let A be a constant matrix of appropriate dimension, then

$$\mathrm{Var}(AX) = A\mathrm{Var}(X)A^\top.$$

This is because

$$\begin{aligned}
\mathrm{Var}(AX) &= \mathrm{E}[(AX - \mathrm{E}(AX))(AX - \mathrm{E}(AX))^\top] \\
&= \mathrm{E}[(AX - A\mu_X)(AX - A\mu_X)^\top] \\
&= \mathrm{E}[A(X - \mu_X)(X - \mu_X)^\top A^\top] \\
&= A\mathrm{E}[(X - \mu_X)(X - \mu_X)^\top]A^\top \\
&= A\mathrm{Var}(X)A^\top.
\end{aligned}$$

Note that A does not need to be a squared matrix. A can even be a vector, such as $a^\top$, then $\mathrm{Var}(a^\top X) = a^\top \mathrm{Var}(X)a$, which is a quadratic form.

3.3 Covariance between two random vectors

We can also define

$$\mathrm{Cov}(X, Y) = \mathrm{E}[(X - \mu_X)(Y - \mu_Y)^\top],$$

then

$$\begin{aligned}
\mathrm{Cov}(AX, BY) &= \mathrm{E}[(AX - A\mu_X)(BY - B\mu_Y)^\top] \\
&= \mathrm{E}[A(X - \mu_X)(Y - \mu_Y)^\top B^\top] \\
&= A\mathrm{E}[(X - \mu_X)(Y - \mu_Y)^\top]B^\top \\
&= A\mathrm{Cov}(X, Y)B^\top
\end{aligned}$$

3.4 Variance matrix diagonalized

$\Sigma = \mathrm{Var}(X) > 0$ because for any non-zero vector a, $a^\top \Sigma a = \mathrm{Var}(a^\top X) > 0$.

Σ can be diagonalized into $\Sigma = Q\Lambda Q^\top$, where $\Lambda = \mathrm{diag}(\lambda_1, ..., \lambda_n)$ is a diagonal matrix. From the viewpoint Q, X becomes Z, i.e., $X = QZ$ or $Z = Q^\top X$, and $\mathrm{Var}(Z) = Q^\top \mathrm{Var}(X) Q = \Lambda$. Thus the elements z_i of Z are uncorrelated. That is, from the viewpoint of Q, the variance matrix Σ becomes Λ.

3.5 Principal component analysis

Assuming $\mathrm{E}(X) = 0$ (otherwise we can let $X \leftarrow X - \mathrm{E}(X)$), and $\mathrm{Var}(X) = \Sigma = Q\Lambda Q^\top$. Then viewed from Q, $\mathrm{E}(Z) = 0$ and $\mathrm{Var}(Z) = \Lambda$.

Assume $\lambda_1 > \lambda_2 > ... > \lambda_n > 0$. If $\lambda_i = \mathrm{Var}(z_i)$ is very small for $i > m$, then $z_i \approx 0$ for $i > m$ (recall $\mathrm{E}(z_i) = 0$). We can represent

$$X \approx \sum_{i=1}^{m} q_i z_i,$$

thus reducing the dimensionality of X from n to m. The $(q_i, i = 1, ..., m)$ are called principal components.

For instance, if X is a face image, then $(q_i, i = 1, ..., m)$ are the eigen faces, which may correspond to different features of a face (e.g., eyes, nose, mouth etc.), and $(z_i, i = 1, ..., m)$ is a low dimensional representation of X.

3.6 Probability density under transformation

Let $X \sim f_X(X)$, the probability density of X. f_X is a single whole notation, where f is like a last name and the subscript X is like a first name.

Let A be an invertible squared matrix, and let $Y = AX$. Let the density of Y be $f_Y(Y)$. Again f_Y is a whole notation.

Consider a small neighborhood around X, denoted D_X. A maps it to a small neighborhood around $Y = AX$, denoted D_Y. The change of volumes caused by A is $|D_Y|/|D_X| = |A|$, the determinant of A.

The density of Y is

$$f_Y(Y) = \frac{P(Y \in D_Y)}{|D_Y|} = \frac{P(X \in D_X)}{|D_Y|} = \frac{P(X \in D_X)}{|A||D_X|} = f_X(A^{-1}Y)/|A|.$$

Symbolically,

$$X \sim f_X(X)dX \sim f_X(A^{-1}Y)dA^{-1}Y \sim f_X(A^{-1}Y)|A|^{-1}dY \sim f_Y(Y)dY.$$

For a non-linear invertible transformation $Y = h(X)$,

$$f_Y(Y) = f_X(X)/|h'(X)|, \;\; X = h^{-1}(Y),$$

where $|h'(X)|$ is the determinant of $\partial Y/\partial X^\top$, i.e., the Jacobian, whose geometric meaning is $|D_Y|/|D_X|$ mentioned above.

For instance, the Jacobian for the polar transformation $X = R\cos\theta$, $Y = R\sin\theta$ is

$$|\partial(X, Y)/\partial(R, \theta)^\top| = r.$$

3.7 Multivariate normal

We start from $Z = (z_1, ..., z_n)^\top$, where $z_i \sim \mathrm{N}(0,1)$ independently. Then $\mathrm{E}(Z) = 0$, and $\mathrm{Var}(Z) = I$. We denote $Z \sim \mathrm{N}(0, I)$. The density of Z is

$$f_Z(Z) = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}\sum_i z_i^2\right] = \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2} Z^\top Z\right].$$

Let $X = \mu + \Sigma^{1/2} Z$, then $Z = \Sigma^{-1/2}(X - \mu)$, which is a matrix version of standardization. Then

$$\begin{aligned} f_Y(Y) &= \frac{1}{(2\pi)^{n/2}} \exp\left[-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right] / |\Sigma^{1/2}| \\ &= \frac{1}{(2\pi)^{n/2}|\Sigma|^{1/2}} \exp\left[-\frac{1}{2}(X-\mu)^\top \Sigma^{-1}(X-\mu)\right]. \end{aligned}$$

Moreover, because $X = \mu + \Sigma^{1/2} Z$, we have $\mathrm{E}(X) = \mu$ and $\mathrm{Var}(X) = \Sigma$. We denote $X \sim \mathrm{N}(\mu, \Sigma)$.

From the viewpoint Q and the metric Λ, X is Z.

In general, if $X \sim \mathrm{N}(\mu, \Sigma)$, and $Y = AX$, then $Y \sim \mathrm{N}(A\mu, A\Sigma A^\top)$. A does not need to be a squared matrix.

3.8 Marginal and conditional of multivariate normal

Assuming $\mathrm{E}(X) = 0$ (otherwise we can always let $X \leftarrow X - \mathrm{E}(X)$). Let us partition X into (X_1, X_2).

$$\begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & \Sigma_{12} \\ \Sigma_{21} & \Sigma_{22} \end{pmatrix}\right).$$

Let $\epsilon = X_2 - AX_1$. We choose A to make $\mathrm{Cov}(\epsilon, X_1) = 0$,

$$\begin{aligned} \mathrm{Cov}(\epsilon, X_1) &= \mathrm{Cov}(X_2 - AX_1, X_1) \\ &= \mathrm{Cov}(X_2, X_1) - A\mathrm{Cov}(X_1, X_1) \\ &= \Sigma_{21} - A\Sigma_{11} = 0, \end{aligned}$$

so $A = \Sigma_{21}\Sigma_{11}^{-1}$, and $X_2 = AX_1 + \epsilon$. This can be considered a regression of X_2 on X_1. The residual variance is

$$\begin{aligned} \mathrm{Var}(\epsilon) &= \mathrm{Cov}(\epsilon, \epsilon) \\ &= \mathrm{Cov}(X_2 - AX_1, \epsilon) \\ &= \mathrm{Cov}(X_2, \epsilon) \\ &= \mathrm{Cov}(X_2, X_2 - AX_1) \\ &= \mathrm{Cov}(X_2, X_2) - \mathrm{Cov}(X_2, AX_1) \\ &= \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}. \end{aligned}$$

Thus

$$\begin{pmatrix} X_1 \\ \epsilon \end{pmatrix} = \begin{pmatrix} I_1 & 0 \\ -A & I_2 \end{pmatrix} \begin{pmatrix} X_1 \\ X_2 \end{pmatrix} \sim \mathrm{N}\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma_{11} & 0 \\ 0 & \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12} \end{pmatrix}\right).$$

So the marginal distribution

$$X_1 \sim \mathrm{N}(0, \Sigma_{11}),$$

and the conditional distribution

$$[X_2 | X_1] \sim \mathrm{N}(\Sigma_{21}\Sigma_{11}^{-1}X_1, \Sigma_{22} - \Sigma_{21}\Sigma_{11}^{-1}\Sigma_{12}),$$

which is a multivariate linear regression.

3.9 Maximum likelihood

Suppose we observe $(X_i, i = 1, ..., n) \sim \mathrm{N}(\mu, \Sigma)$ independently, and X_i is $p \times 1$. Then the log-likelihood function

$$L(\mu, \Sigma) = \log \prod_{i=1}^{n} f(X_i) = -\frac{1}{2} \sum_{i=1}^{n} (X_i - \mu)^\top \Sigma^{-1} (X_i - \mu) - \frac{n}{2} \log |\Sigma|.$$

Let $A = \Sigma^{-1}$, by setting $\partial L / \partial \mu = 0$ and $\partial L / \partial A = 0$, we can find the maximum likelihood estimate

$$\hat{\mu} = \bar{X} = \frac{1}{n} \sum_{i=1}^{n} X_i.$$

$$\hat{\Sigma} = S = \frac{1}{n} \sum_{i=1}^{n} (X_i - \bar{X})(X_i - \bar{X})^\top.$$

3.10 Singular value decomposition via PCA

Assuming $\bar{X} = 0$ (otherwise we can let $X_i \to X_i - \bar{X}$). Diagonalizing

$$S = \sum_{i=1}^{n} X_i X_i^\top / n = Q \Lambda Q^\top.$$

From the viewpoint Q, X_i becomes $Z_i = Q^\top X_i$, and the sample variance S becomes

$$\Lambda = \frac{1}{n} \sum_{i=1}^{n} Z_i Z_i^\top.$$

Let $\tilde{Z}_i = \Lambda^{-1/2} Z_i$, i.e., we scale the elements of Z_i by $\sqrt{\lambda_i}$, then the sample variance of $\tilde{Z}_i$ is

$$\frac{1}{n} \sum_{i=1}^{n} \tilde{Z}_i \tilde{Z}_i^\top = I.$$

Write $\mathbf{X} = (X_1, ..., X_n)^\top$ the $n \times p$ matrix. Let $\tilde{\mathbf{Z}} = (\tilde{Z}_1, ..., \tilde{Z}_n)^\top$ be the transformed matrix by the above standardization. You can imagine $(\mathbf{X}, \tilde{\mathbf{Z}})$ as two data frames side by side. Then

$$\mathbf{X} = \mathbf{Z} D Q,$$

where $D = \mathrm{diag}(\lambda_1^{1/2}, ..., \lambda_n^{1/2})$. The above is a form of singular value decomposition (SVD), with

$$\mathbf{Z}^\top \mathbf{Z} / n = \sum_{i=1}^{n} Z_i Z_i^\top / n = I.$$

The interpretation is as follows. Q rotates the row vectors (p-dimensional observations) of $\mathbf{X}$ to get $\mathbf{Z}$. The column vectors of $\mathbf{Z}$ (corresponding to p variables after transformation) are orthogonal. The squared lengths of the column vectors of $\mathbf{Z}$ are the diagonal elements of Λ.