Proofreading on the Multivariate Note

BHARGAV PARSI(804945591)

MISTAKE 1

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.

Correction → change **multiple** to **multiply.**

MISTAKE 2

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.

$(1/\delta x_j, j = 1,..m)^T$ → $(1/\delta x_j, j = 1,..n)^T$

Because X is a n x 1 matrix.

MISTAKE 3

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 $Y = X^\top SX$, 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.

$\partial y/\partial x$ → $\partial y/\partial x^T$

We always mean to use Transpose in the first derivative.

MISTAKE 4 AND 5

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., $QQ^\top = Q^\top Q = I$.

$u_i = \langle v, q_i \rangle$ → $v = \langle u_i , q_i \rangle$

$u = Q^T u$ → $v = Q^T u$

The above two equations are mistakes because v is defined as the inner product of q and u.

MISTAKE 6

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 $R(\beta)$.

Minumum will be obtained only if double derivative is > 0.

$-2X^TX < 0$ → $2X^TX > 0$

MISTAKE 7

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 $Y = h(X)$,

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

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

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

The yellow highlight must be “R” but not “r”.

MISTAKE 8

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.

X = QZ must be defined in the paragraph.

MISTAKE 9

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.

$f_Y(Y) \rightarrow f_z(Z)$

MISTAKE 10

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 $D = \text{diag}(\lambda_1^{1/2}, ..., \lambda_n^{1/2})$. 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.

The squared lengths of the column vectors of Z are the diagonal elements of Λ. → The squared lengths of the column vectors of D are the diagonal elements of Λ.